DECLARATION OF JUNE ANN MUNFORD

Introduction and Qualifications

1. My name is June Ann Munford. I am over the age of 18, have personal knowledge of the facts set forth herein, and am competent to testify to the same.

2. I earned a Master of Library and Information Science (MLIS) from the University of Wisconsin-Milwaukee in 2009. I have over ten years of experience in the library/information science field. Beginning in 2004, I have served in various positions in the public library sector including Assistant Librarian, Youth Services Librarian and Library Director. I have attached my Curriculum Vitae as Appendix CV.

3. During my career in the library profession, I have been responsible for materials acquisition for multiple libraries. In that position, I have cataloged, purchased and processed incoming library works. That includes purchasing materials directly from vendors, recording publishing data from the material in question, creating detailed material records for library catalogs and physically preparing that material for circulation. In addition to my experience in acquisitions, I was also responsible for analyzing large collections of library

IPR2026-00272
Krisp Technologies EX1010 Page 2

materials, tailoring library records for optimal catalog search performance and creating lending agreements between libraries during my time as a Library Director.

Library Cataloging and Bibliographic Practices

4. I am fully familiar with the catalog record creation process in the library sector. In preparing a material for public availability, a library catalog record describing that material would be created. These records are typically written in Machine Readable Catalog (herein referred to as "MARC") code and contain information such as a physical description of the material, metadata from the material's publisher, and date of library acquisition. In particular, the 008 field of the MARC record is reserved for denoting the date of creation of the library record itself. As this typically occurs during the process of preparing materials for public access, it is my experience that an item's MARC record indicates the date of an item's public availability.

5. Typically, in creating a MARC record, a librarian would gather various bits of metadata such as book title, publisher and subject headings among others and assign each value to a relevant numerical field. For example, a book's physical description is tracked in field 300 while title/attribution is tracked in

field 245. The 008 field of the MARC record is reserved for denoting the creation of the library record itself. As this is the only date reflecting the inclusion of said materials within the library's collection, it is my experience that an item's 008 field accurately indicates the date of an item's public availability.

The Internet Archive and the Wayback Machine

6. I am familiar with the Internet Archive, a digital library accessible to the general public. Among other services that the Internet Archive makes available to the general public is the Wayback Machine, an online archive of the web. The Internet Archive's Wayback Machine service archives webpages as of a certain capture date to track changes in the web over time. The Internet Archive has been in operation as a nonprofit library since 1996 and has hosted the Wayback Machine service since its inception in 2001. During my time as a librarian, I frequently used the Internet Archive's Wayback Machine for research and instruction purposes. This includes teaching instructional classes on using the Wayback Machine to library patrons and using the Wayback Machine to research reference inquiries that require hard-to-find online resources. I consider the Internet Archive's recordkeeping to be as rigorous

and detailed as other formal library recordkeeping practices such as MARC records, OCLC records and Dublin Core.

Accentron

7. I have reviewed Exhibit ####, "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" by Shaojin Ding, et al. ("Accentron") as published by Elsevier in 2021.

8. Attached hereto as Appendix DING01 is a true and correct copy of "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" as published in Computer Speech & Language as held by the Penn State University Library. These scans were secured from the library's collection. In comparing Exhibit #### to Appendix DING01, it is my determination that Exhibit #### is a true and correct copy of Ding as published in Computer Speech & Language.

9. Attached hereto as Appendix DING02 is a true and correct copy of the MARC record for Computer Speech & Language as held by the Penn State University Library. I secured this record myself from the library's public catalog. The

IPR2026-00272
Krisp Technologies EX1010 Page 5

MARC record contained within Appendix DING02 accurately describes the title, publisher and ISSN number of Computer Speech & Language. The Showing Results field present on page 3 of this record indicates the library's collection includes the March 2022 issue of Computer Speech & Language containing “Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning”.

10. The 008 field of the MARC record in Appendix DING02 indicates the date of record creation. The 008 field of Appendix DING02 indicates the Penn State University Library first acquired Computer Speech & Language as of “970531” (May 31, 1997) and held the journal in perpetuity until the publication of Ding. The 260 field of Appendix DING02 indicates this journal was published by Elsevier, the 538 field of Appendix DING02 indicates this record is describing an e-journal accessed via the internet and the Showing Results page visible on page 3 of Appendix DING02 indicates the Computer Speech & Language journal is accessible via an Elsevier platform, ScienceDirect. Considering these fields, it is my determination Penn State University Library users would be able to access the Computer Speech & Language journal via ScienceDirect between the dates of October 1, 2021 and December 31, 2021.

11. Attached hereto as Appendix DING03 is a screen capture of a webpage entitled "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" as hosted on the Elsevier ScienceDirect website. I secured these screen captures myself from https://www.sciencedirect.com/science/article/abs/pii/S0885230821001029.

12. Appendix DING03 is a capture of the website for the publisher of Computer Speech & Language. This capture describes "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" as published in Computer Speech & Language. The tope of page 1 includes link to "View Open Manuscript," which directs a user to Ding, Exhibit ####. Page 1 also contains a Digital Object Identifier (DOI) URL at https://doi.org/10.1016/j.csl.2021.101302, which is meant to serve as a permanent online location for Ding. The "2021" within the URL indicates that the DOI was assigned as of 2021, suggesting its date of creation.

13. Appendix DING03 also contains a sidebar on the righthand side of the page titled "What do these dates mean?" with definitions corresponding to the various dates listed below the authors on Page 1—"Received 12 April 2021,

Revised 19 September 2021, Accepted 21 September 2021, Available online 5 October 2021, Version of Record 3 December 2021." Three of these dates are relevant to understanding terms used below in this Declaration: Accepted represents "[t]he date the article was accepted for publication in the journal," Available online represents "[t]he date a version of the article was made available online in the journal," and Version of Record represents "[t]he date the finalized version of the article was made available in the journal." The "Available online" date indicates that Ding was made available to the public via ScienceDirect as of October 5, 2021, five months prior to its formal publication in the print edition of Computer Speech & Language March 2022.

14. Attached hereto as Appendix DING04 is a screen capture of the Internet Archive Wayback Machine entry for https://service.elsevier.com/app/answers/detail/a_id/11241/supporthub/scopus/. I secured these screen captures myself from https://web.archive.org/web/20220123095825/https://service.elsevier.com/app/answers/detail/a_id/11241/supporthub/scopus/. This record was captured as of January 23, 2022.

15. Appendix DING04 is an Internet Archive capture of a web page entitled What are Articles in Press? as hosted on the Elsevier Support Center website. Appendix DING04 describes how journal articles are made publicly accessible on the ScienceDirect website prior to formal publication in the paper edition of the journal. Those articles that have been accepted for publication in a journal, but have not yet published in that journal, are called "Articles in Press." The following passage provides more detail: "[Elsevier] makes the pre-published versions of documents available online within four days after being transferred from publishers [. . .] Once the final version is published and made available, [Elsevier] will overwrite the Article in Press in the database with the final version." Per the definitions in Appendix DING03, the "final version" is formally known as the "Version of Record." This page also indicates that users would be able to access the articles in press associated with a given journal solely by entering the journal's name into the website's search bar. As the Internet Archive first captured this page as of January 23, 2022, it is my determination that articles in press associated with Computer Speech & Language would have been discoverable to members of the public prior to the publication of Ding in Computer Speech & Language in March 2022.

16. Attached hereto as Appendix DING05 is a screen capture of the Internet Archive Wayback Machine entry for https://www.sciencedirect.com/journal/computer-speech-and-language/articles-in-press. I secured these screen captures myself from https://web.archive.org/web/20200206160426/https://www.sciencedirect.com/journal/computer-speech-and-language/articles-in-press. This record was captured as of February 6, 2020.

17. Appendix DING05 is an Internet Archive capture of a web page entitled Computer Speech & Language – Articles in Press as hosted on the ScienceDirect website. Appendix DING05 is a list of the Articles in Press for the Computer Speech & Language journal as of the capture date. Page 3 defines Articles in Press similarly to Appendix DING04 above, with the additional clarification that when the Article in Press version is replaced by the Version of Record of an article, "[t]he date an article was first made available online will be carried over." Per Appendix DING03, Ding's Version of Record was published December 3, 2021. Per Appendix DING05, the "Available online" date for Ding remains the same as it was prior to the publication of the Version of Record. This further supports Appendix DING03, that Ding was made "Available online" to the public via

ScienceDirect as of October 5, 2021, five months prior to its formal publication in the print edition of Computer Speech & Language March 2022.

18. Appendix DING05 also clarifies that Articles in Press exist in three forms—journal pre-proofs, uncorrected proofs, and corrected proofs. Of particular note, journal pre-proofs are versions of an article that "will undergo additional copyediting, typesetting and review before being published in final form, but *are provided to give early visibility of the article*" (emphasis added). The listing of Articles in Press captured in Appendix DING05 includes several journal pre-proofs. As with Appendix DING04, this capture indicates Elsevier ScienceDirect made Articles in Press accessible to the public, even in their earliest forms. As the Internet Archive first captured this page as of February 6, 2020, this indicates that it was ScienceDirect's regular practice to make Articles in Press associated with Computer Speech & Language publicly available before their formal publication in the journal. As such, it is my determination that Ding would have been discoverable to members of the public as an Article in Press as of October 5, 2021, or at the latest as of December 3, 2021.

19. Attached hereto as Appendix DING06 is a screen capture of the Internet Archive Wayback Machine entry for https://guanlongzhao.github.io/. I secured these screen captures myself from https://web.archive.org/web/20211019060648/https://guanlongzhao.github.io/. This record was captured as of October 19, 2021.

20. Appendix DING06 is an Internet Archive capture of a web page entitled "Guanlong Zhao" as hosted on the GitHub website. GitHub is a website that frequently serves as a curriculum vitae and code repository for many professionals working in the sciences.

21. Appendix DING06 is the GitHub page for Guanlong Zhao, one of the authors of Ding. Page 2 of Appendix DING06 contains a citation for "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning," noting that the paper was accepted to be published in Computer Speech & Language and linking to a "demo" associated with some of the material presented in the paper. As the Internet Archive first captured this page as of October 19, 2021, it is my determination that members of the public would have been alerted to the existence of Ding as of October 19,

2021, via the Internet Archive, and likely earlier by visiting Guanlong Zhao's GitHub page.

22. Attached hereto as Appendix DING07 is a screen capture of a webpage entitled "Journal Pre-proof" as hosted on the GitHub website. I secured these screenshots myself from https://guanlongzhao.github.io/media/publication/ding2021csl.pdf.

23. Appendix DING07 is a capture of a pre-proof edition of Ding as posted on Guanlong Zhao's GitHub page. As captured in Appendix DING05, a journal pre-proof is the earliest form of an Article in Press. Also consistent with the language in Appendix DING05, page 1 of Appendix DING07 is a letter from Elsevier exclusive to the pre-proof edition of this paper, stating, "[t]his version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility to the article." Page 1 of Appendix DING07 also reads "© 2021 Elsevier" and includes Ding's DOI, indicating Elsevier's intent to distribute this paper via their website prior to 2022. Aside from page 1 and cosmetic differences, such as text in the header/footer and the "Journal Pre-Proof"

watermark, the pre-proof version of Ding is identical in content to both Exhibit ####'s open manuscript and Appendix DING01's journal publication.

24. Attached hereto as Appendix DING08 is a screen capture of the Internet Archive Wayback Machine entry for https://psi.engr.tamu.edu/portfolio/accent-conversion/. I secured these screen captures myself from http://web.archive.org/web/20211130062820/https://psi.engr.tamu.edu/portfolio/accent-conversion/. This record was captured as of November 30, 2021.

25. Appendix DING08 is an Internet Archive capture of a web page entitled Accent Conversion as hosted on the PSI Lab website. The PSI Lab is a research lab within Texas A&M University's Department of Computer Science and Engineering, and the PSI Lab website is open to the public. Appendix DING07 lists publications associated with PSI Lab's work in the accent conversion space, with "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" being the first result.

26. Appendix DING08 additionally includes expandable Links and BibTeX tabs within the entry for Ding. The Links tab is expanded to show two URLs by which an interested person may access Ding, one of which is Elsevier's ScienceDirect platform (Appendix DING03). A BibTeX record is a type of metadata submitted by the creator upon publication, similar to the metadata on a PDF or Word document. The BibTeX record for Ding is further discussed in relation to Appendices DING09 and DING10. As the Internet Archive first captured this page as of November 30, 2021, it is my determination that this entry in the lab's bibliography indicates that Ding was accessible via the two URLs as of November 30, 2021, via the Internet Archive, and likely earlier via the PSI Lab website.

27. Attached hereto as Appendix DING09 is a screen capture of the Internet Archive Wayback Machine entry for https://psi.engr.tamu.edu/publications/?tgid=18. I secured these screen captures myself from "http://web.archive.org/web/20211130052733/https://psi.engr.tamu.edu/publications/?tgid=18". This record was captured as of November 30, 2021.

28. Appendix DING09 is an Internet Archive capture of a web page entitled "Publications – Accent Conversion" as hosted on the PSI Lab website. This page captures the results of searching Tamu Lab's bibliography by the topic 'Accent Conversion', with "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" being the first result. Appendix DING07 includes the same Links and BibTeX tabs as in Appendix DING08.

29. The BibTeX record in DING09 contains bibliographic data for Ding, including its years of creation ("year = {2021}"), date of availability ("date = {2021-10-14}"), journal of publication ("journal = {Computer Speech & Language}"), and publication status ("pubstate = {forthcoming}"). The BibTex record also lists the same two URLs list in the Links tab. This suggests that Ding was published on this webpage and thus available online as of October 14, 2021. This also suggests that Ding had been sent to ScienceDirect for publication as early as the same date but was not yet published in Computer Speech & Language. As the Internet Archive first captured this page as of November 30, 2021, it is my determination that users of the PSI Lab website searching for 'Accent Conversion' would be able to discover Ding as of October 14, 2021, or at least as of November 30, 2021, via the

Internet Archive, prior to the paper's formal publication in the March 2022 edition of Computer Speech & Language.

30. Attached hereto as Appendix DING10 is a screen capture of the Internet Archive Wayback Machine entry for https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=&auth=. I secured these screen captures myself from http://web.archive.org/web/20211130053613/https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=&auth=. This record was captured as of November 30, 2021.

31. Appendix DING10 is an Internet Archive capture of a web page entitled Publications – Speech as hosted on the PSI Lab website. This page captures the results of searching Tamu Lab's bibliography by the topic 'Speech' with "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" being the first result. Appendix DING10 includes the same Links and BibTeX tabs as in Appendices DING08 and DING09.

32. The BibTeX record in Appendix DING10 is identical to that shown in Appendix DING09. Thus, Appendix DING10 also suggests that Ding was

published on this webpage on October 14, 2021, and was sent to ScienceDirect for publication as early as the same date but was not yet published in Computer Speech & Language. As the Internet Archive first captured this page as of November 30, 2021, it is my determination that users of the PSI Lab website searching for 'Speech' would be able to discover Ding as of October 14, 2021, or at least as of November 30, 2021, via the Internet Archive, prior to the paper's formal publication in the March 2022 edition of Computer Speech & Language.

33. Attached hereto as Appendix DING11 is a screen capture of a web page entitled Publications. I secured these screen captures myself from https://psi.engr.tamu.edu/publications/. This record was captured as of November 14 2025.

34. Appendix DING11 is a screen capture of PSI Lab's most recent publications, at least as of the capture date. Particularly, the list includes two articles from 2024 that are listed as "Forthcoming" (the orange text boxes) and contain BibTeX records that read "pubstate = {forthcoming}." The entries for both articles also include working URL links to the full articles. This confirms that a publication tagged as "Forthcoming" on the PSI LAB website is only

forthcoming with respect to formal publication in their respective journals. That is, the articles themselves are still publicly available via the links displayed on the screen capture, despite not yet being formally published. This further indicates that, despite being listed as Forthcoming in Appendices DING08 through DING10, Ding was only forthcoming with respect to its publication in Computer Speech & Language March 2022. Thus, it is my determination that Ding was publicly available on the PSI Lab web page at the latest as of November 30, 2021, via the Internet Archive.

35. Exhibit #### indicates as of 2021, Ding is published by Elsevier. Appendix DING01 indicates that Penn State University Libraries holds a true and correct copy of "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" by Shaojin Ding, et al., as published in the journal Computer Speech & Language. Appendix DING02 indicates as of May 31, 1997, an e-journal edition Computer Speech & Language is accessible through Penn State University Libraries and this journal was accessible at Penn State University Libraries via the Elsevier Science Direct platform prior to the Ding paper's formal publication in the March 2022 edition of the Computer Speech & Language journal. Appendix DING03 indicates as of October 5, 2021, Ding and its unique DOI URL is publicly

available on the ScienceDirect platform. Appendix DING04 indicates that Articles in Press are publicly available online after being accepted for publication in a journal but prior to that formal publication. Appendix DING05 indicates as of February 6, 2020, journal pre-proofs are publicly available as Articles in Press for Computer Speech & Language to provide early visibility of the articles. Appendix DING06 indicates as of October 19, 2021, the title of Ding and a sample of material relevant to its content is publicly available on co-author Guanlong Zhao's GitHub page. Appendix DING07 indicates as of 2021, an interested user may access Elsevier's published journal pre-proof of Ding through at least co-author Guanlong Zhao's GitHub page. Appendices DING08, DING09, and DING10 indicate as of October 14, 2021, and at least by November 30, 2021, an interested user may access Ding through its publication on the PSI Lab website. Appendix DING11 indicates as of October 14, 2021, and at least by November 30, 2021, Ding was accessible on the PSI Lab website despite its publication in Computer Speech & Language being listed as "Forthcoming." As such, it is my determination that Exhibit ####, "Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning" by Shaojin Ding, et al., as published in Computer Speech & Language, was first made available to the public as early as October 5, 2021, and no later than December

IPR2026-00272
Krisp Technologies EX1010 Page 20

31, 2021, and could be freely accessed either via at least the ScienceDirect website or GitHub.

Sisman

36. I have reviewed Exhibit ####, "Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion" by Berrak Sisman, et al. as published in IEEE/ACM Transactions on Audio, Speech and Language Processing June 2019.

37. Attached hereto as Appendix SISMAN01 is a true and correct copy of "Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion" as published in IEEE/ACM Transactions on Audio, Speech and Language Processing as held by the Penn State University Library. These scans were secured from the library's collection. In comparing Exhibit #### to Appendix SISMAN01, it is my determination that Exhibit #### is a true and correct copy of "Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion" by Berrak Sisman, et al. as published in IEEE/ACM Transactions on Audio, Speech and Language Processing.

38. Attached hereto as Appendix SISMAN02 is a true and correct copy of the MARC record for IEEE/ACM Transactions on Audio, Speech and Language

IPR2026-00272
Krisp Technologies EX1010 Page 21

Processing as held by the Penn State University Library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix SISMAN02 accurately describes the title, publisher and ISSN number of IEEE/ACM Transactions on Audio, Speech and Language Processing. The Showing Results field present on page 3 of this record indicates the library's collection includes the June 2019 issue of IEEE/ACM Transactions on Audio, Speech and Language Processing containing "Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion".

39. The 008 field of the MARC record in Appendix SISMAN02 indicates the date of record creation. The 008 field of Appendix SISMAN02 indicates the Penn State University Library first acquired IEEE/ACM Transactions on Audio, Speech and Language Processing as of July 9, 2013, and held the journal in perpetuity until the publication of "Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion." Considering this information, it is my determination that IEEE/ACM Transactions on Audio, Speech and Language Processing June 2019 and therefore "Group Sparse Representation with WaveNet Vocoder Adaptation

for Spectrum and Prosody Conversion" was made available to the public at least as early as its initial publication in June 2019.

Loots

40. I have reviewed Exhibit 1006, "Automatic Conversion Between Pronunciations of Different English Accents" by Linsen Loots, et al. as published in Speech Communication January 2011.

41. Attached hereto as Appendix LOOTS01 is a true and correct copy of the MARC record for Speech Communication as held by the Penn State University Library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix LOOTS01 accurately describes the title, publisher and ISSN number of Speech Communication. The Showing Results field present on page 3 of this record indicates the library's collection includes the January 2011 issue of Speech Communication containing "Automatic Conversion Between Pronunciations of Different English Accents".

42. Attached hereto as Appendix LOOTS02 is a true and correct copy of "Automatic Conversion Between Pronunciations of Different English

IPR2026-00272
Krisp Technologies EX1010 Page 23

Accents" as published in Speech Communication as held by the Penn State University Library. These scans were secured from the library's collection. In comparing Exhibit 1006 to Appendix LOOTS02, it is my determination that Exhibit 1006 is a true and correct copy of "Automatic Conversion Between Pronunciations of Different English Accents" by Linsen Loots, et al. as published in Speech Communication.

43. The 008 field of the MARC record in Appendix LOOTS01 indicates the date of record creation. The 008 field of Appendix LOOTS01 indicates the Penn State University Library first acquired Speech Communication as of June 10, 1998 and held the journal in perpetuity until the publication of "Automatic Conversion Between Pronunciations of Different English Accents". Considering this information, it is my determination that Speech Communication January 2011 and therefore "Automatic Conversion Between Pronunciations of Different English Accents" was made available to the public at least as early as its initial publication in January 2011.

Osuna

IPR2026-00272
Krisp Technologies EX1010 Page 24

44. I have reviewed Exhibit 1009, “Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion” by Guanlong Zhao, et al. as published in IEEE / ACM Transactions on Audio, Speech and Language Processing October 2019.

45. Attached hereto as Appendix ZHAO01 is a true and correct copy of the MARC record for IEEE / ACM Transactions on Audio, Speech and Language Processing as held by the Penn State University Library. I secured this record myself from the library’s public catalog. The MARC record contained within Appendix ZHAO01 accurately describes the title, publisher and ISSN number of IEEE / ACM Transactions on Audio, Speech and Language Processing. The Showing Results field present on page 3 of this record indicates the library's collection includes the October 2019 issue of IEEE / ACM Transactions on Audio, Speech and Language Processing containing “Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion”.

46. Attached hereto as Appendix ZHAO02 is a true and correct copy of “Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion” as published in IEEE / ACM Transactions on Audio, Speech and

IPR2026-00272
Krisp Technologies EX1010 Page 25

Language Processing as held by the Penn State University Library. These scans were secured from the library's collection. In comparing Exhibit 1009 to Appendix ZHAO02, it is my determination that Exhibit 1009 is a true and correct copy of "Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion" by Guanlong Zhao, et al. as published in IEEE / ACM Transactions on Audio, Speech and Language Processing.

47. The 008 field of the MARC record in Appendix ZHAO01 indicates the date of record creation. The 008 field of Appendix ZHAO01 indicates the Penn State University Library first acquired IEEE / ACM Transactions on Audio, Speech and Language Processing as of July 9, 2013 and held the journal in perpetuity until the publication of "Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion". Considering this information, it is my determination that IEEE / ACM Transactions on Audio, Speech and Language Processing October 2019 and therefore "Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion" was made available to the public at least as early as its initial publication in October 2019.

IPR2026-00272
Krisp Technologies EX1010 Page 26

Cuturi

48. I have reviewed Exhibit ####, "Soft-DTW: a Differentiable Loss Function for Time-Series" by Marco Cuturi, et al as published in Proceedings of Machine Learning Research.

49. Attached hereto as Appendix CUTURI01 is a screen capture of the Internet Archive Wayback Machine entry for http://proceedings.mlr.press/v70. I secured these screen captures myself from https://web.archive.org/web/20170728161753/http://proceedings.mlr.press/v70. This record was captured as of July 28, 2017.

50. Appendix CUTURI01 is an Internet Archive capture of a web page entitled Volume 70: International Conference on Machine Learning, 6-11 August 2017, International Convention Centre, Sydney, Australia as hosted on the Proceedings of Machine Learning Research website. Appendix CUTURI01 is a table of contents for Volume 70 of Proceedings of Machine Learning Research. This page contains an entry for the Cuturi paper on page 6, including a link entitled 'Download PDF' allowing users to download a copy of this paper. As the Internet Archive first captured this page as of July 28,

2017, it is my determination that “Soft-DTW: a Differentiable Loss Function for Time-Series” by Marco Cuturi, et al as published in Proceedings of Machine Learning Research was first made available to the public as of July 28, 2017 via the original Proceedings of Machine Learning Research website.

51. Attached hereto as Appendix CUTURI02 is a screen capture of the Internet Archive Wayback Machine entry for http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf. I secured these screen captures myself from https://web.archive.org/web/20170917200602/http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf. This record was captured as of September 17, 2017.

52. Appendix CUTURI02 is an Internet Archive capture of a web page entitled “Soft-DTW: a Differentiable Loss Function for Time-Series” as hosted on the Proceedings of Machine Learning Research website. I secured this file by clicking the link entitled ‘Download PDF’ as found Appendix CUTURI01. In comparing Appendix CUTURI02 to Exhibit ####, it is my determination that Exhibit #### is a true and correct copy of the “Soft-DTW: a Differentiable Loss Function for Time-Series” file as hosted on the Proceedings of Machine

Learning Research website. As the Internet Archive first captured this page as of September 17, 2017, it is my determination that "Soft-DTW: a Differentiable Loss Function for Time-Series" was first made available to the public as of September 17, 2017 via the Internet Archive and likely earlier via the original Proceedings of Machine Learning Research website.

53. Attached hereto as Appendix CUTURI03 is a screen capture of the Internet Archive Wayback Machine entry for https://arxiv.org/abs/1703.01541. I secured these screen captures myself from https://web.archive.org/web/20190213125708/https://arxiv.org/abs/1703.01541. This record was captured as of February 13, 2019.

54. Appendix CUTURI03 is an Internet Archive capture of a web page entitled Soft-DTW: a Differentiable Loss Function for Time-Series as hosted on the arxiv.org website. Appendix CUTURI03 contains a description of the Cuturi paper, as well as a link entitled 'PDF' allowing users to download a copy of this paper. As the Internet Archive first captured this page as of February 13, 2019, it is my determination that "Soft-DTW: a Differentiable Loss Function for Time-Series" was first made available to the public as of February 13, 2019 via the original arxiv.org website.

55. Attached hereto as Appendix CUTURI04 is a screen capture of the Internet Archive Wayback Machine entry for https://arxiv.org/pdf/1703.01541.pdf. I secured these screen captures myself from https://web.archive.org/web/20211108222341/https://arxiv.org/pdf/1703.01541.pdf. This record was captured as of November 8, 2021.

56. Appendix CUTURI04 is an Internet Archive capture of a web page entitled "Soft-DTW: a Differentiable Loss Function for Time-Series" as hosted on the arxiv.org website. I secured this file by clicking the link entitled 'PDF' as present in Appendix CUTURI03. In comparing Appendix CUTURI04 to Exhibit ####, it is my determination that Exhibit #### is a true and correct copy of the "Soft-DTW: a Differentiable Loss Function for Time-Series" web page as hosted on the arxiv.org website. As the Internet Archive first captured this page as of November 8, 2021, it is my determination that "Soft-DTW: a Differentiable Loss Function for Time-Series" was first made available to the public as of November 8, 2021 via the Internet Archive and likely earlier via the original arxiv.org website.

57. Attached hereto as Appendix CUTURI05 is a true and correct copy of the MARC record for Proceedings of the 2017 International Conference on

IPR2026-00272
Krisp Technologies EX1010 Page 30

Machine Learning as held by the University of Saskatchewan Library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix CUTURI05 accurately describes the title, publisher and ISBN number of Proceedings of the 2017 International Conference on Machine Learning.

58. The 008 field of the MARC record in Appendix CUTURI05 indicates the date of record creation. The 008 field of Appendix CUTURI05 indicates the University of Saskatchewan Library first acquired Proceedings of the 2017 International Conference on Machine Learning as of March 29, 2018. Considering this information, it is my determination that Proceedings of the 2017 International Conference on Machine Learning and therefore "Soft-DTW: a Differentiable Loss Function for Time-Series" was made available to the public at least as early as March 29, 2018.

Kamruzzaman

59. I have reviewed Exhibit ####, "Speaker Identification using MFCC-Domain Support Vector Machine" by S.M. Kamruzzaman, et al as published in International Journal of Electrical and Power Engineering April-May 2007.

60. Attached hereto as Appendix KAMRUZZAMAN01 is a screen capture of the Internet Archive Wayback Machine entry for https://arxiv.org/abs/1009.4972. I secured these screen captures myself from https://web.archive.org/web/20210417200127/https://arxiv.org/abs/1009.4972. This record was captured as of April 17, 2021.

61. Appendix KAMRUZZAMAN01 is an Internet Archive capture of a web page entitled Speaker Identification using MFCC-Domain Support Vector Machine as hosted on the arxiv.org website. Appendix KAMRUZZAMAN01 contains a description of the Kamruzzaman paper, as well as a link entitled 'PDF' allowing users to download a copy of this paper. As the Internet Archive first captured this page as of April 17, 2021, it is my determination "Speaker Identification using MFCC-Domain Support Vector Machine" was first made available to the public as of April 17, 2021 via the original arxiv.org website.

62. Attached hereto as Appendix KAMRUZZAMAN02 is a screen capture of the Internet Archive Wayback Machine entry for https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf. I secured these screen captures myself from

https://web.archive.org/web/20200825105542/https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf. This record was captured as of August 25, 2020.

63. Appendix KAMRUZZAMAN02 is an Internet Archive capture of a web page entitled "Speaker Identification using MFCC-Domain Support Vector Machine" as hosted on the arxiv.org website. I secured this file by clicking the link entitled 'PDF' as present in Appendix KAMRUZZAMAN01. In comparing Appendix KAMRUZZAMAN02 to Exhibit ####, it is my determination that Exhibit #### is a true and correct copy of the "Speaker Identification using MFCC-Domain Support Vector Machine" web page as hosted on the arxiv.org website. As the Internet Archive first captured this page as of August 25, 2020, it is my determination that "Speaker Identification using MFCC-Domain Support Vector Machine" was first made available to the public as of August 25, 2020 via the Internet Archive and likely earlier via the original arxiv.org website.

64. Attached hereto as Appendix KAMRUZZAMAN03 is a true and correct copy of the MARC record for International Journal of Electrical and Power Engineering as held by the Glasgow Caledonian University Library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix KAMRUZZAMAN03 accurately describes the

title, publisher and ISSN number of International Journal of Electrical and Power Engineering. The Full Text Availability field present on page 2 of this record indicates the library's collection includes the April-May 2007 issue of International Journal of Electrical and Power Engineering containing "Speaker Identification using MFCC-Domain Support Vector Machine".

65. The 008 field of the MARC record in Appendix KAMRUZZAMAN03 indicates the date of record creation. The 008 field of Appendix KAMRUZZAMAN03 indicates the Glasgow Caledonian University Library first acquired International Journal of Electrical and Power Engineering as of September 20, 2007 and the collection details indicate this acquisition includes all issues of the journal from January 2007 forward. Considering this information, it is my determination that International Journal of Electrical and Power Engineering April-May 2007 and therefore "Speaker Identification using MFCC-Domain Support Vector Machine" was made available to the public via the Glasgow Caledonian University Library at least as early as September 20, 2007.

Heckmann

IPR2026-00272
Krisp Technologies EX1010 Page 34

66. I have reviewed Exhibit ####, "A Hierarchical Framework for Spectro-Temporal Feature Extraction" by Martin Ernst Heckmann, et al as published in Speech Communication May 2011.

67. Attached hereto as Appendix HECKMANN01 is a true and correct copy of the MARC record for Speech Communication as held by the Penn State University Library. I secured this record myself from the library's public catalog. The MARC record contained within Appendix HECKMANN01 accurately describes the title, publisher and ISSN number of Speech Communication. The Showing Results field present on page 3 of this record indicates the library's collection includes the May 2011 issue of Speech Communication containing "A Hierarchical Framework for Spectro-Temporal Feature Extraction".

68. Attached hereto as Appendix HECKMANN02 is a true and correct copy of "A Hierarchical Framework for Spectro-Temporal Feature Extraction" as published in Speech Communication as held by the Penn State University Library. These scans were secured from the library's collection. In comparing Exhibit #### to Appendix HECKMANN02, it is my determination that Exhibit #### is a true and correct copy of "A

IPR2026-00272
Krisp Technologies EX1010 Page 35

Hierarchical Framework for Spectro-Temporal Feature Extraction" by Martin Ernst Heckmann, et al as published in Speech Communication.

69. The 008 field of the MARC record in Appendix HECKMANN01 indicates the date of record creation. The 008 field of Appendix HECKMANN01 indicates the Penn State University Library first acquired Speech Communication as of June 10, 1998 and held the journal in perpetuity until the publication of "A Hierarchical Framework for Spectro-Temporal Feature Extraction". Considering this information, it is my determination that Speech Communication May 2011 and therefore "A Hierarchical Framework for Spectro-Temporal Feature Extraction" was made available to the public at least as early as its initial publication in May 2011.

70. I have been retained on behalf of the Petitioner to provide assistance in the above-illustrated matter in establishing the authenticity and public availability of the documents discussed in this declaration. I am being compensated for my services in this matter at the rate of $200.00 per hour plus reasonable expenses. My statements are objective, and my compensation does not depend on the outcome of this matter.

71. I declare under penalty of perjury that the foregoing is true and correct. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: 3/18/2026



June Ann Munford

Appendix CV

June A. Munford
Curriculum Vitae

Education

University of Wisconsin-Milwaukee - MS, Library & Information Science, 2009
Milwaukee, WI

- Coursework included cataloging, metadata, data analysis, library systems, management strategies and collection development.
- Specialized in library advocacy, cataloging and public administration.

Grand Valley State University - BA, English Language & Literature, 2008
Allendale, MI

- Coursework included linguistics, documentation and literary analysis.
- Minor in political science with a focus in local-level economics and government.

Professional Experience

Researcher / Expert Witness, October 2017 – present
Freelance ● Pittsburgh, Pennsylvania & Grand Rapids, Michigan

- Material authentication and public accessibility determination. Declarations of authenticity and/or public accessibility provided upon research completion. Experienced with appeals and deposition process.

- Research provided on topics of public library operations, material publication history, digital database services and legacy web resources.

- Past clients include Alston & Bird, Arnold & Porter, Baker Botts, Fish & Richardson, Erise IP, Irell & Manella, O'Melveny & Myers, Perkins-Coie, Pillsbury Winthrop Shaw Pittman and Slayden Grubert Beard.

Library Director, February 2013 - March 2015
Dowagiac District Library ● Dowagiac, Michigan

- Executive administrator of the Dowagiac District Library. Located in Southwest Michigan, this library has a service area of 13,000, an annual operating budget of over $400,000 and total assets of approximately $1,300,000.

● Developed careful budgeting guidelines to produce a 15% surplus during the 2013-2014 & 2014-2015 fiscal years while being audited.

● Using this budget surplus, oversaw significant library investments including the purchase of property for a future building site, demolition of existing buildings and building renovation projects on the current facility.

● Led the organization and digitization of the library's archival records.

● Served as the public representative for the library, developing business relationships with local school, museum and tribal government entities.

● Developed an objective-based analysis system for measuring library services - including a full collection analysis of the library's 50,000+ circulating items and their records.

November 2010 - January 2013
Librarian & Branch Manager, Anchorage Public Library ● Anchorage, Alaska

● Headed the 2013 Anchorage Reads community reading campaign including event planning, staging public performances and creating marketing materials for mass distribution.

● Co-led the social media department of the library's marketing team, drafting social media guidelines, creating original content and instituting long-term planning via content calendars.

● Developed business relationships with The Boys & Girls Club, Anchorage School District and the US Army to establish summer reading programs for children.

June 2004 - September 2005, September 2006 - October 2013
Library Assistant, Hart Area Public Library
Hart, MI

● Responsible for verifying imported MARC records and original MARC cataloging for the local-level collection as well as the Michigan Electronic Library.

● Handled OCLC Worldcat interlibrary loan requests & fulfillment via ongoing communication with lending libraries.

Professional Involvement

Alaska Library Association - Anchorage Chapter

● Treasurer, 2012

Library Of Michigan
● Level VII Certification, 2008
● Level II Certification, 2013

Michigan Library Association Annual Conference 2014
● New Directors Conference Panel Member

Southwest Michigan Library Cooperative
● Represented the Dowagiac District Library, 2013-2015

Professional Development

Library Of Michigan Beginning Workshop, May 2008
Petoskey, MI
● Received training in cataloging, local history, collection management, children's literacy and reference service.

Public Library Association Intensive Library Management Training, October 2011
Nashville, TN
● Attended a five-day workshop focused on strategic planning, staff management, statistical analysis, collections and cataloging theory.

Alaska Library Association Annual Conference 2012 - Fairbanks, February 2012
Fairbanks, AK
● Attended seminars on EBSCO advanced search methods, budgeting, cataloging, database usage and marketing.

Depositions

2019 ● Fish & Richardson

Apple v. Qualcomm (Case No. IPR2018-001281, 39521-00421IP, IPR2018-01282 and 39521-00421IP2)

2019 ● Erise IP

Implicit, LLC v. Netscout Systems, Inc (Case No. 2:18-cv-53-JRG)

2019 ● Perkins-Coie

Adobe Inc. v. RAH Color Technologies LLC (Case No. IPR2019-00627, IPR2019-00628, IPR2019-00629 and IPR2019-00646)

2020 ● O'Melveny & Myers

Maxell, Ltd. v. Apple Inc. (Case No. 5:19-cv-00036-RWS)

2021 ● Pillsbury Winthrop Shaw Pittman LLP

Intel v. SRC (Case No. IPR2020-1449)

2022 ● Perkins-Coie

Realtek v. Future Link (Case No. IPR2021-01182)

2023 ● Fish & Richardson

Neuroderm Ltd. v. Abbvie, Inc (Case No. PGR2022-00040)

2023 ● Fish & Richardson

Nearmap US Inc. v. Pictometry International Corp. (Case No. IPR2022-00735)

2023 ● Fish & Richardson

Samsung Electronics v. MemoryWeb LLC (Case No. 39843-0136PS1)

2023 ● Pillsbury Winthrop Shaw Whitman LLP

Gravel Rating Systems v. Costco Wholesale Corp. (Case No. 4:21-cv-149-ALM)

2024 ● Willkie-Farr

Netflix, Inc. v. VideoLabs. Inc. (Case No. IPR2023-00628)

2024 ● Quinn Emanuel Urquhart & Sullivan

Purdue University v. Wolfspeed (Case No. 1:21-CV-840)

2024 ● Fish & Richardson

Dish Network v. Entropic Communications (Case No. IPR2024-00393)

2025 ● Perkins Coie

Amazon.com, Inc. v. Nokia Technologies Oy (Case No. IPR2024-00572)

2025 ● Fish & Richardson

Dish Network v. Entropic Communications (Case No. IPR2024-00373)

Limited Case and Clientele History

Alston & Bird

- Ericsson

  v. Collision Communications (Case No. IPR2022-01233)

- Nokia

  v. Neptune Subsea, Xtera (Case No. 1:17-cv-01876)

- Universal Electronics Inc

  v. Roku Inc (Case No. IPR2022-00818)

Alavi & Anaipakos PLLC

- Stingray Group Inc. and Stingray Music USA, Inc.

  v. Edwin A. Hernandez-Mondragon and Egla Corp. (Case No. 1:24-cv-21226-RAR)

Arnold & Porter

- Ivantis

  v. Glaukos (Case No. 8:18-cv-00620)

- Samsung

  v. Jawbone (Case No. 2:21-cv-00186)

Benesch Friedlander Coplan & Aronoff

- Voyis

  v. Cathx (Case No. 5:21-cv-00077-RWS)

Buchanan, Ingersoll & Rooney PC

- Google LLC, AT&T Service Inc., T Mobile USA Inc., Cellco Partnership, Ericsson Inc. & Nokia of America Corp.

  v. KT Corp & Pegasus Wireless Innovation LLC (Case No. IPR2025-00293)

Deschert LLP

- Smaxtec Animal Care

  v. ST Reproductive Technologies, LLC (Case No. IPR2024-00885)

Duane Morris

- Kangxi Communication Technologies

  v. Skyworks Solutions Inc. (Case No. IPR2025-00373)

- Microsoft

  v. Edge Networking Systems LLC (Case No. 1:24-cv-00215)

Erise I.P.

- Apple

  v. Ericsson Inc. (Case No. IPR2022-00715)

  v. Future Link Systems (Case No. IPRs 6317804, 6622108, 6807505, and 7917680)

  v. INVT (Case No. 20-1881)

  v. Navblazer LLC (Case No. IPR2020-01253)

  v. Qualcomm (Case No. IPR2018-001281, 39521-00421IP, IPR2018-01282, 39521-00421IP2)

v. Quest Nettech Corp (Case No. 2:19-cv-00118-JRG)

v. Telefonaktiebolaget LM Ericsson (Case No. IPR2022-00275)

v. Theta IP, LLC (Case No. IPR2024-00818)

v. THL Holding Company LLC (Case No. 1:23-cv-00548)

- Fanduel

  v. CGT (Case No. 19-1393)

- Garmin

  v. Phillips North America LLC (Case No. 2:19-cv-6301-AB-KS)

- Netscout

  v. Longhorn HD LLC (Case No. 2:20-cv-00349)

  v. Implicit, LLC (Case No. 2:18-cv-53-JRG)

- Sony Interactive Entertainment LLC

  v. Bot M8 LLC (Case No. IPR2020-01288)

  v. Infernal Technology LLC (Case No. 2:19-CV-00248-JRG)

- Tesla

  v. Charge Fusion Technologies LLC (Case No. IPR2025-00032)

- Unified Patents

  v. GE Video Compression (Case No. 2:19-cv-248)

Fish & Richardson

- Apple

  v. AliveCor (Case No. 3:21-cv-03958)

  v. LBS Innovations (Case No. 2:19-cv-00119-JRG-RSP)

  v. Koss Corporation (Case No. IPR2021-00305)

v. Masimo (Case No. IPR 50095-0012IP1, 50095-0012IP2, 50095-0013IP1, 50095-0013IP2, 50095-0006IP1, 50095-0135IP1)

v. Neonode (Case No. 21-cv-08872-EMC)

v. Qualcomm (Case No. IPR2018-001281, 39521-00421IP, IPR2018-01282, 39521-00421IP2)

v. Resonant Sys. (Case No. 7:23-cv-00077-ADA)

- AsusTek Computer Inc.
  v. Videolabs, Inc. (Case No. 22-CV-00720-ADA)

- Dell
  v. Neo Wireless (Case No. IPR2022-00616)

- Dish Network

  v. Entropic Communications, LLC (Case No. 2:2023-CV-01043)

  v. Entropic Communications LLC (Case No. IPR2024-00393)

  v. Realtime Adaptive Streaming (Case No. 1:17-CV-02097-RBJ)

  v. TQ Delta LLC (Case No. 18-1798)

- Evapco Dry Cooling

  v. SPG Dry Cooling (Case No. IPR2021-00688)

- Genetec

  v. Sensormatic Electronics (Case No. 1:20-CV-00760)

- Huawei

  v. Bell Northern Research LLC (Case No. IPR2019-01174)

- Kianxis

  v. Blue Yonder (Case No. 3:20-cv-03636)

- LG Electronics

  v. Bell Northern Research LLC (Case No. 3:18-cv-2864-CAB-BLM)

- MediaTek

  v. MOSAID (Case No. IPR2024-00718)

- Metaswitch

  v. Sonus Networks (Case No. IPR2018-01719)

- Microsoft

  v. Throughputer Inc (Case No. IPR2022-00757)

- Mom Enterprises

  v. Ddrops Company (Case No. 1:22-cv-00332)

- MLC Intellectual Property

  v. MicronTech (Case No. 3:14-cv-03657-SI)

- Nearmap Inc

  v. EagleView Technologies (Case No. IPR2022-01009)

- Neuroderm Ltd.

  v. Abbvie, Inc (Case No. PGR2022-00040)

- Posco Co Ltd.

  v. Pascal Drillet et al. (Case No. IPR2025-00370)

- Realtek Semiconductor

  v. Future Link (Case No. IPR2021-01182)

- Quectel
  - v. Koninklijke Philips (Case No. 1:20-cv-01710)

- Samsung
  - v. Aire Technology (Case No. IPR2022-00877)
  - v. Bell Northern Research (Case No. 2:19-cv-00286-JRG)
  - v. Communication Technologies Inc (Case No. IPR2022-01221)
  - v. Jawbone Innovations (Case No. IPR2022-00865)
  - v. MemoryWeb LLC (Case No. IPR2022-00885)
  - v. Dodots Licensing Solutions (Case No. IPR2023-00701)
  - v. Telefonaktiebolaget LM Ericsson (Case No. IPR2021-00615)

- Texas Instruments
  - v. Vantage Micro LLC (Case No. IPR2020-01261)

- Xilinx
  - v. Sentient Sensors LCC (Case No. 1:22-cv-00173)

Goodwin Proctor

- Intel
  - v. Proxense, LLC (Case No. 6:24-cv-00283)

- Samsung Electronics
  - v. Mobile Data Technologies (Case No. 2:24-cv-00435)

- Taiwan Semiconductor Manufacturing Company Ltd.
  - v. Advanced Integrated Circuit Process LLC (Case No. 2:24-cv-00623)

Groombridge Wu

- Nearmap

  v. Eagle View Technologies Inc. and Pictometry International Corp. (Case No. 2:21-cv-00283-TS-DAO)

Irell & Manella

- Curium

Latham & Watkins LLP

- Apple

  v. Advanced Coding Research LLC

Leydig, Voit & Mayer, Ltd.

- Public Availability Consultancy

O'Melveny & Myers

- Apple

  v. Maxell (Case No. 5:19-cv-00036-RWS)

- Disney Media & Entertainment Distribution LLC

  v. Digital Media Technology Holdings, LLC (Case No. 2:22-cv-01642)

- Micron Technology Inc.

  v. Besang Inc (Case No. IPR2023-00900)

- Samsung

  v. Daedalus Prime LLC (1335 Investigation)

Perkins-Coie

- Amazon
  - v. Nokia (Case No. IPR2024-00847 and IPR2024-00848)
  - v. Nokia Technologies Oy (Case No. IPR2024-00572)
- Heru Industries
  - v. The UAB Research Foundation (Case No. IPR2022-01148)
- Intel Corporation
  - v. BeSang Inc. (Case No. IPR2023-00991)
  - v. BitMICRO (Case No. 5:23-cv-00625)
  - v. Telefonaktiebolaget L M Ericsson (Case No. IPR2024-00610)
- Realtek Semiconductor
  - v. Future Link (Case No. IPR2021-01182)
- r-pac
  - v. Adasa, Inc. (Case No. 1:2024cv06102)
- Twitter Inc
  - v. VOIP-Pal.com (Case No. 3:20-cv-02397-JD)
- TCL Industries
  - v. Koninklijke Philips NV (Case No. IPR2021-00495, IPR2021-00496 and IPR2021-00497)
- VusionGroup
  - v. Hanshow (Case No. IPR2024-00857)

Pillsbury Winthrop Shaw Pittman

- Intel

  v. FG SRC LLC (Case No. 6:20-cv-00315)

- Gravel Rating Systems

  v. Costco (Case No. 4:21-cv-149)
  v. Lowe's Home Centers (Case No. 4:21-cv-150)
  v. T-Mobile USA (Case No. 4:21-cv-152)
  v. Kohl's Inc. (Case No. 4:21-cv-258)
  v. Under Armor (Case No. 4:21-cv-356)

Quinn Emanuel

- Exact Sciences Corporation

  v. Geneoscopy, Inc. (Case No. IPR2024-00459 and IPR2024-1330)

- ServiceNow, Inc.

  v. InQuisient, Inc. (Case No. 22-900-CJB)

- Wolfspeed

  v. The Trustees of Purdue University (Case No. 1:2021-cv-00840)

Sheppard, Mullin, Richter & Hampton LLP

- Advanced Micro Devices, Inc and Pensando Systems, Inc.

  v. Concurrent Ventures LLC and Xtreamedge, Inc. (Case No. IPR2025-00478)

- Cadence Design Systems Inc.

  v. Semiconductor Design Technologies (Case No. 3:23-cv-01001)

Shook, Hardy & Bacon LLP

- HP Inc.

  v. Universal Connectivity Technologies (Case No. 1:23-cv-01177)

Stern, Kessler, Goldstein & Fox

- Preliminary Research

Wilmer, Cutler, Pickering, Hale and Dorr

- Apple

  v. Shunock (Case No. 1:23-cv-08598)

- Quest Diagnostics Inc., Haystack Oncology Inc. and Haystack Oncology GmbH

  v. Natera, Inc. (Case No. 1:23-cv-08598)

- Roche Diabetes, Inc.

  v. Trividia Health, Inc. (Case No. 1:24-cv-00668)

Willkie, Farr & Gallagher LLP

- Lenovo, Dell, HP

  v. Universal Connectivity Technologies Inc. (Case No. 2:2023cv00449)

- Neurent

  v. The Foundry, LLC (Case No. IPR2024-00669)

- Netflix, Inc.

  v. VideoLabs, Inc. (Case No. IPR2023-00628)

Appendix DING01

Computer Speech and Language 72 (2022) 101302

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning☆

Shaojin Ding *, Guanlong Zhao[1], Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, USA

ARTICLE INFO

Keywords:
Foreign accent conversion
Speech synthesis
Computer-assisted pronunciation training

ABSTRACT

Foreign accent conversion (FAC) aims to create a new voice that has the *voice identity* of a given second-language (L2) speaker but with a native (L1) *accent*. Previous FAC approaches usually require training a separate model for each L2 speaker and, more importantly, generally require considerable speech data from each L2 speaker for training. To address these limitations, we propose Accentron, an approach that can generate accent-converted speech for arbitrary L2 speakers unseen during training. In the proposed approach, we first train a speaker-independent acoustic model on L1 corpora to extract bottleneck features that represent the linguistic content of utterances. Then, we develop a speaker encoder and an accent encoder to generate embedding vectors for the desired voice identity (L2 speaker's) and accent (L1 accent), respectively. Lastly, we use a sequence-to-sequence model to transform bottleneck-features to Mel-spectrograms, conditioned on the L2 speaker embedding and the L1 accent embedding. We conducted experiments on the L2-ARCTIC corpus under two testing conditions: the *standard* FAC setting where test L2 speakers were seen during training, and a *zero-shot* FAC setting where test L2 speakers were unseen during training. Accentron achieves over 27% relative improvement in accentedness ratings compared to two state-of-the-art FAC systems in the standard FAC setting. More importantly, our results show that Accentron generalizes to the zero-shot FAC setting with no performance loss. Therefore, in practical use scenarios (e.g., computer-assisted pronunciation training software), Accentron can effectively avoid the need to adapt or retrain the model, which significantly reduces computations and the users' waiting time.

1. Introduction

Foreign accent conversion (FAC) (Felps et al., 2009) aims to create a new voice that has the voice identity of a given L2 speaker and the accent of an L1 speaker. In pronunciation training, FAC can serve as a "golden speaker" for the L2 speaker to practice with: their own voice, but with a native accent (Felps et al., 2009; Probst et al., 2002; Ding et al., 2019a; Wang and Lu, 2011). FAC also finds applications in movie dubbing (Turk and Arslan, 2002), personalized text-to-speech (TTS) synthesis (Sun et al., 2016b; Oshima et al., 2015), and improving speech recognition performance (Biadsy et al., 2019). A variety of techniques have been proposed to perform FAC, including voice morphing (Felps et al., 2009; Aryal et al., 2013; Huckvale and Yanagisawa, 2007), frame pairing (Aryal and Gutierrez-Osuna, 2014; Zhao and Gutierrez-Osuna, 2019), articulatory synthesis (Aryal and Gutierrez-Osuna, 2015a,b), and sequence-to-sequence (seq2seq) modeling (Zhao et al., 2019; Liu et al., 2020). However, previous FAC approaches

☆ This work was supported by NSF awards 1619212 and 1623750.
* Corresponding author.
E-mail addresses: shjd@tamu.edu (S. Ding), gzhao@tamu.edu (G. Zhao), rgutier@cse.tamu.edu (R. Gutierrez-Osuna).
[1] Ding and Zhao are now with Google. This work was done solely with TAMU resources.

https://doi.org/10.1016/j.csl.2021.101302
Received 12 April 2021; Received in revised form 19 September 2021; Accepted 21 September 2021
Available online 5 October 2021
0885-2308/© 2021 Elsevier Ltd. All rights reserved.

IPR2026-00272
Krisp Technologies EX1010 Page 54

have two major limitations. First, they operate in a one-to-one fashion, i.e., they require training a separate model for each pair of L1 and L2 speakers. Second, they need a considerable amount of speech data (~1000 utterances) for each L2 speaker. Thus, when using these conventional FAC methods in real-world applications such as pronunciation training, L2 learners need to record a large number of utterances and then wait for a dedicated model to be trained, which can be tedious and demotivating.

To address this issue, we propose Accentron, a zero-shot learning (Lampert et al., 2009) approach to FAC that can synthesize speech for arbitrary L2 speakers who were unseen during training. Accentron consists of four independent models: (1) a speaker-independent acoustic model that captures the linguistic content of an L1 utterance as a sequence of *bottleneck feature vectors*, (2) a speaker encoder that captures the voice identity of the L2 speaker, denoted as a *speaker embedding*, (3) an accent encoder that captures the desired L1 accent, denoted as an *accent embedding*, and (4) a seq2seq model that generates a Mel-spectrogram from the sequence of bottleneck features, conditioned on the desired speaker and accent embeddings. These components can be trained independently, at which point the system can generate accent conversions to arbitrary L2 speakers given a few seconds of audio (i.e., enough speech to compute a speaker embedding), without the need to have any model re-training or adaptation process.

To our knowledge, ours is the first work to apply zero-shot learning for the task of FAC. Though zero-shot learning has been used for voice conversion (Liu et al., 2018; Hsu et al., 2016; Kameoka et al., 2018; Sun et al., 2016a) and voice cloning (Jia et al., 2018; Arık et al., 2018), previous studies (Liu et al., 2018; Hsu et al., 2016; Kameoka et al., 2018; Sun et al., 2016a; Jia et al., 2018; Arık et al., 2018) have focused exclusively on manipulating voice identity, ignoring the speaker's accent, which holds important cues to speaker recognition (Das et al., 2020) and speech perception (Leikin et al., 2009; Major et al., 2002; Van Heugten et al., 2015; Cristia et al., 2012). Incorporating accent into the conversion process requires changes to the conventional encoder–decoder structure of sequence-to-sequence (seq2seq) models for voice conversion. Our encoder takes a sequence of L1 bottleneck feature vectors as the input, and produces a hidden representation sequence. In a conventional voice conversion system (Liu et al., 2018; Hsu et al., 2016; Ding and Gutierrez-Osuna, 2019; Saito et al., 2018; Hsu et al., 2017b; Mohammadi and Kim, 2019; Lu et al., 2019; Ding et al., 2020), this hidden representation sequence is then concatenated with the speaker embedding of the target speaker. In our case, however, the system also concatenates the accent embedding, which is treated as an additional independent and controllable factor during synthesis. The combined bottleneck/speaker/accent embedding is consumed by a decoder coupled with a location-sensitive attention mechanism (Chorowski et al., 2015). During each decoding step, the decoder autoregressively predicts a Mel-spectrogram frame based on the output from the previous decoding step and a context vector produced by the attention mechanism. Finally, the output Mel-spectrogram is converted back into a waveform through either the Griffin–Lim algorithm (Griffin and Lim, 1984) or a separately trained vocoder (e.g., WaveNet (Tamamori et al., 2017), WaveRNN (Kalchbrenner et al., 2018)).

We thoroughly evaluated Accentron on the L2-ARCTIC corpus (Zhao et al., 2018). First, we visualized the speaker and accent embedding distributions for the accent-converted speech and natural speech, and the results show that Accentron syntheses can successfully capture the L2 voice identity along with an L1 accent. Second, we conducted a series of listening tests under two different settings: (1) a *standard* FAC setting, where the test L2 speakers were available during training, and (2) a *zero-shot* FAC setting, which assumes that the test L2 speakers were not available during training. Accentron achieves 27% relative improvement in accentedness while retaining the acoustic quality and voice identity, compared to two state-of-the-art FAC systems in standard FAC settings. In addition, Accentron showed no performance degradation when tested under zero-shot FAC setting.

The manuscript is organized as follows. Section 2 reviews prior approaches to foreign accent conversion, many-to-many voice conversion, and sequence-to-sequence models. Section 3 describes the architecture of Accentron in detail. Section 4 provides the experimental setup, including the corpora and implementation details. Section 5 presents the objective and subjective evaluations of Accentron and analyzes the results. We discuss the implications of the results in Section 6. Lastly, we conclude the findings of this work and point out potential future directions in Section 7.

2. Related work

2.1. Foreign accent conversion

The problem of foreign accent conversion was first formulated by Felps et al. (2009) as the means to provide implicit feedback in computer assisted pronunciation training. Early approaches (Aryal and Gutierrez-Osuna, 2015b; Felps et al., 2012; Aryal and Gutierrez-Osuna, 2016; Brand, 1999) involved building an articulatory synthesizer for the L2 speaker. The articulatory synthesizer was trained to map the L2 speaker's articulatory trajectories (e.g., tongue and lip movements) into his or her acoustics features (e.g., Mel Cepstra) using GMMs (Aryal and Gutierrez-Osuna, 2015b), unit-selection models (Felps et al., 2012), and DNNs (Aryal and Gutierrez-Osuna, 2016). Once the synthesizer was built, it could be driven with articulatory trajectories from an L1 speaker to synthesize FAC speech. However, these approaches were impractical for pronunciation training since collecting articulatory data is expensive and requires specialized equipment.[2] As a result, later work on FAC has focused on acoustic methods, since they only require recording speech with a microphone. Previous acoustic methods can be grouped into two categories: frame-pairing methods (Aryal and Gutierrez-Osuna, 2014; Zhao and Gutierrez-Osuna, 2019) and seq2seq methods (Zhao et al., 2019; Liu et al., 2020). As the name suggests, frame-pairing methods builds a lookup table where L1 and L2 speech frames are paired based on their similarity, and then use a statistical model (e.g., a GMM) to convert from L1 frames to their corresponding L2 frames in the

[2] Articulatory measurements can be performed via electromagnetic articulography (Felps et al., 2012), ultrasound imaging (Denby and Stone, 2004), palatography (Mumtaz et al., 2014), and more recently real-time MRI (Toutios et al., 2016).

IPR2026-00272
Krisp Technologies EX1010 Page 55

lookup table. Aryal and Gutierrez-Osuna (2014) first proposed a technique to pair L1–L2 frames based on their acoustic similarity (in MFCC space), after applying vocal tract length normalization to reduce global differences between the L1 and L2 spectra. Following this, Zhao and Gutierrez-Osuna (2019) argued that the L1 and L2 frames should be paired based on their linguistic content, and consequently, they used Phonetic-PosteriorGram (PPG) similarity instead of MFCC similarity. More recently, methods based on seq2seq models have been shown to significantly improve synthesis quality. In a previous study (Zhao et al., 2019), we proposed a seq2seq PPG-to-Mel synthesizer for FAC. During training, the system learns a seq2seq model to convert PPGs to Mel-spectra extracted from utterances of an L2 speaker. During inference, the model is driven by PPGs extracted from a reference L1 utterance, which then produces FAC synthesis. In related work, Liu et al. (2020) proposed a novel recognizer–synthesizer framework to remove the need for a reference L1 utterance. Their system trained a speaker recognizer, a multi-speaker text-to-speech (TTS) model, and an accent-sensitive automatic speech recognition (ASR) system. During inference, they feed L2 Mel-spectra to the ASR system with the corresponding accent, and then feed the output of the ASR system and the L2 speaker embedding to the multi-speaker TTS model to generate accent-converted utterances. These seq2seq model based FAC approaches can convert segmental and prosody features simultaneously, producing syntheses with higher speech naturalness and acoustic quality.

2.2. Many-to-many voice conversion

Foreign accent conversion is related to the more general problem of voice conversion (VC) (Mohammadi and Kain, 2017; Sisman et al., 2020), which aims to synthesize a voice that has the linguistic content of an utterance from a source speaker and the voice identity of a target speaker. Traditional VC approaches use GMMs (Toda et al., 2007; Erro et al., 2009), sparse representations (Takashima et al., 2012; Ding et al., 2019b), and DNNs (Chen et al., 2014; Desai et al., 2010; Nakashika et al., 2016; Xie et al., 2016; Miyoshi et al., 2017; Kaneko et al., 2019; Sun et al., 2016a; Zhou et al., 2019) to transform the spectra from a source speaker to that of the target speaker. These methods require training a separate model for each pair of source-target speakers. More recently, several studies have proposed many-to-many VC approaches based on Variational Autoencoders (VAE) (Hsu et al., 2016; Ding and Gutierrez-Osuna, 2019; Saito et al., 2018; Hsu et al., 2017b; Kameoka et al., 2018; Huang et al., 2019) and the PPG-to-speech synthesizer (Liu et al., 2018; Mohammadi and Kim, 2019; Lu et al., 2019; Kaneko et al., 2017; Zhang et al., 2019). Hsu et al. (2016, 2017a) first proposed to use a VAE for many-to-many VC. During training, the encoder learns a speaker-independent latent embedding from input speech signals, and the decoder reconstructs the input speech signals given the latent embedding and the corresponding speaker embedding. During inference, the speaker embedding is replaced with that of a target speaker to produce a VC synthesis. A number of subsequent studies have been conducted to improve performance through various techniques, such as auxiliary classifiers (Kameoka et al., 2018), WaveNet vocoder adaption (Huang et al., 2019), and discrete latent space (Ding and Gutierrez-Osuna, 2019; Van Den Oord et al., 2017). Other studies (Liu et al., 2018; Mohammadi and Kim, 2019; Lu et al., 2019) have used a PPG-to-speech synthesizer approach to perform many-to-many VC. The PPG-to-speech synthesizer is a neural network that takes PPGs as an input, and predict spectra conditioned on the speaker embedding of the target speaker. Early many-to-many VC models used one-hot vectors as the speaker embedding due to its simplicity, but recent studies (Saito et al., 2018; Liu et al., 2018; Mohammadi and Kim, 2019; Lu et al., 2019) have used learned speaker embeddings (e.g., i-vector Dehak et al., 2010, d-vector Variani et al., 2014) to generalize to unseen speakers, which make it possible to perform VC in a zero-shot fashion.

2.3. Seq2seq models

The seq2seq model was originally proposed by Sutskever et al. (2014) for machine translation. The seq2seq model usually has an encoder–decoder architecture. The encoder learns a hidden representation sequence from an input sequence, and the decoder learns to autoregressively generate the output sequence given the hidden representation. To capture contextual information and handle length mismatches between the input and output sequences, an attention mechanism (Bahdanau et al., 2015) is added between the encoder and the decoder. In recent years, there has been growing interest in applying seq2seq model to speech synthesis. Wang et al. (2017) first proposed a seq2seq based TTS synthesizer (Tacotron), which significantly improved the acoustic quality of the syntheses over previous methods. Following this, Shen et al. (2018) proposed Tacotron2, which further improved the acoustic quality of Tacotron by using a novel model architecture and a WaveNet vocoder. Jia et al. (2018) extended Tacotron2 to voice cloning by conditioning a speaker embedding on the decoder. Besides, Biadsy et al. (2019) and Jia et al. (2019) also explored the use of seq2seq model in end-to-end speech-to-speech transformation, which avoids the need of multi-stage recognizer–synthesizer framework in related tasks, such as hearing-impaired speech synthesis and speech-to-speech translation. Seq2seq model has also been applied to voice conversion (Miyoshi et al., 2017; Kaneko et al., 2017; Zhang et al., 2019) and foreign accent conversion (Zhao et al., 2019; Liu et al., 2020), which significantly improved the performance on these tasks compared to conventional approaches.

3. Methods

Accentron consists of four modules: (1) a speaker-independent acoustic model that generates a linguistic representation of an utterance, (2) a speaker encoder that captures the voice identity of the desired speaker, (3) an accent encoder that captures the desired accent, and (4) a seq2seq model that consumes the previous three representations to synthesize Mel-spectrogram for an arbitrary L2 speaker.

The workflow for training Accentron is shown in Fig. 1(a). The acoustic model, speaker encoder, and accent encoder are trained separately, and then are used as feature extractors for the seq2seq model. The seq2seq model is trained on a parallel corpus with

IPR2026-00272
Krisp Technologies EX1010 Page 56

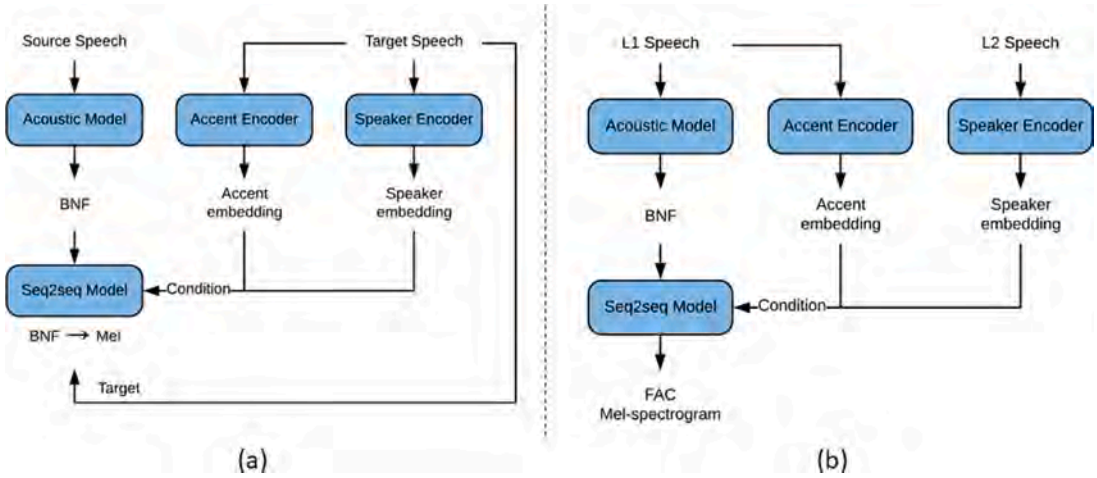


Fig. 1. (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

multiple L1 and L2 speakers, capturing the voice characteristics of different speakers and accents. In what follows, we define a *"source"* speaker to be a selected reference L1 speaker, and a *"target"* speaker to be any L1/L2 speaker. To train the seq2seq model, we pair the source speaker with each target speaker. Then, for each pair of speakers, we feed source utterances to the speaker-independent acoustic model to extract bottleneck features (BNFs), which we assume capture only the linguistic content. Next, we feed an utterance from the target speaker to the speaker encoder and the accent encoder, which extract their speaker embedding and accent embedding, respectively. Finally, we train the seq2seq model to convert the source BNFs to the target Mel-spectrogram, conditioning on the target speaker's speaker and accent embeddings.

The workflow during inference is illustrated in Fig. 1(b). Accentron requires a source utterance from an L1 speaker and an utterance from the L2 speaker. First, we extract BNFs and accent embedding from the L1 utterance, which encode the desired linguistic content and native accent, and a speaker embedding from the L2 utterance, which encodes the desired voice identity of the L2 speaker. Then, we pass the L1 BNFs, L1 accent embedding, and L2 speaker embedding to the seq2seq model, which generates the accent-converted Mel-spectrogram. Finally, the Mel-spectrogram is converted back to a waveform, in our case using a separately trained WaveRNN (Kalchbrenner et al., 2018), though other vocoders or the Griffin–Lim algorithm (Griffin and Lim, 1984) can also be used.

3.1. Speaker-independent acoustic model

To capture the linguistic content of an utterance, we use the output of the last hidden layer of a speaker-independent acoustic model (AM) as BNFs, rather than the output of the final layer of the AM, which represent the PPGs (i.e., the probabilities of each senone/tri-phone). BNFs contain similar linguistic information as PPGs but have much lower dimensionality (e.g., Senone-PPG: 6024 dimensions; BNF: 256 dimensions), which avoids the need to perform dimensionality reduction in the seq2seq model.

Our AM is based on a Factorized Time Delayed Neural Network (TDNN-F) (Povey et al., 2018; Peddinti et al., 2015), a feed-forward neural network acting as a sequential classifier. Given an input acoustic feature vector (i.e., 40-dimensional MFCCs), the TDNN-F produces the probabilities of the vector belonging to each senone/triphone (6024 senones). The TDNN-F takes time-delayed input frames as side inputs to its hidden layers to model long-term temporal dependencies, concatenated with a 100-dimensional i-vector (Dehak et al., 2010) of the corresponding speaker.[3] Additionally, the TDNN-F uses factorized layers with semi-orthogonal constraints as hidden layers and dilated connections between hidden layers, which are more efficient during training and inference than recurrent layers due to their feed-forward nature (Povey et al., 2018). The TDNN-F model is composed of five hidden layers. Each of the first four hidden layers has 1280 neurons, followed by ReLU activation and batch normalization (Ioffe and Szegedy, 2015), whereas the last hidden layer has 256 neurons, corresponding to the dimensionality of the BNFs. We train the model through a supervised 6024-way senone classification task. To promote that the AM produces speaker-independent BNFs, we train the model on speech data from several thousands of speakers (Librispeech corpus Panayotov et al., 2015, 2484 native English speakers; see Section 4.1).

[3] As noted by Peddinti et al. (2015), this allows the model to capture both speaker and environment specific information, which is useful for neural network adaption.

IPR2026-00272
Krisp Technologies EX1010 Page 57

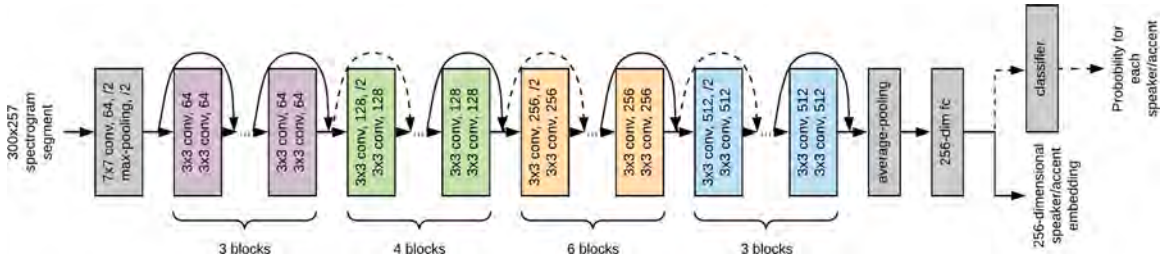


Fig. 2. Speaker/accent encoder model architecture. The model is based on ResNet-34 (He et al., 2016). Each convolution block is illustrated as the kernel size and channel numbers. "/2" means the layer divides the spatial resolution by 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Speaker and accent encoders

We use two independent encoders to compute the desired voice identity and accent. The speaker encoder is built as a speaker recognition model, which is trained to determine the identity of a speaker from an input utterance, whereas the accent encoder is based on an accent recognition model, which is trained to recognize accent/dialect patterns (e.g., pronunciation and prosody). For this work, we use a convolutional neural network (CNN) based on ResNet-34 (He et al., 2016) for both the speaker encoder and the accent encoder, following a previous speaker-recognition study (Li et al., 2017). We use the same CNN architecture for both models, so here we provide a detailed workflow only for the speaker encoder; the training and inference workflows of the accent encoder can be derived similarly.

The architecture of the speaker encoder is shown in Fig. 2. The model takes 300×257 in time $\times$ frequency magnitude spectrogram segments as inputs. The inputs are first fed to a convolution layer containing 64 7×7 kernels with 2×2 stride, followed by a 2×2 max-pooling layer. These layers decrease the spatial resolution of the feature maps, reducing model complexity and improving training speed. On top of them, there are 16 convolution residual blocks, which extract more abstract features. Each convolution block consists of two convolution layers with 3×3 kernels. The first convolution layer in each block has 2×2 stride to further decrease the spatial resolution of the feature maps. More importantly, each block has a skip connection as an alternative path to avoid gradient vanishing in a very deep model. The 16 convolution blocks have different numbers of kernels, as highlighted in different colors in Fig. 2 (Purple: 64 kernels; Green: 128 kernels; Orange: 256 kernels; Blue: 512 kernels). Next, there is an average pooling layer that produces a 512-dimensional vector, followed by a 256-dimensional fully-connected layer. All the layers are followed by ReLU activations and batch-normalization (Ioffe and Szegedy, 2015).

The model is trained through a supervised speaker-classification task. During training, a classifier on top of the 256-dimensional fully-connected layer produces the probabilities that the segment belongs to each speaker. The network is then optimized by minimizing the cross-entropy loss between the prediction and the target speaker label. During inference, we discard the final classifier layer and directly use the 256-dimensional bottleneck feature as the segment-wise speaker embedding. To obtain utterance-level speaker embeddings for a speaker that does not appear during training, we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors Variani et al., 2014).

3.3. Seq2seq foreign accent conversion model

Our seq2seq model is inspired by the text-to-speech Tacotron2 model (Shen et al., 2018). As shown in Fig. 3, the seq2seq model has an encoder–decoder architecture. During training, inputs to the network consist of a triplet: (1) a sequence of BNFs from the source (L1) speaker, $\mathbf{x} \in \mathbb{R}^{T_i \times D_{BNF}}$, (2) a speaker embedding of the target (L1 or L2) speaker $\mathbf{s} \in \mathbb{R}^{D_{speaker}}$ extracted from the speaker encoder, and (3) the accent embedding of the target speaker $\mathbf{a} \in \mathbb{R}^{D_{accent}}$ extracted from the accent encoder. T_i is the length of the sequence $\mathbf{x}$. D_{BNF} is the dimensionality of the BNFs (e.g., 256 in this work). $D_{speaker}$ and D_{accent} are the dimensionalities of the speaker embedding $\mathbf{s}$ and accent embedding $\mathbf{a}$, respectively (both of them are 256 in this work). The ground-truth target of the model is a sequence of Mel-spectrogram frames $\mathbf{y} \in \mathbb{R}^{T_o \times D_{Mel}}$, where T_o is the length of the sequence and D_{Mel} is the number of Mel-filterbanks (e.g., 80 in this work). First, the encoder accepts a BNF sequence $\mathbf{x}$ and produces a hidden representation $\mathbf{h}$:

$$\mathbf{h} = \text{Encoder}(\mathbf{x}) \tag{1}$$

Then, to condition the decoder on the voice identity and the accent of the target speaker, we concatenate the target speaker's speaker embedding and accent embedding to the hidden representation:

$$\mathbf{h}_{\mathbf{concat}} = [\mathbf{h}, \mathbf{s}, \mathbf{a}] \tag{2}$$

Finally, the decoder autoregressively predicts the Mel-spectrogram of the target speech using the attention context computed based on the concatenated hidden representation:

$$\hat{\mathbf{y}}^t = \text{Decoder}(\hat{\mathbf{y}}^{t-1}, \mathbf{h}_{\mathbf{concat}}) \tag{3}$$

where $\hat{\mathbf{y}}^t$ is the tth frame of the predicted Mel-spectrogram.

IPR2026-00272
Krisp Technologies EX1010 Page 58

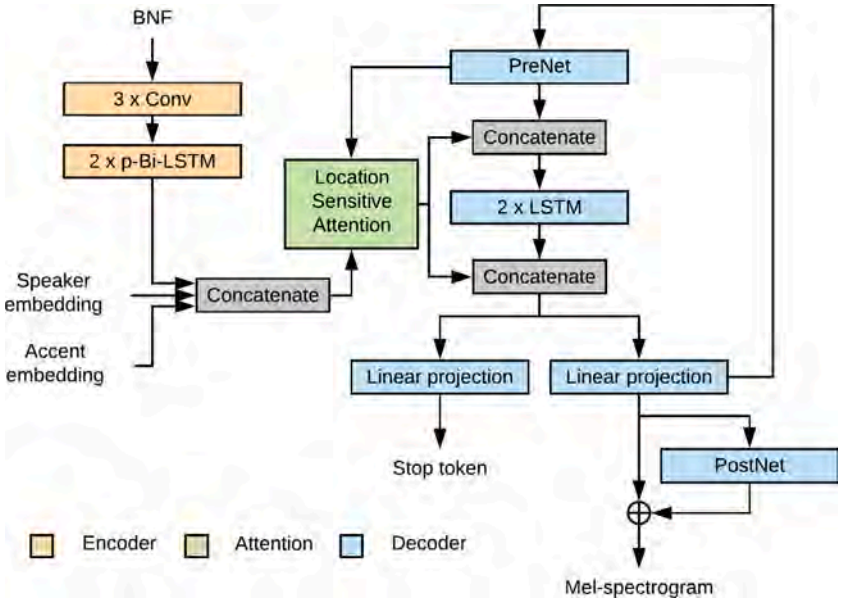


Fig. 3. The seq2seq model in Accentron.

Table 1
Hyperparameters of the seq2seq foreign accent conversion model.

Block	Component	Parameters
Inputs	BNF	256-dimensional
	Speaker d-vector	256-dimensional
	Accent d-vector	256-dimensional
Encoder	3 × Conv layers	512 5 × 1 kernel with 1 × 1 stride; ReLU; batch norm
	2 × p-Bi-LSTM	256 cells per direction;
Attention	Location-sensitive attention	128-dim attention context; 32 31 × 1 attention conv kernels
Decoder	PreNet	2 × Fully-connected layer ; 256 neurons; ReLU
	2 × LSTM	1024 cells
	Linear (Mel)	1 × Fully-connect layer; 80 units; no activation
	Linear (stop token)	1 × Fully-connect layer; 1 unit; no activation
	PostNet	5 × Conv layers; 512 5 × 1 kernel with 1 × 1 stride; tanh; batch norm
Outputs	Mel-spectrogram	80-dimensional
	Stop token	2-dimensional

During inference, the inputs to the network are also a triplet: (1) a sequence of BNFs from an L1 speaker, $\mathbf{x}$, (2) a speaker embedding of an L2 speaker, $\mathbf{s}_{\mathbf{L2}}$, and (3) an accent embedding of an L1 speaker, $\mathbf{a}_{\mathbf{L1}}$. The network produces a hidden representation $\mathbf{h}$, concatenates it with $\mathbf{s}_{\mathbf{L2}}$ and $\mathbf{a}_{\mathbf{L1}}$, and feeds it to the decoder to produce the predicted FAC Mel-spectrogram. We describe each component in the following subsections. The hyper-parameters of each component are summarized in Table 1.

3.3.1. Encoder

The encoder converts a BNF sequence to a hidden representation sequence. The original text-to-speech Tacotron2 encoder contains three 1-dimensional convolution layers and one Bidirectional Long Short-Term Memory (Bi-LSTM) layer. However, in our case, the inputs of the seq2seq model are BNF sequences instead of text embeddings, which are usually significantly longer. To capture the high-level phonetic and contextual information in an input BNF sequence, we replace the LSTM layer in the encoder with two pyramidal Bidirectional LSTM (p-Bi-LSTM) layers (Chan et al., 2016). Each p-Bi-LSTM reduces the time resolution by a factor of two, and therefore our encoder produces four times shorter hidden representation sequences compared with the input sequences. A convolution layer has 512 kernels with 5 × 1 shape in time×frequency and 1 × 1 stride, followed by ReLU activation

IPR2026-00272
Krisp Technologies EX1010 Page 59

and batch normalization. Each convolution kernel spans five BNF frames, which models the local context information. A p-Bi-LSTM layer has 256 cells in each direction, followed by ReLU activation and batch normalization, producing a 512-dimensional hidden representation sequence.

3.3.2. Decoder

The decoder is an autoregressive recurrent neural network coupled with a local sensitive attention mechanism (Chorowski et al., 2015). The decoder accepts the concatenated hidden representation sequences as inputs, and produces an 80-dimensional Mel-spectrogram as the prediction of the L2 speech. During each decoding step, the predicted Mel-spectrogram frame from the previous step is first passed into a pre-net that has two 256-dimensional fully-connected layers with ReLU activations:

$$\mathbf{q}^t = \text{PreNet}(\hat{\mathbf{y}}^{t-1}) \tag{4}$$

The pre-net acts as an information bottleneck, which is essential for learning attentions (Shen et al., 2018). Next, the location-sensitive attention mechanism computes a 128-dimensional attention context vector $\mathbf{c}^t$ based on the pre-net output, the concatenated hidden representations, and the attention context from the previous step:

$$\mathbf{c}^t = \text{Attention}(\mathbf{q}^t, \mathbf{h}_{\mathbf{concat}}, \mathbf{c}^{t-1}) \tag{5}$$

Following this, the pre-net output is concatenated with the context vector and fed to two unidirectional LSTM layers with 256 cells. Then, the output of the second LSTM layer is concatenated again with the context vector $\mathbf{c}^t$ and passed through an 80-unit linear layer to make a prediction of the 80-dimensional L2 Mel-spectrogram frame:

$$\hat{\mathbf{y}}^t_{\mathbf{pre}} = \text{Linear}(\text{LSTM}(\mathbf{q}^t, \mathbf{c}^t), \mathbf{c}^t) \tag{6}$$

More importantly, the network also predicts if the generating process should stop at the current decoding step at the same time, i.e., a stop token $\hat{\mathbf{t}} \in \mathbb{R}^{T_o}$. Finally, to incorporate the spectral residual and improve synthesis quality, the predicted Mel-spectrogram is passed through a post-net consisting of five convolution layers to predict the residual. Each of these layers has 512 kernels with 5×1 shape and 1×1 stride, followed by *tanh* activation and batch normalization. The residual is added back to the original prediction to form the final prediction:

$$\hat{\mathbf{y}}^t = \hat{\mathbf{y}}^t_{\mathbf{pre}} + \text{PostNet}(\hat{\mathbf{y}}^t_{\mathbf{pre}}) \tag{7}$$

The model is optimized by minimizing the Euclidean distance between the target Mel-spectrogram and the prediction before/after the post-net. We also jointly minimize an extra cross-entropy loss to learn the stop token for model inference.

$$L = \|\hat{\mathbf{y}}^t_{\mathbf{pre}} - \mathbf{y}\|_2^2 + \|\hat{\mathbf{y}}^t - \mathbf{y}\|_2^2 + \lambda \text{CrossEntropy}(\hat{\mathbf{t}}, \mathbf{t}) \tag{8}$$

where $\|\cdot\|_2^2$ is the Euclidean distance; $\hat{\mathbf{t}}$ is the sequence of predicted stop tokens, and $\mathbf{t}$ is the sequence of target stop tokens; λ is the weight controlling the relative importance of the cross-entropy loss. Additionally, we use the teacher-forcing procedure during training by feeding in the correct output instead of the predicted output on the decoder side, which has been shown to improve the efficiency of the model training (Williams and Zipser, 1989).

4. Experimental setup

4.1. Acoustic model

We trained the TDNN-F acoustic model using the Librispeech corpus (Panayotov et al., 2015), which consists of 960 h of 16 kHz audiobook speech data produced by 2484 native English speakers, the majority being American English. The training set consists of two "clean" subsets and a "noisy" subset.[4] We used both sets in training to ensure that the BNF was speaker-independent. In addition, we used a subset (200 h) of the training set to train the i-vector extractor. We implemented the training following the official "tdnn_1d" recipe of the TDNN-F model in Kaldi.[5] The trained model achieves 3.76% word error rate (WER) on Librispeech's test-clean subset and 8.92% WER on the test-other subset.

4.2. Speaker encoder

We trained the speaker encoder using the VoxCeleb1 corpus (Nagrani et al., 2017), which contains 153,516 utterances of 16 kHz speech produced by 1251 speakers. Specifically, we used the training set from the official identification split, which is comprised of 138,316 utterances (~300 h) from these speakers. We extracted 257-dimensional magnitude spectrograms with a 25 ms window and 10 ms shift. We trained the model on a single NVIDIA Tesla V100 GPU with a batch size of 128. We used Adam Optimizer with an initial learning rate of 10^{-2}, which was annealed down to zero following a cosine schedule (Loshchilov and Hutter, 2017). The trained model achieves 81.34% Top-1 accuracy and 94.49% Top-5 accuracy on the official VoxCeleb1 identification testing set.

[4] We use the term "noisy" subset to refer to the test-other test subset in the LibriSpeech corpus. LibriSpeech has two test subsets: test-clean and test-other. The recordings in test-other subset have significantly higher background noise level than the test-clean subset.

[5] https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/local/chain/tuning/run_tdnn_1d.sh.

IPR2026-00272
Krisp Technologies EX1010 Page 60

4.3. Accent encoder

We trained the accent encoder using the Speech Accent Archive dataset (Weinberger, 2015), which consists of recordings of the "Please call Stella" paragraph (Weinberger, 2015) produced by speakers in 386 native and non-native English accents. For most of the accents, however, the number of speakers is limited, which may degrade the performance of the accent encoder. To address this issue, we selected a subset of accents where each accent has at least 30 speakers. The resulting subset has 18 accents,[6] with an average of 107 speakers in each accent. The total length of the selected subset is around 16 h. We randomly selected 90% utterances from each accent as the training set and used the remaining 10% utterances as the testing set. The audio waveforms in the original dataset have 8 kHz sampling rate. To match it with other modules, we resample them to 16 kHz. Other configurations were the same as that for speaker recognition. Our trained model achieves 79.36% Top-1 accuracy and 95.42% Top-5 accuracy on the testing set.

4.4. Seq2seq foreign accent conversion model

To evaluate the proposed approach, we conducted experiments with the ARCTIC (Kominek and Black, 2004) and L2-ARCTIC corpora (Zhao et al., 2018). We used four native English speakers from ARCTIC (BDL, RMS, SLT, CLB) and all 24 non-native English speakers from L2-ARCTIC. For each speaker, we divided their utterances into three subsets: a training set of 1032 utterances (~1 h of speech), a validation set of 50 utterances, and a testing set of 50 utterances. During training, we set BDL as the source speaker and paired it with all 28 speakers, including himself. During inference, we used both BDL (male) and CLB (female; used as an unseen L1 speaker for the zero-shot FAC setting) as the native reference speakers, and we performed FAC on four L2 speakers whose first languages were different: NJS (Spanish, female), TXHC (Mandarin, male), YKWK (Korean, male), and ZHAA (Arabic, female).

The original L2-ARCTIC audio waveforms have a 44.1 kHz sampling rate, so we resampled them to 16 kHz to match the ARCTIC recordings. We extracted 80-dimensional Mel-spectrogram with a 25 ms window and 10 ms shift. Following the same frame shift, we extracted BNFs for each utterance using the acoustic model (Section 3.1). In addition, we extracted utterance-level speaker and accent d-vectors from the speaker encoder and accent encoder, respectively. We implemented the model using TensorFlow (Abadi et al., 2016) and trained it on a single NVIDIA Tesla V100 GPU. The hyperparameter λ (Eq. (8)) in the loss function was set to 0.005 empirically. We set the batch size to 48, and we used an Adam Optimizer with an initial learning rate of 10^{-3}, which was then annealed down to 10^{-5} following exponential scheduling. The model converged after 200,000 steps, and the entire training time was around 100 h.[7] During model inference, we used a separately trained speaker-independent WaveRNN vocoder to invert the Mel-spectrogram back to the time-domain waveform. We trained the WaveRNN model on the Librispeech dataset.

5. Results

We have validated Accentron through a series of objective and subjective evaluations. For the objective evaluation, we used t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) to analyze the distribution of speaker embeddings and accent embeddings for the original speech and the accent-converted speech. This allowed us to examine whether the two encoder networks can decouple speaker identity and accent. For the first of two sets of subjective evaluations, we tested the system when the test L1 and L2 speakers were seen during training (*standard* FAC setting) and compared it against two state-of-the-art FAC systems (Zhao et al., 2019; Liu et al., 2020). We also tested whether our system could be used in the reverse direction, i.e., to impart a non-native accent to a native speaker's voice. In the second set of subjective evaluation, we explored the effectiveness of Accentron when the test L1 and L2 speakers were unseen during training (*zero-shot* FAC setting). Further, we characterized the performance of Accentron as a function of the number of available L2 test utterances during inference (i.e., which are used to extract the L2 speaker's voice identity footprint), and we compared it against a system that uses these utterances to fine-tune a pre-trained FAC system to understand the tradeoff between using L2 test utterances to compute a speaker embedding (zero-shot learning) and using them to refine a pre-trained model (fine-tuning).

5.1. Objective evaluation: speaker and accent embedding spaces

We used t-SNE (Van der Maaten and Hinton, 2008) to visualize the speaker and accent embedding spaces,[8] which helped provide a qualitative and intuitive explanation of how our proposed system operates. First, we visualized the speaker and accent embeddings of 20 FAC utterances for TXHC, a male Mandarin speaker. We used the system in the zero-shot condition when both L1 and L2 speakers were unseen (see Section 5.3.1) to generate the syntheses, since it acts as a performance lower-bound for all our systems, and it can also provide insights for zero-shot FAC. We also plotted the embeddings of natural speech from ten L1 and L2 speakers as references (20 utterances for each speaker). Results are shown in Fig. 4, where we use colors and shapes to represent speaker and accent, respectively (e.g., orange denotes speaker BDL, and diamond denotes L1 accent). As shown, the speaker embeddings of

[6] These 18 accents were Arabic, Cantonese, Dutch, English, Farsi, French, German, Hindi, Italian, Japanese, Korean, Mandarin, Polish, Portuguese, Russian, Spanish, Turkish, and Vietnamese.

[7] Audio samples from this work can be found at https://shaojinding.github.io/samples/accentron/.

[8] The axes in a t-SNE plot have no actual meaning. t-SNE projects high dimensional vectors into a lower dimensional plane (e.g., 2-D), such that similar points in the projection are also similar in the original space, and vice versa.

IPR2026-00272
Krisp Technologies EX1010 Page 61

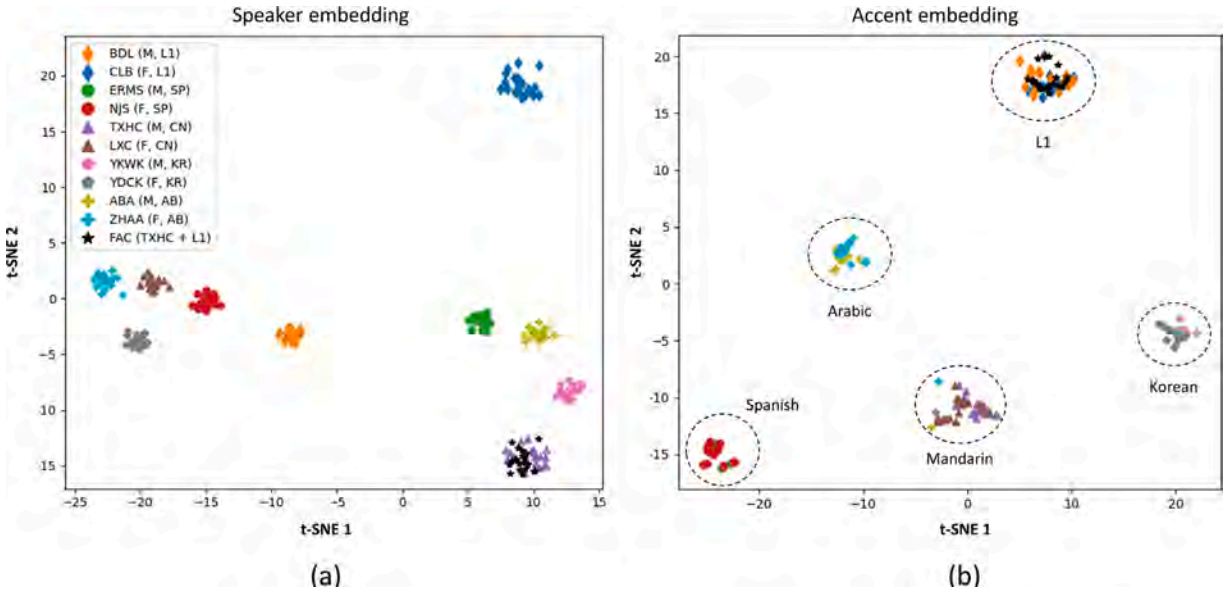


Fig. 4. Speaker and accent embedding visualization of FAC syntheses for TXHC using t-SNE. (a): speaker embedding; (b): accent embedding. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

utterances from the same speaker form a cluster, and the boundary between different clusters are clear. Similarly, accent embeddings from speakers with the same accent form a cluster. These two results indicate that the speaker and accent encoders are operating as expected. In terms of the FAC utterances, their speaker embeddings are distributed in the cluster of TXHC, which indicate that Accentron can generate speech that matches the voice identity of the target speaker, and the accent embeddings lie in a cluster of L1 speakers (BDL and CLB), which also indicate the Accentron can generate the desired accent. Together, these visualizations indicate that Accentron can successfully generate speech with the same voice identity as TXHC but with a native accent.

We also conducted t-SNE visualizations on a "reverse" FAC task (Zhao and Gutierrez-Osuna, 2019; Henter et al., 2018), where the goal was to synthesize speech with the voice identity of a given L1 speaker but with an L2 accent. This is a straightforward process in Accentron, since we only need to change the inputs of the seq2seq model to use an L2 accent embedding and an L1 speaker embedding during inference. For these visualizations, we synthesized a voice that had the voice identity of CLB but with an L2 (Mandarin) accent. As shown in Fig. 5, the speaker embeddings of the reverse FAC syntheses lie in the same cluster as CLB utterances, whereas the accent embeddings lie in the same cluster as the two Mandarin speakers (TXHC and LXC), indicating that the reverse FAC syntheses have a voice identity of CLB and a Mandarin accent.

Another interesting question that we would like to investigate here is whether the speaker embedding also carries accent cues, as these two aspects are closely related in the recognition of speakers (Leikin et al., 2009; Major et al., 2002; Shen et al., 2018; Cristia et al., 2012). If the speaker embedding is entangled with accent information, then there is the risk that Accentron would generate speech with an incorrect accent that is introduced by speaker embeddings. To examine this potential issue, we generated t-SNE visualizations of speaker embeddings for natural utterances from 16 speakers with 4 different accents in L2-ARCTIC corpus (See Fig. 6 for details of the speakers and accents). Results in Fig. 6 show no adjacency patterns among speakers sharing the same accent. Instead, t-SNE shows a distinct separation among speakers based on gender. These results suggest that speaker embedding mainly encodes information such as voice quality/timbre and pitch, and no evidence of potential entanglement between speaker and accent cues.

5.2. Subjective evaluations under standard FAC setting

We also evaluated Accentron through two sets of perceptual experiments. In the first set (this sub-section), we evaluated the system under the standard FAC setting, i.e., the test L1 and L2 speakers were seen during training. In the second set (Section 5.3), we evaluated the system under the zero-shot FAC setting, where the L1 speaker and/or the L2 speaker were unseen during training.

For the standard FAC setting, we used the union of the training sets of all 28 speakers to train the system. During inference, we used BDL as the L1 speaker, who then had been "seen" during training. First, we compared the proposed approach against two state-of-the-art FAC approaches:

- **Baseline1:** the FAC system proposed by Zhao et al. (2019), a one-to-one FAC approach based on seq2seq model. This baseline system trains a seq2seq PPG-to-speech synthesizer for each L2 speaker, and drives the synthesizer with PPGs extracted from an L1 speaker. As such, the baseline system requires training separate models for each L2 speaker.

IPR2026-00272
Krisp Technologies EX1010 Page 62

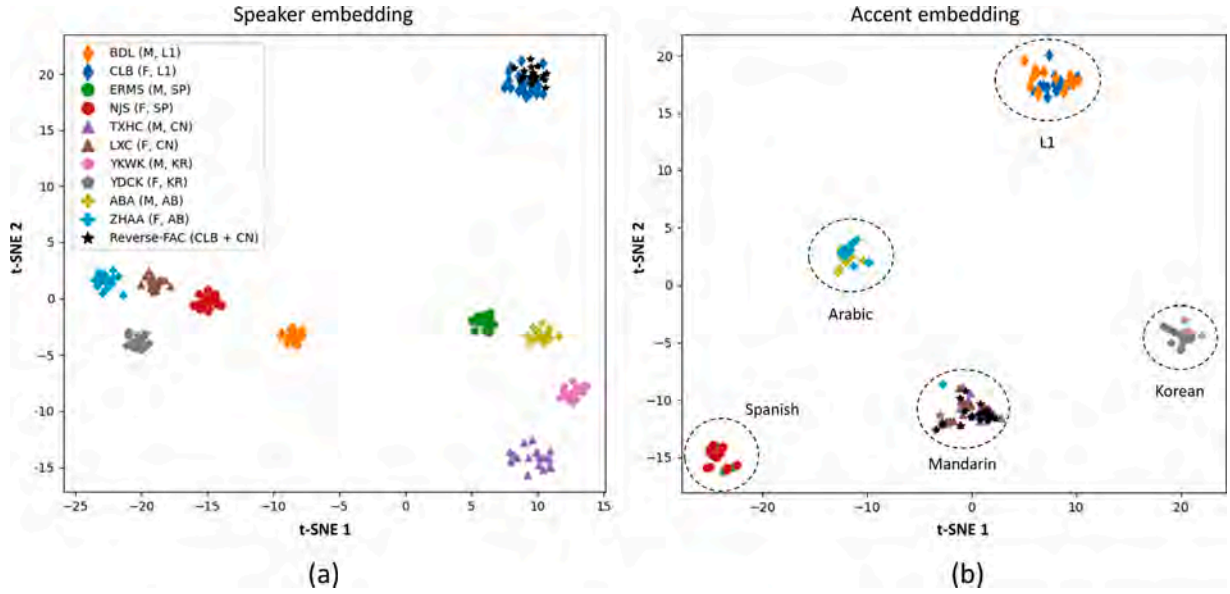


Fig. 5. Speaker and accent embedding visualization of *reverse* FAC syntheses (CLB with a Mandarin accent) using t-SNE. (a): speaker embedding; (b): accent embedding. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

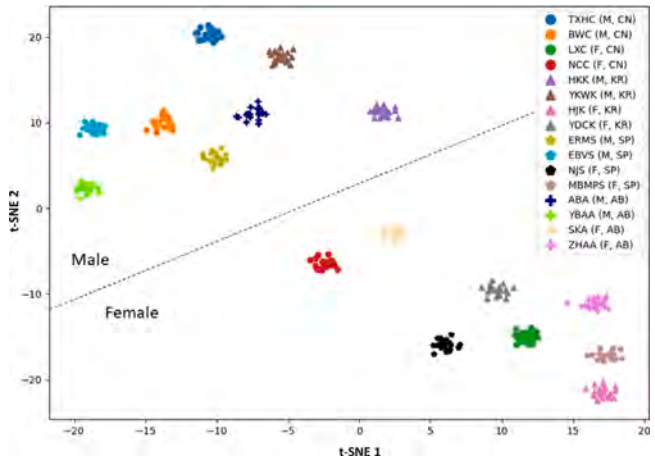


Fig. 6. t-SNE visualization of the speaker embeddings from 16 speakers with 4 accents. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- **Baseline2:** the system proposed by Liu et al. (2020), a reference-free many-to-many FAC approach based on a novel recognizer–synthesizer architecture. The system is trained on 105 speakers from CSTR VCTK dataset (Veaux et al., 2017). Audio samples were produced by feeding the test utterances through their system, which is provided as a courtesy by Liu et al. Due to the implementation differences between the systems, we conducted two post-processing steps to ensure a fair comparison. First, as the accent conversion model of Liu et al. was trained on VCTK speakers, the stop-token predictions on L2-ARCTIC test utterances are not robust, occasionally resulting in a few seconds of white noise at the end of speech in accent conversion syntheses. To solve this issue, we manually removed the trailing white noises in these test utterances. Second, we resampled the syntheses of Liu system from 22.05 kHz to 16 kHz to make the sampling rate be consistent with other systems.

We conducted listening tests through Amazon Mechanical Turk[9] to rate three perceptual attributes of the synthesized speech:

[9] https://www.mturk.com.

IPR2026-00272
Krisp Technologies EX1010 Page 63

Table 2
Accentedness (1-no foreign accent, 9-very strong foreign accent) results and acoustic quality (1-bad, 5-excellent) results under standard FAC setting. All the results are shown as average ± 95% confidence intervals.

System	Accentedness	Acoustic quality
Original L2	7.11 ± 0.21	3.67 ± 0.28
Original L1	1.06 ± 0.12	4.90 ± 0.10
Baseline1	4.63 ± 0.10	3.47 ± 0.14
Baseline2	6.25 ± 0.39	3.12 ± 0.13
Proposed	**3.39 ± 0.14**	**3.51 ± 0.15**

- **Accentedness:** : The test asked participants to rate the degree of foreign accentedness of each utterance in a 9-point scale (1-no foreign accent; 9-very strong foreign accent), which is commonly used in the pronunciation literature (Munro and Derwing, 2002). Participants were told that the native accent in this task was General American.
- **Acoustic quality:** The test asked participants to rate the acoustic quality of each utterance through a standard 5-point Mean Opinion Score (MOS; 1-bad, 5-excellent).
- **Voice identity:** The test asked participants to rate the voice similarity between the FAC syntheses and the original L2 speech through a 14-point Voice Similarity Score (VSS) (Felps and Gutierrez-Osuna, 2010). For each FAC-L2 utterances pair, participants were required to decide whether the two utterances were from the same speaker and then rate their confidence in the decision on a 7-point scale (1: not confident at all; 3: somewhat confident; 5: quite a bit confident; 7: extremely confident). The VSS was computed by collapsing the above two fields into a 14-point scale: −7 (definitely different speakers) to +7 (definitely the same speaker). To minimize the influence of accent, the two utterances had different linguistic content and were played in *reverse*, following Felps et al. (2009).

Instructions were given in each test to help participants focus on the target speech attribute. For example, in the accentedness test, participants were asked to *"Try to ignore the audio quality (noise, distortions). Please focus only on the speaker's accent, for example, their pronunciation, rhythm, and fluency"*. Test utterances were randomly selected from our test set, and the presentation order was counter-balanced. Additionally, in each listening test, we included five calibration utterances to detect if participants were cheating. We excluded ratings of the calibration utterances from the data analysis (Buchholz and Latorre, 2011). We recruited 18 participants for each listening test. All participants resided in the United States, and they passed a qualification test that asked them to identify different regional accents in the United States.

5.2.1. Comparison to the baseline system

Accentedness. Participants rated 20 utterances per system (5 utterances for each test L2 speaker). These utterances shared the same linguistic content across all the systems to ensure a fair comparison. Additionally, participants also rated the same set of sentences from the original L1 and L2 speakers as a reference. Results are shown in the first column of Table 2. Accentron received significantly lower ratings of foreign accentedness (3.39) than the original L2 utterance (7.11), though not as low as those of the original L1 utterance (1.06). These results suggest that our proposed seq2seq FAC model can effectively reduce foreign accentedness from the L2 speech. Accentron also outperformed the two baseline systems (Baseline1: 4.63, 27% relative improvement, $p \ll 0.001$; Baseline2: 6.25, 46% relative improvement, $p \ll 0.001$).

Acoustic quality. As shown in the second column of Table 2, the proposed method achieved an MOS of 3.51, which is comparable to Baseline1 (3.47, $p > 0.5$) but significantly higher than Baseline2 (3.12, 13% relative improvement, $p \ll 0.001$). The original L1 speech received the highest MOS (4.90), followed by the original L2 speech (3.67). Note that the MOS ratings of Accentron are closer to those of the original L2 speech than to the original L1 speech, possibly due to native listeners confounding acoustic quality with intelligibility (Felps et al., 2009), and therefore, they may be influenced by the intelligibility and provide lower ratings for non-native speech. Thus, the proposed system achieves similar acoustic quality as the baseline systems (or better), but unlike them does not require training a separate model for each new test L2 speaker.

Voice identity. Participants rated 20 pairs of utterances per system (5 pairs of utterances for each test L2 speaker). Each pair consisted of a FAC utterance and an utterance randomly selected from the L2 speaker. Voice identity results are shown in Table 3. Accentron achieved a 5.05 VSS, indicating that participants were "quite confident" that the FAC syntheses and the L2 speech were produced by the same speaker. These ratings are comparable to those of Baseline1 (5.05 VSS, $p > 0.5$) and significantly higher than those of Baseline2 (3.81 VSS, 33% relative improvement, $p \ll 0.001$). It is worth noting that the L1 speaker in Baseline1 had the same gender as the L2 speaker, whereas Accentron used the same L1 speaker for all L2 speakers. As a result, syntheses from Accentron included both intra (same)-gender FAC pairs and inter (different)-gender FAC pairs, the latter being more challenging due to the differences in prosody and pitch range. Although the VSS on inter-gender pairs (4.80) was lower than that on intra-gender pairs (5.29) and Baseline1, the difference was not significant ($p = 0.14$). These results suggest that the proposed system can generate FAC syntheses that greatly resemble the voice identity of L2 speakers of any gender, using a canonical reference L1 speaker.

IPR2026-00272
Krisp Technologies EX1010 Page 64

Table 3
Voice identity results under standard FAC setting. Voice Similarity Score ranges from −7 (definitely different speakers) to +7 (definitely the same speaker). All the results are shown as average ± 95% confidence intervals.

System	Voice similarity
Baseline1	5.05 ± 0.28
Baseline2	3.81 ± 0.29
Proposed (All pairs)	**5.05 ± 0.31**
Proposed (Intra-gender)	5.29 ± 0.30
Proposed (Inter-gender)	4.80 ± 0.35

Table 4
Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (−7-definitely different speakers, +7-definitely the same speaker) of *reverse* foreign accent conversion under standard condition. All the results are shown as average ± 95% confidence intervals.

System	Accentedness	Acoustic quality	Voice similarity (All pairs)	Voice similarity (Intra-gender)	Voice similarity (Inter-gender)
Proposed	5.58 ± 0.35	3.24 ± 0.17	4.91 ± 0.34	5.11 ± 0.35	4.71 ± 0.32

Table 5
The four conditions in zero-shot FAC experiment.

		L1 speaker	
		Seen	Unseen
L2 speaker	Seen	Condition SS	Condition US
	Unseen	Condition SU	Condition UU

5.2.2. Performance on reverse FAC

To evaluate Accentron on the reverse FAC task, we synthesized testing utterances using the accent embeddings from NJS, TXHC, YKWK, and ZHAA, and the speaker embedding from BDL. Table 4 shows the accentedness, acoustic quality, and voice identity results of the reverse FAC evaluation. Accentron received a 5.58 accentedness rating, much closer to that of the original L2 speech (7.11) than to the original L1 speech (1.06), indicating that our approach was able to impart an L2 accent to utterances from an L1 speaker. Accentron also received a 3.24 MOS, significantly lower ($p = 0.02$) than the MOS of the "direct" FAC syntheses (3.51), a result that is likely due to the correlation between acoustic quality and intelligibility — see Section 5.2.1. Finally, Accentron received a 4.91 VSS, indicating that raters were "quite confident" that the reverse FAC syntheses and the L1 speech were produced by the same speaker; we found no significant differences between the voice identity ratings of reverse and direct FAC syntheses. Thus, we can conclude that Accentron can also operate in the reverse direction, generating non-native utterances with the voice identity of a native speaker.

5.3. Subjective evaluations under zero-shot FAC setting

In the second set of subjective evaluations, we evaluated the proposed system under the zero-shot FAC setting, where the L1 speaker and/or the L2 speaker were unseen during training. The zero-shot FAC setting is appealing for real-world applications since it requires minimal data from the target speaker. First, we compared the performance of Accentron when using seen/unseen L1 or L2 speakers during inference. Then, we characterize its performance as a function of the number of available L2 utterances.

5.3.1. Comparing different conditions in zero-shot foreign accent conversion

We considered four different conditions in this experiment, as summarized in Table 5. In condition SS, the L1 speaker and the L2 speaker were both seen during training. Note this condition is the same as the system evaluated in Section 5.2.1, so it serves as a best-case scenario. In condition US, the L1 speaker was unseen during training, and the L2 speaker was seen during training. In condition SU, the L1 speaker was seen during training, and the L2 speaker was unseen during training. Finally, in condition UU, the L1 speaker and the L2 speaker were both unseen during training. Thus, condition UU was the most challenging of the four.

To ensure that the test speakers were unseen during training, we trained four models using different training sets. In Condition SS, we used the same model as in the standard FAC condition. In Condition US, we excluded CLB from the training set and used it as the test L1 speaker. In Condition SU, we excluded the four test L2 speakers from the training set. In Condition UU, we excluded the four test L2 speakers and CLB from the training set, and we also used CLB as the test L1 speaker. For unseen L1/L2 speakers, we used the 50 utterances from the test set to generate the accent/speaker embedding. As before, we conducted three types of listening tests through Amazon Mechanical Turk to rate the accentedness, acoustic quality, and voice similarity of the synthesized speech. In addition, we kept the participants the same as those in the first experiment, so that the results are comparable between different experiments (e.g., participants in the accentedness test for the two experiments were the same).

IPR2026-00272
Krisp Technologies EX1010 Page 65

Table 6
Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (−7-definitely different speakers, +7-definitely the same speaker) under zero-shot FAC condition. All the results are shown as average ± 95% confidence intervals.

System	Accentedness	Acoustic quality	Voice Similarity
Condition SS	3.39 ± 0.14	3.51 ± 0.15	5.05 ± 0.28
Condition US	3.33 ± 0.26	3.47 ± 0.13	4.99 ± 0.30
Condition SU	3.35 ± 0.25	3.50 ± 0.12	4.92 ± 0.28
Condition UU	3.30 ± 0.26	3.43 ± 0.12	4.59 ± 0.34

Table 7
Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (−7-definitely different speakers, +7-definitely the same speaker) with different numbers of available L2 (non-native) utterances during inference. All the results are shown as average ± 95% confidence intervals.

#L2 utterances	Accentedness		Acoustic quality		Voice similarity	
	Proposed	Fine-tuned	Proposed	Fine-tuned	Proposed	Fine-tuned
50	3.30 ± 0.26	3.03 ± 0.24	3.43 ± 0.12	3.54 ± 0.11	4.59 ± 0.34	4.97 ± 0.27
20	3.30 ± 0.22	3.47 ± 0.18	3.45 ± 0.11	3.48 ± 0.11	4.68 ± 0.30	4.65 ± 0.23
10	3.34 ± 0.26	3.84 ± 0.18	3.44 ± 0.12	3.46 ± 0.11	4.59 ± 0.29	4.06 ± 0.26
5	3.32 ± 0.23	4.58 ± 0.10	3.43 ± 0.11	3.38 ± 0.10	4.42 ± 0.33	3.49 ± 0.34
1	3.31 ± 0.25	4.72 ± 0.08	3.43 ± 0.12	3.24 ± 0.13	4.57 ± 0.29	3.73 ± 0.35

Results from the accentedness, acoustic quality, and voice identity tests are shown in Table 6. We found no statistically significant differences between condition SS (best-case scenario) and the three more challenging conditions (US, SU, UU); $p > 0.5$ in all cases. This result suggests that Accentron has no trouble generalizing to unseen L1 or (and) L2 speakers during inference without any degradation in accentedness, acoustic quality, and voice identity.

5.3.2. Influence of the number of available L2 utterances

For practical FAC applications, it is important to understand the minimum amount of data needed from a target speaker. Requiring L2 learners to record a large amount of speech before they can hear their "golden speaker" voice can be tedious and demotivating. On the other hand, training the system with insufficient speech data might significantly degrade synthesis quality. To characterize the data requirements of Accentron under zero-shot FAC condition, we measured its performance of the UU codition (unseen L1 speaker and unseen L2 speaker) as a function of the number of available L2 utterances. We used the UU condition since it is the most flexible for real-world applications, and also the most challenging, which provides a lower bound of performance. For these experiments, we used 50 test L2 utterances to produce the speaker embedding during inference, and reduced the number from 50 to 1 (N = 50, 20, 10, 5, 1) and re-evaluated system performance. Results are shown in Table 7. Reducing the number of utterances from 50 to 1 has no impact on any of the three perceptual measures ($p > 0.5$ in all cases). These results indicate that as little as a single utterance (∼3 s of speech) is sufficient to generate accent conversions for a new unseen L2 speaker, with no impact on performance.

To some extent, the above result is to be expected since test utterances are only used to compute the speaker embedding. Thus, we also examined whether test utterances could instead be more beneficial if they were used to fine-tune a pre-trained FAC system. Starting with a pre-trained UU model, we fine-tuned the model on each unseen L1–L2 speaker pair (i.e., CLB-NJS, CLB-TXHC, CLB-YKWK, CLB-ZHAA) with N = 50, 20, 10, 5, 1 test utterances, resulting in 20 fine-tuned models (4 speakers; 5 models with different number of training utterances for each speaker) for the unseen L2 speakers. Results are also shown in Table 7. We observe performance degradations in all three measurements when reducing the number of utterances from 50 to 1. When there are 50 test utterances, the fine-tuned system shows a marginal improvement compared to the zero-shot model (i.e., without fine-tuning), though the differences are not statistically significant (Accentedness: 3.03 vs. 3.30, $p = 0.03$; Acoustic quality: 3.54 vs. 3.43, $p = 0.18$; Voice identity: 4.97 vs. 4.59, $p = 0.25$). When decreasing the number from 50 to 20, the fine-tuned system achieves comparable performance as the zero-shot system ($p > 0.5$). Surprisingly, however, fine-tuning the systems with fewer than 20 utterances degrades performance compared to the zero-shot model. In the extreme case (with only 1 utterance), the zero-shot model significantly outperforms the fine-tuned model in all three measurements (Accentedness: 4.72 vs. 3.31, $p \ll 0.001$; Acoustic quality: 3.24 vs. 3.43, $p = 0.01$; Voice identity: 3.73 vs. 4.57, $p \ll 0.001$). These results further speak of the robustness of Accentron in the zero-shot condition, and they also illustrate the tradeoff between zero-shot learning models and fine-tuning models in FAC.

6. Discussion

We have proposed Accentron, a zero-shot many-to-many speech synthesizer that can convert utterances from a source speaker to appear as if someone else, and with a different accent, had produced it. We thoroughly evaluated the system through a series of objective and perceptual listening experiments. Visualizations through t-SNE show that Accentron captures the target voice identity and accent, and that the speaker and accent embeddings are independent of each other and effectively summarize the speaker and accent characteristic of an utterance.

IPR2026-00272
Krisp Technologies EX1010 Page 66

We evaluated Accentron in a standard FAC setting and compared it against two state-of-the-art baseline FAC systems. Compared to baseline 1 (Zhao et al., 2019), Accentron achieves significantly better (i.e., lower) ratings of foreign accentedness, and similar acoustic quality and voice identity ratings. This is an important finding since baseline 1 builds a dedicated model for each pair of L1–L2 speakers, which one would expect would help capture voice identity more faithfully than Accentron's many-to-many mapping. Although baseline 1 and Accentron use the same backbone architecture for the seq2seq model, Accentron achieves significantly better (lower) ratings of accentedness. A possible explanation for this result is that baseline 1 is trained exclusively on L2 utterances. Thus, if the L2 speaker has systematic substitution or deletion errors (e.g., Mandarin speakers from certain areas systematically substitute /SH/ with /S/), the correct pronunciations will be missing in their utterances. Thus, when the baseline 1 model is driven by L1 BNFs during inference, it has to interpolate these missing pronunciations, which leads to noticeable segmental errors. In contrast, Accentron avoids this potential issue since it is trained using both L1 and L2 speech. Compared to baseline 2 (Liu et al., 2020), Accentron achieves better ratings in all three measurements (accentedness, acoustic quality, and voice identity). Though Accentron and baseline 2 both use a seq2seq architecture, Accentron has two additional components: an acoustic model and an accent encoder to extract the linguistic content and accent embeddings, respectively, from an L1 reference during inference. These two embeddings capture the L1 segmental and prosodic patterns, respectively, which are shown to be essential to achieve advanced FAC performance.

We also evaluated Accentron on a "reverse FAC" task, where the goal was to impart an L2 accent to a native utterance. Results on this task corroborates the t-SNE visualizations, and suggest that Accentron can also preserve an L2 accent and implant it into an L1 speaker's utterance. Combined, results from the direct and reverse FAC tasks indicate that the proposed system can disentangle linguistic content, voice identity, and accent in speech signals, instead of merely memorizing mappings between different speaker pairs.

Finally, we also evaluated Accentron in a zero-shot FAC setting. First, we compared all four combinations in which the L1 speaker and L2 speakers could have been seen/unseen during training (i.e., SS, SU, US, UU). And we found no significant differences among the four conditions in terms of accentedness, acoustic quality, or voice identity. Thus, Accentron performs equally well under a standard FAC setting (condition SS) and a zero-shot FAC setting (condition UU), which indicates that it generalizes to unseen L1 and L2 speakers without the need to re-train or fine-tune the model.

Can Accentron generalize to unseen accents? In principle, the Accentron architecture can naturally generalize to unseen accents — in the same way that it is able to model unseen speakers. However, this requires access to sufficient training data on a large number of accents. The Speech Accent Archive has hundreds of accents, but unfortunately each speaker only produces less than one minute of speech, which is insufficient for training the FAC model (empirically, at least one hour per speaker is needed). Its counterpart, the L2-ARCTIC corpus, has substantially more speech data per subject (around one hour), but only has six accents, which is insufficient to achieve reasonable transferability to an unseen accent. Thus, while we believe that Accentron can be used to generate unseen accents, this ability can only be truly assessed when more accented speech corpora become available.

Our results under both standard and zero-shot FAC setting indicate that Accentron can generate FAC synthesis with high-quality, and can achieve it with limited data from new L1 and L2 speakers (one utterance, or around three seconds of speech). This capability can dramatically simplify the deployment of pronunciation-training tools (our envisioned target application). When deploying conventional FAC models, one needs to design and implement the model training pipeline, as it requires training a dedicated model for each target L2 speaker. Running the training pipeline is usually resource- and time-consuming, and therefore, one has to include specific modules to manage the server's computational resources (e.g., an asynchronous queue was used in Ding et al., 2019a). Such systems are hard to scale to an increasing number of users, due to the heavy resource demands. Instead, Accentron only requires running the model inference pipeline when producing FAC synthesis, and the inference can be accomplished in real time, thus essentially reducing the resource demands and simplifying the application design and deployment. In addition, in previous FAC studies (e.g., Aryal and Gutierrez-Osuna, 2014; Zhao and Gutierrez-Osuna, 2019; Zhao et al., 2019), inter-gender conversion usually achieved inferior performance than intra-gender conversion, due to the mismatch in pitch ranges. As a result, when deploying these methods into practical applications, they have to use a reference L1 speaker that has the same gender as the target L2 speaker. In contrast, Accentron achieves similar voice identity ratings for intra-gender pairs and inter-gender pairs, which gives more flexibility when choosing the reference L1 speakers in the pronunciation-training tool. Finally, Accentron significantly reduces the data required for each new L2 speaker from hours to seconds. As a result, the L2 learners only need to go through a simple speaker enrollment process (recording several seconds of their speech), before practicing with their "golden speaker" voices, which makes Accentron an ideal system for performing pronunciation-training studies at scale.

7. Conclusion and future work

In this paper, we have proposed Accentron, a zero-shot learning system that can generate accent conversion for any L2 speaker (seen or unseen). Our proposed approach is in contrast to most of existing FAC approaches, which require building a separate model for each L2 speaker. The proposed approach first trains separate models to extract L1 bottleneck features, L1 accent embeddings, and L2 speaker embeddings. Then it uses a seq2seq model to transform L1 bottleneck features to accent-converted Mel-spectrogram, conditioned on an L1 accent embedding and L2 speaker embedding. Our results suggest that the system can successfully transform L1 speech to match the voice identity of an L2 speaker while using a small amount of data from the L2 speaker.

One possible future direction of this work is to improve its robustness in generating long utterances. Currently, our system uses a location-sensitive attention mechanism (Chorowski et al., 2015) in the seq2seq model, which can fail when the utterances are too long (Skerry-Ryan et al., 2018) (e.g., longer than 10 s). To solve this problem, an alternative attention mechanism could be used, such as Gaussian mixture attention mechanism (Graves, 2013), which has been shown to be more robust in generating

IPR2026-00272
Krisp Technologies EX1010 Page 67

long utterances (Skerry-Ryan et al., 2018). An additional potential improvement would be to add an auxiliary decoder to perform phoneme recognition (during training) (Biadsy et al., 2019; Jia et al., 2019). Such auxiliary decoder would guide the hidden representation produced by the encoder to preserve phonetic information, enforcing the synthesized speech to be phonetically reasonable and improving synthesis quality (Biadsy et al., 2019; Jia et al., 2019).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Arık, S.O., Chen, J., Peng, K., Ping, W., Zhou, Y., 2018. Neural voice cloning with a few samples. In: International Conference on Neural Information Processing Systems. pp. 10040–10050.

Aryal, S., Felps, D., Gutierrez-Osuna, R., 2013. Foreign accent conversion through voice morphing. In: INTERSPEECH. pp. 3077–3081.

Aryal, S., Gutierrez-Osuna, R., 2014. Can voice conversion be used to reduce non-native accents? In: ICASSP. IEEE, pp. 7879–7883.

Aryal, S., Gutierrez-Osuna, R., 2015a. Articulatory-based conversion of foreign accents with deep neural networks. In: INTERSPEECH.

Aryal, S., Gutierrez-Osuna, R., 2015b. Reduction of non-native accents through statistical parametric articulatory synthesis. J. Acoust. Soc. Am. 137 (1), 433–446.

Aryal, S., Gutierrez-Osuna, R., 2016. Data driven articulatory synthesis with deep neural networks. Comput. Speech Lang. 36, 260–273.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations.

Biadsy, F., Weiss, R.J., Moreno, P.J., Kanvesky, D., Jia, Y., 2019. Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. In: INTERSPEECH. pp. 4115–4119.

Brand, M., 1999. Voice puppetry. In: Annual Conference on Computer Graphics and Interactive Techniques. pp. 21–28.

Buchholz, S., Latorre, J., 2011. Crowdsourcing preference tests, and how to detect cheating. In: INTERSPEECH. pp. 3053–3056.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: ICASSP. IEEE, pp. 4960–4964.

Chen, L.-H., Ling, Z.-H., Liu, L.-J., Dai, L.-R., 2014. Voice conversion using deep neural networks with layer-wise generative training. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1859–1872.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems. pp. 577–585.

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., Floccia, C., 2012. Linguistic processing of accented speech across the lifespan. Front. Psychol. 3, 479.

Das, A., Zhao, G., Levis, J., Chukharev-Hudilainen, E., Gutierrez-Osuna, R., 2020. Understanding the effect of voice quality and accent on talker similarity. In: INTERSPEECH. pp. 1763–1767.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2010. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Denby, B., Stone, M., 2004. Speech synthesis from real time ultrasound images of the tongue. In: ICASSP, Vol. 1. IEEE, pp. I–685.

Desai, S., Black, A.W., Yegnanarayana, B., Prahallad, K., 2010. Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio Speech Lang. Process. 18 (5), 954–964.

Ding, S., Gutierrez-Osuna, R., 2019. Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion. In: INTERSPEECH. pp. 724–728.

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis, J., Gutierrez-Osuna, R., 2019a. Golden speaker builder–an interactive tool for pronunciation training. Speech Commun. 115, 51–66.

Ding, S., Zhao, G., Gutierrez-Osuna, R., 2020. Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition. In: INTERSPEECH. pp. 776–780.

Ding, S., Zhao, G., Liberatore, C., Gutierrez-Osuna, R., 2019b. Learning structured sparse representations for voice conversion. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 343–354.

Erro, D., Moreno, A., Bonafonte, A., 2009. INCA algorithm for training voice conversion systems from nonparallel corpora. IEEE Trans. Audio Speech Lang. Process. 18 (5), 944–953.

Felps, D., Bortfeld, H., Gutierrez-Osuna, R., 2009. Foreign accent conversion in computer assisted pronunciation training. Speech Commun. 51 (10), 920–932.

Felps, D., Geng, C., Gutierrez-Osuna, R., 2012. Foreign accent conversion through concatenative synthesis in the articulatory domain. IEEE Trans. Audio Speech Lang. Process. 20 (8), 2301–2312.

Felps, D., Gutierrez-Osuna, R., 2010. Developing objective measures of foreign-accent conversion. IEEE Trans. Audio Speech Lang. Process. 18 (5), 1030–1040.

Graves, A., 2013. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time fourier transform. IEEE Trans. Acoust. Speech Signal Process. 32 (2), 236–243.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Henter, G.E., Lorenzo-Trueba, J., Wang, X., Kondo, M., Yamagishi, J., 2018. Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody. In: ICASSP. IEEE, pp. 4799–4803.

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M., 2016. Voice conversion from non-parallel corpora using variational auto-encoder. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. IEEE, pp. 1–6.

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M., 2017a. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. In: INTERSPEECH, Vol. 2017.

Hsu, W.-N., Zhang, Y., Glass, J., 2017b. Unsupervised learning of disentangled and interpretable representations from sequential data. In: Advances in Neural Information Processing Systems. pp. 1878–1889.

Huang, W.-C., Wu, Y.-C., Hwang, H.-T., Tobing, P.L., Hayashi, T., Kobayashi, K., Toda, T., Tsao, Y., Wang, H.-M., 2019. Refined WaveNet vocoder for variational autoencoder based voice conversion. In: European Signal Processing Conference. IEEE, pp. 1–5.

Huckvale, M., Yanagisawa, K., 2007. Spoken language conversion with accent morphing. In: ISCA Workshop on Speech Synthesis. pp. 64–70.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456.

IPR2026-00272
Krisp Technologies EX1010 Page 68

Jia, Y., Weiss, R.J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., Wu, Y., 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In: INTERSPEECH.
Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Moreno, I.L., Wu, Y., et al., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Advances in Neural Information Processing Systems. pp. 4480–4490.
Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., Kavukcuoglu, K., 2018. Efficient neural audio synthesis. In: International Conference on Machine Learning.
Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N., 2018. ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder. arXiv preprint arXiv:1808.05092.
Kaneko, T., Kameoka, H., Hiramatsu, K., Kashino, K., 2017. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In: INTERSPEECH, Vol. 2017. pp. 1283–1287.
Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N., 2019. CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion. In: ICASSP. IEEE, pp. 6820–6824.
Kominek, J., Black, A.W., 2004. The CMU ARCTIC speech databases. In: Fifth ISCA Workshop on Speech Synthesis.
Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 951–958.
Leikin, M., Ibrahim, R., Eviatar, Z., Sapir, S., 2009. Listening with an accent: Speech perception in a second language by late bilinguals. J. Psycholinguist. Res. 38 (5), 447.
Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., Zhu, Z., 2017. Deep speaker: an end-to-end neural speaker embedding system. arXiv preprint arXiv:1705.02304.
Liu, S., Wang, D., Cao, Y., Sun, L., Wu, X., Kang, S., Wu, Z., Liu, X., Su, D., Yu, D., et al., 2020. End-to-end accent conversion without using native utterances. In: ICASSP. IEEE, pp. 6289–6293.
Liu, S., Zhong, J., Sun, L., Wu, X., Liu, X., Meng, H., 2018. Voice conversion across arbitrary speakers based on a single target-speaker utterance. In: INTERSPEECH. pp. 496–500.
Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations.
Lu, H., Wu, Z., Dai, D., Li, R., Kang, S., Jia, J., Meng, H., 2019. One-shot voice conversion with global speaker embeddings. In: INTERSPEECH. pp. 669–673.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9 (11).
Major, R.C., Fitzmaurice, S.F., Bunta, F., Balasubramanian, C., 2002. The effects of nonnative accents on listening comprehension: Implications for esl assessment. TESOL Quart. 36 (2), 173–190.
Miyoshi, H., Saito, Y., Takamichi, S., Saruwatari, H., 2017. Voice conversion using sequence-to-sequence learning of context posterior probabilities. In: INTERSPEECH. pp. 1268–1272.
Mohammadi, S.H., Kain, A., 2017. An overview of voice conversion systems. Speech Commun. 88, 65–82.
Mohammadi, S.H., Kim, T., 2019. One-shot voice conversion with disentangled representations by leveraging phonetic posteriorgrams. In: INTERSPEECH. pp. 704–708.
Mumtaz, R., Preuß, S., Neuschaefer-Rube, C., Hey, C., Sader, R., Birkholz, P., 2014. Tongue contour reconstruction from optical and electrical palatography. IEEE Signal Process. Lett. 21 (6), 658–662.
Munro, M.J., Derwing, T.M., 2002. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Lang. Learn. 49 (s1), 285–310.
Nagrani, A., Chung, J.S., Zisserman, A., 2017. VoxCeleb: A large-scale speaker identification dataset. In: INTERSPEECH. pp. 2616–2620.
Nakashika, T., Takiguchi, T., Minami, Y., 2016. Non-parallel training in voice conversion using an adaptive restricted boltzmann machine. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (11), 2032–2045.
Oshima, Y., Takamichi, S., Toda, T., Neubig, G., Sakti, S., Nakamura, S., 2015. Non-native speech synthesis preserving speaker individuality based on partial correction of prosodic and phonetic characteristics. In: INTERSPEECH.
Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. In: ICASSP. IEEE, pp. 5206–5210.
Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: INTERSPEECH.
Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S., 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks.. In: INTERSPEECH. pp. 3743–3747.
Probst, K., Ke, Y., Eskenazi, M., 2002. Enhancing foreign language tutors–in search of the golden speaker. Speech Commun. 37 (3–4), 161–173.
Saito, Y., Ijima, Y., Nishida, K., Takamichi, S., 2018. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In: ICASSP. IEEE, pp. 5274–5278.
Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al., 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: ICASSP. IEEE, pp. 4779–4783.
Sisman, B., Yamagishi, J., King, S., Li, H., 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. arXiv preprint arXiv:2008.03648.
Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R.J., Clark, R., Saurous, R.A., 2018. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. arXiv preprint arXiv:1803.09047.
Sun, L., Li, K., Wang, H., Kang, S., Meng, H., 2016a. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In: IEEE International Conference on Multimedia and Expo. IEEE, pp. 1–6.
Sun, L., Wang, H., Kang, S., Li, K., Meng, H.M., 2016b. Personalized, cross-lingual TTS using phonetic posteriorgrams. In: INTERSPEECH. pp. 322–326.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112.
Takashima, R., Takiguchi, T., Ariki, Y., 2012. Exemplar-based voice conversion in noisy environment. In: 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 313–317.
Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K., Toda, T., 2017. Speaker-dependent wavenet vocoder. In: INTERSPEECH, Vol. 2017. pp. 1118–1122.
Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2222–2235.
Toutios, A., Sorensen, T., Somandepalli, K., Alexander, R., Narayanan, S.S., 2016. Articulatory synthesis based on real-time magnetic resonance imaging data.. In: INTERSPEECH. pp. 1492–1496.
Turk, O., Arslan, L.M., 2002. Subband based voice conversion. In: International Conference on Spoken Language Processing.
Van Den Oord, A., Vinyals, O., et al., 2017. Neural discrete representation learning. In: Advances in Neural Information Processing Systems. pp. 6306–6315.
Van Heugten, M., Bergmann, C., Cristia, A., 2015. The effects of talker voice and accent on young children's speech perception. In: Individual Differences in Speech Production and Perception. Peter Lang, pp. 57–88.
Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. In: ICASSP. IEEE, pp. 4052–4056.
Veaux, C., Yamagishi, J., MacDonald, K., et al., 2017. Superseded-Cstr Vctk Corpus: English Multi-Speaker Corpus for Cstr Voice Cloning Toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR).
Wang, R., Lu, J., 2011. Investigation of golden speakers for second language learners from imitation preference perspective by voice modification. Speech Commun. 53 (2), 175–184.

IPR2026-00272
Krisp Technologies EX1010 Page 69

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al., 2017. Tacotron: Towards end-to-end speech synthesis. In: INTERSPEECH. pp. 4006–4010.
Weinberger, S., 2015. Speech Accent Archive. George Mason University.
Williams, R.J., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1 (2), 270–280.
Xie, F.-L., Soong, F.K., Li, H., 2016. A KL divergence and DNN-based approach to voice conversion without parallel training sentences.. In: INTERSPEECH. pp. 287–291.
Zhang, J.-X., Ling, Z.-H., Dai, L.-R., 2019. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 540–552.
Zhao, G., Ding, S., Gutierrez-Osuna, R., 2019. Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. In: INTERSPEECH. pp. 2843–2847.
Zhao, G., Gutierrez-Osuna, R., 2019. Using phonetic posteriorgram based frame pairing for segmental accent conversion. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (10), 1649–1660.
Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., Gutierrez-Osuna, R., 2018. L2-ARCTIC: A non-native english speech corpus. In: INTERSPEECH. pp. 2783–2787.
Zhou, Y., Tian, X., Xu, H., Das, R.K., Li, H., 2019. Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling. In: ICASSP. IEEE, pp. 6790–6794.

IPR2026-00272
Krisp Technologies EX1010 Page 70

Appendix DING02

MARC View

LEADER 02265nas a2200541 a 4500			
001 49130348			
003 SIRSI			
005 20251014090047.0			
006 m o d			
007 cr n			
008 970531c19869999ne qx pss 0 a0eng c			
010			a| sn 97001928
015			a| GBB540824 2| bnb
016	7		a| 1462883-1 2| DE-600
016	7		a| 012432536 2| Uk
019			a| DO NOT EDIT this 360 MARC RECORD
022			a| 1095-8363 l| 0885-2308 y| 0885-2308
035			a| (WaSeSS)ssj0006547
040			a| OH1 b| eng c| OH1 d| NSD d| OCL d| OCLCQ d| GUA d| OCLCQ d| MUQ d| DLC d| OCLCQ d| YUS d| CSU d| BUF d| OCLCQ d| DLC d| OCLCF d| OCLCO d| OCLCQ d| OCLCO d| OCLCQ d| Z5A d| OCLCO d| OCLCA d| NLE d| EZC d| AU@ d| OCLCO d| WYU d| UKMGB d| VT2 d| U3W d| OCLCO d| OCLCQ d| OCLCO d| OCL d| OCLCQ d| UAB d| OCLCL d| WaSeSS
042			a| nsdp a| pcc
050		4	a| TK7882.S65 b| C63
082	1	0	a| 006 2| 12 1| https://id.oclc.org/worldcat/ddc/E379xBCBchbdmrqYKbR6BJFPFX
130	0		a| Computer speech & language (Online)
210	0		a| Comput. speech lang. b| (Online)
222		0	a| Computer speech & language b| (Online)
245	1	0	a| Computer speech & language h| [electronic resource].
246	3		a| Computer speech and language
260			a| London : b| Academic Press, c| ©1986-
260	3		3| <2004-> : a| [Amsterdam] : b| Elsevier
310			a| Five no. a year, b| 2012-
321			a| Quarterly, b| 1986-2011
362	1		a| Print began with v. 1, no. 1 (Mar. 1986).
515			a| Vol. 1 complete in 2 issues.
538			a| Mode of access: World Wide Web.
588	0		a| Vol. 9, no. 1 (Jan. 1995); title from home page.
588	1		a| V. 26, issue 5 (Oct. 2012) (ScienceDirect, viewed July 30, 2012).
650		0	a| Speech processing systems v| Periodicals.
650		0	a| Automatic speech recognition v| Periodicals.
650		0	a| Linguistics.
650		2	a| Computers
650		2	a| Linguistics
650		2	a| Speech-Language Pathology
655		0	a| Electronic journals.
655		2	a| Periodical
776	0	8	i| Print version: t| Computer speech & language x| 0885-2308 w| (DLC) 87658573 w| (OCoLC)12597555
856	4	0	u| http://SK8ES4MC2L.search.serialssolutions.com/?V=1.0&L=SK8ES4MC2L&S=JCs&C=COMPSPELA&T=marc
910			a| CONSER online
949			a| Electronic resource w| ASIS c| 1 i| 49130348-1001 l| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 10/18/2025 o| .STAFF. 360 MARC

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

Computer speech & language [electronic resource].

Uniform Title	Computer speech & language (Online)
Additional Titles	Comput. speech lang. (Online) and Computer speech and language
Published	London : Academic Press, ©1986- <2004-> : [Amsterdam] : Elsevier

Access Online

serialssolutions.com

Dates of Publication and/or Sequential Designation

Print began with v. 1, no. 1 (Mar. 1986).

Subject(s)

Speech processing systems—Periodicals
Automatic speech recognition—Periodicals
Linguistics
Computers
Speech-Language Pathology

Genre(s)

Electronic journals

ISSN

1095-8363

Publication Frequency

Five no. a year, 2012-
Quarterly, 1986-2011

Special Numbering

Vol. 1 complete in 2 issues.

Technical Details

Mode of access: World Wide Web.

ASK A LIBRARIAN

View MARC record | catkey: 49130348

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

(814) 865-6368

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

Online Journals, Books and more

All Journals Only Books Only Videos Only

Title begins with Search

Show more search options ▼

Showing results 1 through 1 of 1
for the search: "COMPSPELA"

Refine Results: All | Journals Only | Books Only | Videos Only

Limit by: Peer Reviewed | Open Access



Computer speech & language

ISSN: 0885-2308

Peer Reviewed

1986 to 1994 in Backfile Package - Computer Science (Legacy) [YCS]
Backfile Package - Psychology including Supplement 1 [YUP]
Backfile Package - Social Science (Legacy) [YST]
Backfile Package - Social Science including Supplement 1 [YUS]

03/01/1986 to Present in ScienceDirect Journals

Search inside this journal Search

Back to Top

Copyright ©2025 The Pennsylvania State University.
All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Last Updated on 10-28-2025 © Ex Libris, a ProQuest Company, 2000-2025

Appendix DING03

ScienceDirect Journals & Books Help Search My account Sign in

Access through your organization View Open Manuscript Purchase PDF

Article preview

Abstract

Introduction

Section snippets

References (91)

Cited by (25)

Computer Speech & Language

Volume 72, March 2022, 101302



Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning ☆

Shaojin Ding, Guanlong Zhao [1], Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, USA

Received 12 April 2021, Revised 19 September 2021, Accepted 21 September 2021, Available online 5 October 2021, Version of Record 3 December 2021.

What do these dates mean?

Check for updates

Show less

Add to Mendeley Share Cite

https://doi.org/10.1016/j.csl.2021.101302 Get rights and content

Abstract

Foreign accent conversion (FAC) aims to create a new voice that has the *voice identity* of a given second-language (L2) speaker but with a native (L1) *accent*. Previous FAC approaches usually require training a separate model for each L2 speaker and, more importantly, generally require considerable speech data from each L2 speaker for training. To address these limitations, we propose Accentron, an approach that can generate accent-converted speech for arbitrary L2 speakers unseen during training. In the proposed approach, we first train a speaker-independent acoustic model on L1 corpora to extract bottleneck features that represent the linguistic content of utterances. Then, we develop a speaker encoder and an accent encoder to generate embedding vectors for the desired voice identity (L2 speaker's) and accent (L1 accent), respectively. Lastly, we use a sequence-to-sequence model to transform bottleneck-features to Mel-spectrograms, conditioned on the L2 speaker embedding and the L1 accent embedding. We conducted experiments on the L2-ARCTIC corpus under two testing conditions: the *standard* FAC setting where test L2 speakers were seen during training, and a *zero-shot* FAC setting where test L2 speakers were unseen during training. Accentron achieves over 27% relative improvement in accentedness ratings compared to two state-of-the-art FAC systems in the standard FAC setting. More importantly, our results show that Accentron generalizes to the zero-

Recommended articles

Cross-language article linking with deep neural network based...

Computer Speech & Language, Volume 72,...

Yu-Chun Wang, ..., Richard Tzong-Han Tsai

Analysis of out-of-breath speech for assessment of person's...

Computer Speech & Language, Volume 76,...

Suman Deb, Samarendra Dandapat

Treatment of postictal headache

Epilepsy & Behavior, Volume 123, 2021, Arti...

Carlo Cianchetti

Show 3 more articles

Article Metrics

Citations

Citation Indexes 25

Captures

Mendeley Readers 21

PLUMX View details

shot FAC setting with no performance loss. Therefore, in practical use scenarios (e.g., computer-assisted pronunciation training software), Accentron can effectively avoid the need to adapt or retrain the model, which significantly reduces computations and the users' waiting time.

Introduction

Foreign accent conversion (FAC)(Felps et al., 2009) aims to create a new voice that has the voice identity of a given L2 speaker and the accent of an L1 speaker. In pronunciation training, FAC can serve as a "golden speaker" for the L2 speaker to practice with: their own voice, but with a native accent(Felps et al., 2009, Probst et al., 2002, Ding et al., 2019a, Wang and Lu, 2011). FAC also finds applications in movie dubbing(Turk and Arslan, 2002), personalized text-to-speech (TTS) synthesis(Sun et al., 2016b, Oshima et al., 2015), and improving speech recognition performance(Biadsy et al., 2019). A variety of techniques have been proposed to perform FAC, including voice morphing(Felps et al., 2009, Aryal et al., 2013, Huckvale and Yanagisawa, 2007), frame pairing(Aryal and Gutierrez-Osuna, 2014, Zhao and Gutierrez-Osuna, 2019), articulatory synthesis(Aryal and Gutierrez-Osuna, 2015a, Aryal and Gutierrez-Osuna, 2015b), and sequence-to-sequence (seq2seq) modeling(Zhao et al., 2019, Liu et al., 2020). However, previous FAC approaches have two major limitations. First, they operate in a one-to-one fashion, i.e.,they require training a separate model for each pair of L1 and L2 speakers. Second, they need a considerable amount of speech data (~1000 utterances) for each L2 speaker. Thus, when using these conventional FAC methods in real-world applications such as pronunciation training, L2 learners need to record a large number of utterances and then wait for a dedicated model to be trained, which can be tedious and demotivating.

To address this issue, we propose Accentron, a zero-shot learning(Lampert et al., 2009) approach to FAC that can synthesize speech for arbitrary L2 speakers who were unseen during training. Accentron consists of four independent models: (1) a speaker-independent acoustic model that captures the linguistic content of an L1 utterance as a sequence of *bottleneck feature vectors*, (2) a speaker encoder that captures the voice identity of the L2 speaker, denoted as a *speaker embedding*, (3) an accent encoder that captures the desired L1 accent, denoted as an *accent embedding*, and (4) a seq2seq model that generates a Mel-spectrogram from the sequence of bottleneck features, conditioned on the desired speaker and accent embeddings. These components can be trained independently, at which point the system can generate accent conversions to arbitrary L2 speakers given a few seconds of audio (i.e.,enough speech to compute a speaker embedding), without the need to have any model re-training or adaptation process.

To our knowledge, ours is the first work to apply zero-shot learning for the task of FAC. Though zero-shot learning has been used for voice conversion(Liu et al., 2018, Hsu et al., 2016, Kameoka et al., 2018, Sun et al., 2016a) and voice cloning(Jia et al., 2018, Arık et al., 2018), previous studies(Liu et al., 2018, Hsu et al., 2016, Kameoka et al., 2018, Sun et al., 2016a, Jia et al., 2018, Arık et al., 2018) have focused exclusively on manipulating voice identity, ignoring the speaker's accent, which holds important cues to speaker recognition(Das et al., 2020) and speech perception(Leikin et al., 2009, Major et al., 2002, VanHeugten et al., 2015, Cristia et al., 2012). Incorporating accent into the conversion process requires changes to the conventional encoder–

IPR2026-00272
Krisp Technologies EX1010 Page 77

decoder structure of sequence-to-sequence (seq2seq) models for voice conversion. Our encoder takes a sequence of L1 bottleneck feature vectors as the input, and produces a hidden representation sequence. In a conventional voice conversion system(Liu et al., 2018, Hsu et al., 2016, Ding and Gutierrez-Osuna, 2019, Saito et al., 2018, Hsu et al., 2017b, Mohammadi and Kim, 2019, Lu et al., 2019, Ding et al., 2020), this hidden representation sequence is then concatenated with the speaker embedding of the target speaker. In our case, however, the system also concatenates the accent embedding, which is treated as an additional independent and controllable factor during synthesis. The combined bottleneck/speaker/accent embedding is consumed by a decoder coupled with a location-sensitive attention mechanism(Chorowski et al., 2015). During each decoding step, the decoder autoregressively predicts a Mel-spectrogram frame based on the output from the previous decoding step and a context vector produced by the attention mechanism. Finally, the output Mel-spectrogram is converted back into a waveform through either the Griffin–Lim algorithm(Griffin and Lim, 1984) or a separately trained vocoder (e.g.,WaveNet(Tamamori et al., 2017), WaveRNN(Kalchbrenner et al., 2018)).

We thoroughly evaluated Accentron on the L2-ARCTIC corpus(Zhao et al., 2018). First, we visualized the speaker and accent embedding distributions for the accent-converted speech and natural speech, and the results show that Accentron syntheses can successfully capture the L2 voice identity along with an L1 accent. Second, we conducted a series of listening tests under two different settings: (1) a *standard* FAC setting, where the test L2 speakers were available during training, and (2) a *zero-shot* FAC setting, which assumes that the test L2 speakers were not available during training. Accentron achieves 27% relative improvement in accentedness while retaining the acoustic quality and voice identity, compared to two state-of-the-art FAC systems in standard FAC settings. In addition, Accentron showed no performance degradation when tested under zero-shot FAC setting.

The manuscript is organized as follows. Section2 reviews prior approaches to foreign accent conversion, many-to-many voice conversion, and sequence-to-sequence models. Section3 describes the architecture of Accentron in detail. Section4 provides the experimental setup, including the corpora and implementation details. Section5 presents the objective and subjective evaluations of Accentron and analyzes the results. We discuss the implications of the results in Section6. Lastly, we conclude the findings of this work and point out potential future directions in Section7.

Section snippets

Foreign accent conversion

The problem of foreign accent conversion was first formulated by Felps et al. (2009) as the means to provide implicit feedback in computer assisted pronunciation training. Early approaches(Aryal and Gutierrez-Osuna, 2015b, Felps et al., 2012, Aryal and Gutierrez-Osuna, 2016, Brand, 1999) involved building an articulatory synthesizer for the L2 speaker. The articulatory synthesizer was trained to map the L2 speaker's articulatory trajectories (e.g.,tongue and lip movements) into his or her ...

IPR2026-00272
Krisp Technologies EX1010 Page 78

Methods

Accentron consists of four modules: (1) a speaker-independent acoustic model that generates a linguistic representation of an utterance, (2) a speaker encoder that captures the voice identity of the desired speaker, (3) an accent encoder that captures the desired accent, and (4) a seq2seq model that consumes the previous three representations to synthesize Mel-spectrogram for an arbitrary L2 speaker.

The workflow for training Accentron is shown in Fig. 1(a). The acoustic model, speaker encoder, ...

Acoustic model

We trained the TDNN-F acoustic model using the Librispeech corpus(Panayotov et al., 2015), which consists of 960 h of 16 kHz audiobook speech data produced by 2484 native English speakers, the majority being American English. The training set consists of two "clean" subsets and a "noisy" subset.[4] ...

Results

We have validated Accentron through a series of objective and subjective evaluations. For the objective evaluation, we used t-distributed stochastic neighbor embedding (t-SNE)(Vander Maaten and Hinton, 2008) to analyze the distribution of speaker embeddings and accent embeddings for the original speech and the accent-converted speech. This allowed us to examine whether the two encoder networks can decouple speaker identity and accent. For the first of two sets of subjective evaluations, we ...

Discussion

We have proposed Accentron, a zero-shot many-to-many speech synthesizer that can convert utterances from a source speaker to appear as if someone else, and with a different accent, had produced it. We thoroughly evaluated the system through a series of objective and perceptual listening experiments. Visualizations through t-SNE show that Accentron captures the target voice identity and accent, and that the speaker and accent embeddings are independent of each other and effectively summarize the ...

Conclusion and future work

In this paper, we have proposed Accentron, a zero-shot learning system that can generate accent conversion for any L2 speaker (seen or unseen). Our proposed approach is in contrast to most of existing FAC approaches, which require building a separate model for each L2 speaker. The proposed approach first trains separate models to extract L1 bottleneck features, L1 accent embeddings, and L2 speaker embeddings. Then it uses a seq2seq model to transform L1 bottleneck features to accent-converted ...

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. ...

References (91)

AryalS. *et al.*
Data driven articulatory synthesis with deep neural networks
Comput. Speech Lang. (2016)

DingS. *et al.*
Golden speaker builder–an interactive tool for pronunciation training
Speech Commun. (2019)

FelpsD. *et al.*
Foreign accent conversion in computer assisted pronunciation training
Speech Commun. (2009)

HsuC.-C. *et al.*
Voice conversion from non-parallel corpora using variational auto-encoder

MohammadiS.H. *et al.*
An overview of voice conversion systems
Speech Commun. (2017)

ProbstK. *et al.*
Enhancing foreign language tutors–in search of the golden speaker
Speech Commun. (2002)

WangR. *et al.*
Investigation of golden speakers for second language learners from imitation preference perspective by voice modification
Speech Commun. (2011)

AbadiM. *et al.*
Tensorflow: Large-scale machine learning on heterogeneous distributed systems
(2016)

ArıkS.O. *et al.*
Neural voice cloning with a few samples

AryalS. *et al.*
Foreign accent conversion through voice morphing

View more references

Cited by (25)

The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation ↗
2023, Frontiers in Psychology

Show abstract

Controllable Accented Text-to-Speech Synthesis With Fine and

Coarse-Grained Intensity Rendering ↗
2024, IEEE ACM Transactions on Audio Speech and Language Processing

Voice-Preserving Zero-Shot Multiple Accent Conversion ↗
2023, ICASSP IEEE International Conference on Acoustics Speech and Signal Processing Proceedings

TTS-Guided Training for Accent Conversion Without Parallel Data ↗
2023, IEEE Signal Processing Letters

Towards Zero-Shot Multi-Speaker Multi-Accent Text-to-Speech Synthesis ↗
2023, IEEE Signal Processing Letters

Zero-Shot Foreign Accent Conversion without a Native Reference ↗
2022, Proceedings of the Annual Conference of the International Speech Communication Association Interspeech

View all citing articles on Scopus ↗

☆ This work was supported by NSF awards 1619212 and 1623750.

1 Ding and Zhao are now with Google. This work was done solely with TAMU resources.

View full text

© 2021 Elsevier Ltd. All rights reserved.

PDF
Help

About ScienceDirect ↗ Remote access Contact and support ↗ Terms and conditions ↗ Privacy policy ↗

RELX™

Cookies are used by this site. **Cookie settings**

All content on this site: Copyright © 2025 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the relevant licensing terms apply.

Appendix DING04

https://web.archive.org/web/20220201000000*/https://service.elsevier.com/app/answers/detail/a_id/11241/supporthub/scopus/ October 28, 2025

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE

INTERNET ARCHIVE WayBackMachine

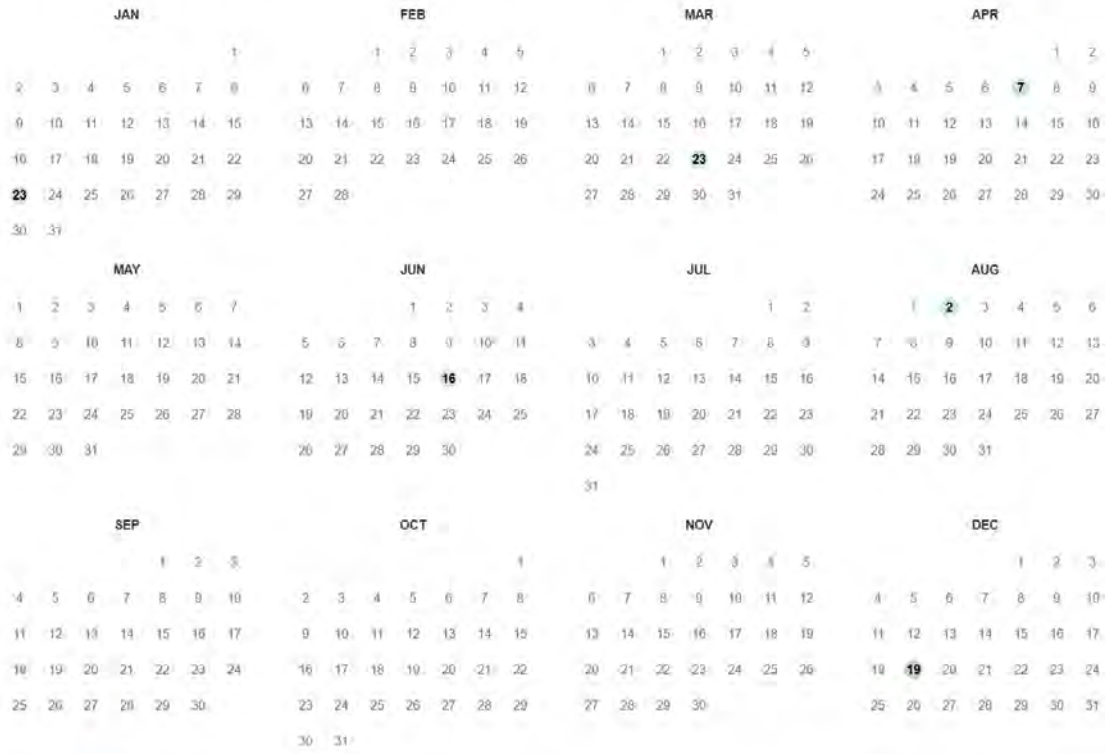
Explore more than 1 trillion web pages saved over time

https://service.elsevier.com/app/answers/detail/a_id/11241/suppc

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **6 times** between January 23, 2022 and December 19, 2022.





Note

This calendar view maps the number of times **https://service.elsevier.com/app/answers/detail/a_id/11241/supporthub/scopus/** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ. Green indicates redirects (3xx).

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20220123095825/https://service.elsevier.com/app/answers/detail/a_id/11241/supporthub/scopus/ October 28, 2025
INTERNET ARCHIVE WaybackMachine https://service.elsevier.com/app/answers/detail/a_id/11241/supporthub/scopus/ Go DEC JAN 23 2022 MAR 6 captures 23 Jan 2022 - 19 Dec 2022 2021 2023 About this capture

Orders & Renewals
Access
Onboarding
Training
Using the product
Content
New and updated FAQs

What are Articles in Press?

Last updated on June 30, 2020

Articles in Press are documents that have been accepted for publication, but have not yet been assigned to a journal issue. They are indicated by the Articles in Press symbol on document pages and in search result lists.

Scopus makes the pre-published versions of documents available online within four days after being transferred from publishers.

Note: Corrections to Articles in Press cannot be made because they are not yet published. Scopus can only correct a document if the data on Scopus deviates from the original *published* article. For more information see page 24 of the Scopus Content Coverage Guide.

Once the final version is published and made available, Scopus will overwrite the Article in Press in the database with the final version.

You can view Articles in Press for a Scopus source from the Source Information page.

How

To view Articles in Press for a publication

1. From the Scopus toolbar, click '*Sources*'. The Scopus sources page opens.
2. From the Scopus sources page, use Search for a source or Browse sources to locate a publication. A results list of sources opens.
3. Once you have created a source results list, click a publication title. The source information page opens.
4. From the source information page, click '*Articles in press*'.

To search for articles in press

Was this answer helpful?

Yes No

Related Articles:

- What can I do on a Source details page?
- How do I search for a document?
- How do I use the Scopus Sources feature?
- What is Scopus Preview?
- How can I best use the Advanced search?

For further assistance:

Email

Chat

Terms and Conditions Privacy Policy Cookie Settings

Copyright ©2022 Elsevier B.V.

ELSEVIER

RELX™

Appendix DING05

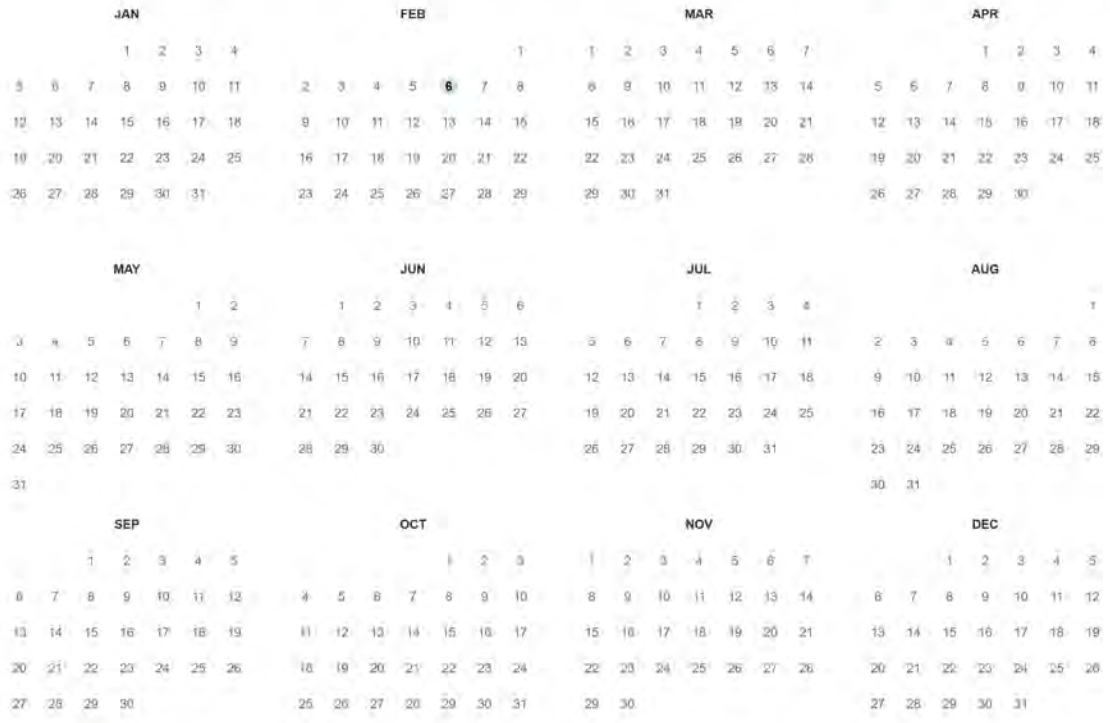
INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE INTERNET ARCHIVE WayBackMachine Explore more than 1 trillion web pages saved over time
https://www.sciencedirect.com/journal/computer-speech-and-lan

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **3 times** between February 6, 2020 and February 10, 2024.





Note

This calendar view maps the number of times **https://www.sciencedirect.com/journal/computer-speech-and-language/articles-in-press** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use

https://web.archive.org/web/20200206160426/https://www.sciencedirect.com/journal/computer-speech-and-language/articles-in-press November 14, 2025

INTERNET ARCHIVE WaybackMachine https://www.sciencedirect.com/journal/computer-speech-and-language/articles-in-press Go JAN FEB JAN 06 2019 2020 2023 About this capture
3 captures 6 Feb 2020 - 10 Feb 2024

ADVERTISEMENT



Computer Speech & Language

Supports *open access*

Articles in press

Articles in press are accepted, peer reviewed articles that are not yet assigned to volumes/issues, but are citable using DOI. Note to users

Actions for selected articles
Select all / Deselect all

Download PDFs

Export citations

Show all article previews

Receive an update when the latest issues in this journal are published

Sign in to set up alerts

Help improve this page

Research article ○ Abstract only
Learning Multi-Level Information for Dialogue Response Selection by Highway Recurrent Transformer
Ting-Rui Chiang, Chao-Wei Huang, Shang-Yu Su, Yun-Nung Chen
In Press, Journal Pre-proof, Available online 3 February 2020
Purchase PDF Article preview

Research article ● *Open access*
Noetic End-to-End Response Selection with Supervised Neural Network Based Classifiers and Unsupervised Similarity Models
Paweł Skórzewski, Weronika Sieińska, Marek Kubis
In Press, Journal Pre-proof, Available online 3 February 2020
Download PDF Article preview

Research article ○ Abstract only
Sequential Neural Networks for Noetic End-to-End Response Selection
Qian Chen, Wen Wang
In Press, Journal Pre-proof, Available online 30 January 2020
Purchase PDF Article preview

Research article ○ Abstract only
Generating Diverse Conversation Responses by Creating and Ranking Multiple Candidates
Yu-Ping Ruan, Zhen-Hua Ling, Xiaodan Zhu, Quan Liu, Jia-Chen Gu
In Press, Journal Pre-proof, Available online 24 January 2020
Purchase PDF Article preview

Research article ● *Open access*
Context and Knowledge Aware Conversational Model and System Combination for Grounded Response Generation
Ryota Tanaka, Akihide Ozeki, Shugo Kato, Akinobu Lee
In Press, Journal Pre-proof, Available online 24 January 2020

Download PDF Article preview

Research article ○ Abstract only

Overview of the seventh Dialog System Technology Challenge: DSTC7

Luis Fernando D'Haro, Koichiro Yoshino, Chiori Hori, Tim K. Marks, ... Xiang Gao

In Press, Journal Pre-proof, Available online 15 January 2020

Purchase PDF Article preview

Research article ○ Abstract only

13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE

Pavel Matějka, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, ... Jan "Honza" Černocký

In Press, Journal Pre-proof, Available online 17 December 2019

Purchase PDF Article preview

Research article ○ Abstract only

Recurrent neural network language generation for spoken dialogue systems

Tsung-Hsien Wen, Steve Young

In Press, Uncorrected Proof, Available online 18 June 2019

Purchase PDF Article preview

Erratum ● Full text access

Corrigendum to 'Extensions to Hybrid Code Networks for FAIR Dialog Data' Computer Speech & Language volume 53 (2019) Pages 80-91

Jiyeon Ham, Soohyun Lim, Kyeng-Hun Lee, Kee-Eung Kim

In Press, Corrected Proof, Available online 13 November 2018

Download PDF

Note to users:

Articles in press are peer reviewed, accepted articles to be published in this publication. When the final article is assigned to volumes/issues of the publication, the article in press version will be removed and the final version will appear in the associated published volumes/issues of the publication. The date an article was first made available online will be carried over. Please be aware that, although articles in press do not have all bibliographic details available yet, they can already be cited using the year of online publication and the DOI, as follows: author(s), article title, Publication (year), DOI.

Please consult the journal's reference style for the exact appearance of these elements, abbreviation of journal names and use of punctuation.

There are three types of Articles in Press:

- **Journal pre-proofs**: versions of an article that have undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but are not yet definitive versions of record. These versions will undergo additional copyediting, typesetting and review before being published in final form, but are provided to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
- **Uncorrected proofs**: articles that have been copy edited and formatted, but have not been finalized yet. They still need to be proof-read and corrected by the author(s) and the text could still change before final publication.
- **Corrected proofs**: articles that contain the authors' corrections. Final citation details, e.g. volume and/or issue number, publication year and page numbers, still need to be added and the text might change before final publication.

ISSN: 0885-2308

Copyright © 2020 Elsevier Ltd. All rights reserved

We use cookies to help provide and enhance our service and tailor content and ads. By continuing you agree to the **use of cookies**.
Copyright © 2020 Elsevier B.V. or its licensors or contributors. ScienceDirect ® is a registered trademark of Elsevier B.V.
ScienceDirect ® is a registered trademark of Elsevier B.V.

Appendix DING06

https://web.archive.org/web/20210801000000*/https://guanlongzhao.github.io/ October 28, 2025

INTERNET ARCHIVE | WEB | TEXTS | VIDEO | AUDIO | SOFTWARE | IMAGES | SIGN UP | LOG IN | UPLOAD | Search

ABOUT | BLOG | EVENTS | PROJECTS | HELP | DONATE | CONTACT | JOBS | VOLUNTEER

DONATE

INTERNET ARCHIVE WayBackMachine

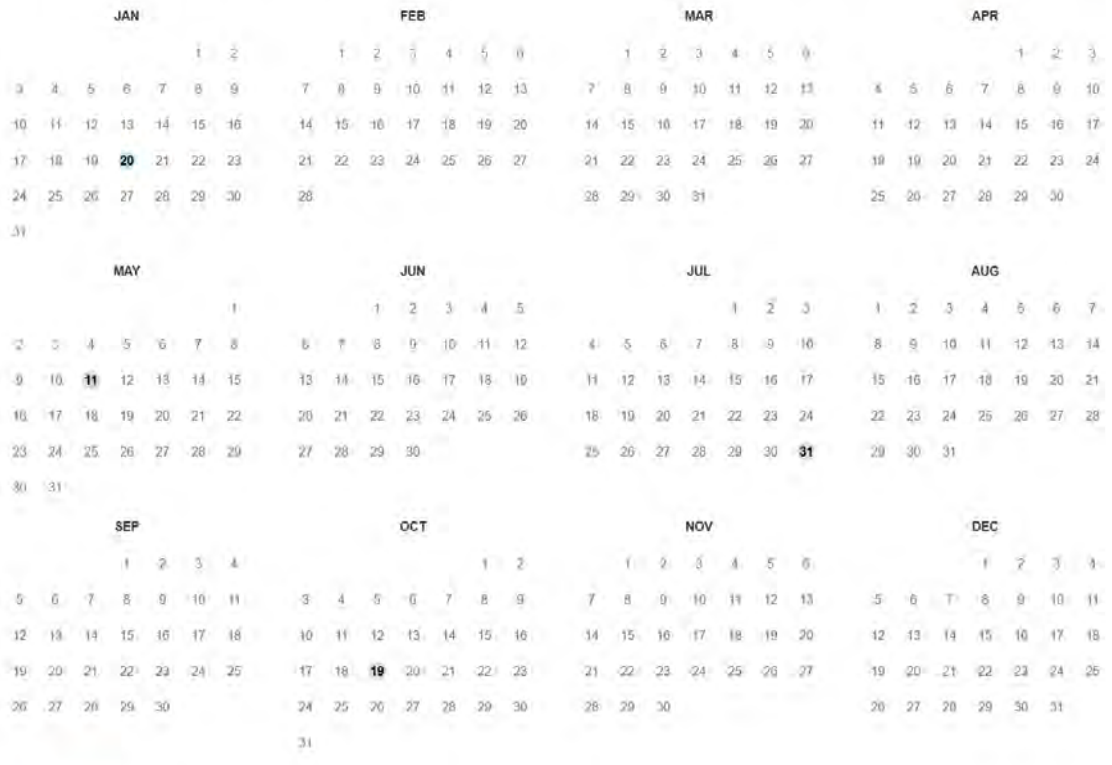
Explore more than 1 trillion web pages saved over time

https://guanlongzhao.github.io/

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **19 times** between November 9, 2020 and August 7, 2025.





Note

This calendar view maps the number of times **https://guanlongzhao.github.io/** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20211019060648/https://guanlongzhao.github.io/ October 28, 2025
https://guanlongzhao.github.io/ Go JUL OCT 19 JAN
19 captures 9 Nov 2020 - 7 Aug 2025 2020 2021 2022 About this capture

Guanlong Zhao, Ph.D. 赵冠龙

About

I am a speech researcher and engineer. My research interests broadly focus on the intersection of speech modification and speech perception. I am also interested in speech synthesis (e.g., TTS) and speech recognition (E2E models in particular). Specifically, I spent five years working on topics in accent and voice conversion during my Ph.D. training.

If you would like to reach out to me, my email address is *zhao at aggienetwork dot com*. How do I pronounce my name? In PinYin, it is written as Guàn-Lóng Zhào; the tones are fourth, second, and fourth. Mapping to American English phonemes, it roughly sounds like Guan-Loan Chao. 🌈 Cheers!

Education

Ph.D. in Computer Science, Texas A&M; University, 2020

- Dissertation: Foreign accent conversion with neural acoustic modeling
- Advisor: Dr. Ricardo Gutierrez-Osuna

B.S. in Applied Physics (minor in Computer Science), University of Science and Technology of China, 2015

Work Experience

Software Engineer @ Google (Speech), August 2020–Present

Research and Teaching Assistant @ Texas A&M; University (Department of Computer Science and Engineering), September 2015–May 2020

Software Engineering Intern @ Google (Geo Machine Perception), May–August 2019

Software Engineering Intern @ Google (Speech), June–August 2018

Publications

Journal Articles

1. S. Ding, **G. Zhao**, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, 2021. (Accepted) demo
2. **G. Zhao**, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2367–2381, 2021. pdf demo
3. I. Lučić Rehman, A. Silpachai, J. Levis, **G. Zhao**, and R. Gutierrez-Osuna, "The English pronunciation of Arabic speakers: A data-driven approach to segmental error identification," *Language Teaching Research*, 2020. pdf summary
4. **G. Zhao** and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649–1660, 2019. pdf code demo
5. S. Ding, **G. Zhao**, C. Liberatore, and R. Gutierrez-Osuna, "Learning structured sparse representations for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 343–354, 2019. pdf demo
6. S. Ding, C. Liberatore, S. Sonsaat, I. Lučić Rehman, A. Silpachai, **G. Zhao**, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "Golden speaker builder–An interactive tool for pronunciation training," *Speech Communication*, vol. 115, pp. 51–66, 2019. pdf code demo

Conference Proceedings

7. A. Hair, **G. Zhao**, B. Ahmed, K. Ballard, and R. Gutierrez-Osuna, "Assessing posterior-based mispronunciation detection on field-collected recordings from child speech therapy sessions," in *Interspeech*, 2021. pp. 2936–2940. pdf
8. A. Silpachai, I. Lučić Rehman, T. A. Barriuso, J. Levis, E. Chukharev-Khudilaynen, **G. Zhao**, and R. Gutierrez-Osuna, "Effects of voice type and task on L2 learners' awareness of pronunciation errors," in *Interspeech*, 2021. pp. 1952–1956. pdf
9. S. Ding, **G. Zhao**, and R. Gutierrez-Osuna, "Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition," in *Interspeech*, 2020. pp. 776–780. pdf code demo video
10. A. Das, **G. Zhao**, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Understanding the effect of voice quality and accent on talker similarity," in *Interspeech*, 2020. pp. 1763–1767. pdf video
11. **G. Zhao**, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Interspeech*, 2019,

pp. 2843–2847. pdf code demo slides

12. **G. Zhao**, S. Sonsaat, A. Silpachai, I. Lučić Rehman, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," in *Interspeech*, 2018, pp. 2783–2787. pdf data code slides
13. S. Ding, **G. Zhao**, C. Liberatore, and R. Gutierrez-Osuna, "Improving sparse representations in exemplar-based voice conversion with a phoneme-selective objective function," in *Interspeech*, 2018, pp. 476–480. pdf
14. C. Liberatore, **G. Zhao**, and R. Gutierrez-Osuna, "Voice conversion through residual warping in a sparse, anchor-based representation of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5284–5288. pdf poster
15. **G. Zhao**, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5314–5318. pdf code demo poster
16. G. Angello, A. B. Manam, **G. Zhao**, and R. Gutierrez-Osuna, "Training behavior of successful tacton-phoneme learners," in *IEEE Haptics Symposium (WIP)*, 2018. pdf
17. **G. Zhao** and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 5525–5529. pdf demo poster

Book Chapter

18. Y. Liu, **G. Zhao**, B. Gong, Y. Li, R. Raj, N. Goel, S. Kesav, S. Gottimukkala, Z. Wang, W. Ren, and D. Tao, "Chapter 10 — Image dehazing: Improved techniques," in *Deep Learning through Sparse and Low-Rank Modeling*: Elsevier, 2019, pp. 251–262. link code

Preprints

19. Y. Liu, **G. Zhao**, B. Gong, Y. Li, R. Raj, N. Goel, S. Kesav, S. Gottimukkala, Z. Wang, W. Ren, and D. Tao, "Improved techniques for learning to dehaze and beyond: A collective study," *arXiv preprint arXiv:1807.00202*, 2018. pdf code
20. Y. Liu and **G. Zhao**, "PAD-Net: A perception-aided single image dehazing network," *arXiv preprint arXiv:1805.03146*, 2018. pdf code
21. A. Datta, **G. Zhao**, B. Ramabhadran, E. Weinstein, "LSTM acoustic models learn to align and pronounce with graphemes," *arXiv preprint arXiv:2008.06121*, 2020. (Work done as an intern at Google NYC during summer 2018.) pdf

Abstracts

22. I. Lučić Rehman, A. Silpachai, J. Levis, **G. Zhao**, and R. Gutierrez-Osuna, "Pronunciation errors — A systematic approach to diagnosis," in *L2 Pronunciation Research Workshop: Bridging the Gap between Research and Practice*, 2019, pp. 23–24. pdf
23. S. Sonsaat, E. Chukharev-Hudilainen, I. Lučić Rehman, A. Silpachai, J. Levis, **G. Zhao**, S. Ding, C. Liberatore, and R. Gutierrez-Osuna, "Golden Speaker Builder, an interactive tool for pronunciation training: User studies," in *6th International Conference on English Pronunciation: Issues & Practices (EPIP6)*, 2019, p. 72. pdf
24. S. Ding, C. Liberatore, **G. Zhao**, S. Sonsaat, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "Golden Speaker Builder: an interactive online tool for L2 learners to build pronunciation models," in *Pronunciation in Second Language Learning and Teaching (PSLLT)*, 2017, pp. 25–26. pdf

Professional Service

Reviewer for:

- IEEE Transactions on Image Processing
- Heliyon
- Language Learning & Technology
- IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP): 2020, 2022
- International Conference on Advances in Signal, Image and Video Processing (SIGNAL): 2020, 2021
- Annual Conference of the International Speech Communication Association (Interspeech): 2019, 2021
- IEEE Automatic Speech Recognition and Understanding Workshop (ASRU): 2021

Teaching

Teaching Assistant: CSCE 482: Senior Capstone Design (Spring 2016)

Honors

- Graduate Student Travel Award (for Interspeech'19), Department of Computer Science and Engineering, Texas A&M; University, 2019
- Graduate Student Presentation Grant (for ICASSP'17), Office of Graduate and Professional Studies, Texas A&M; University, 2017
- Outstanding Graduate Award, University of Science and Technology of China, 2015
- Outstanding Undergraduate Student Scholarship, University of Science and Technology of China, 2011–2014
- Second Prize @ Chinese Chemistry Olympiad (Provincial Level), Chinese Chemical Society, 2010

L2-ARCTIC Corpus

The L2-ARCTIC corpus is a multi-purpose non-native English speech dataset. I took a leading role in this project, where I designed the data collection schemes and the annotation standards. I also spent a lot of time manually cleaning the raw speech recordings and performing quality control to ensure that the speech data and annotations were consistent and high-quality. The recordings were collected at Iowa State University (ISU), led by Dr. John Levis and his students in the Department of English. The annotations were mostly done by Alif Silpachai and Ivana Lučić Rehman. The project spanned around two years. We released the first version at Interspeech 2018, and we continued to add more data to the corpus. Its most recent version is almost 2.4x the size of the initial release.

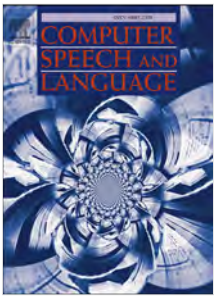
We initially designed the corpus for the accent conversion task, and that was why we chose to use the CMU-ARCTIC prompts in the first place. Along the way, we were also working on some projects related to mispronunciation detection (MPD) and realized that there were limited open-source resources for MPD. We noticed that many of the CMU-ARCTIC sentences were hard for the participants to speak, which, on the one hand, made the recording sessions difficult, on the other hand, elicited rich pronunciation errors in non-native speech productions. As a result, we decided to annotate part of the corpus for phonetic errors. All the sentences we annotated were carefully selected by Dr. Levis to reflect the pronunciation issues that might happen given the speakers' native languages.

I use this corpus in all my publications on accent conversion. I found it is well-suited for the task because it allows me to test the algorithms on speakers with different accents, fluency level, age, and gender. I also use this corpus for MPD research. To the best of my knowledge, this is probably the largest open-source annotated MPD corpus. If you are interested in using the corpus for your projects, you can find access guidelines on its official project site. I would be happy to see it being used in more projects. If you have any questions regarding downloading/using the corpus, please feel free to drop me an email.

Copyright © 2021 Guanlong Zhao. All Rights Reserved.

Appendix DING07

Journal Pre-proof



Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning

Shaojin Ding, Guanlong Zhao, Ricardo Gutierrez-Osuna

PII: S0885-2308(21)00102-9
DOI: https://doi.org/10.1016/j.csl.2021.101302
Reference: YCSLA 101302

To appear in: *Computer Speech & Language*

Received date : 12 April 2021
Revised date : 19 September 2021
Accepted date : 21 September 2021

Please cite this article as: S. Ding, G. Zhao and R. Gutierrez-Osuna, Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language* (2021), doi: https://doi.org/10.1016/j.csl.2021.101302.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Elsevier Ltd. All rights reserved.

Revised manuscript Journal Pre-proof

Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers Using Zero-Shot Learning*

Shaojin Ding*, Guanlong Zhao[1], Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, USA

Abstract

Foreign accent conversion (FAC) aims to create a new voice that has the *voice identity* of a given second-language (L2) speaker but with a native (L1) *accent*. Previous FAC approaches usually require training a separate model for each L2 speaker and, more importantly, generally require considerable speech data from each L2 speaker for training. To address these limitations, we propose Accentron, an approach that can generate accent-converted speech for arbitrary L2 speakers unseen during training. In the proposed approach, we first train a speaker-independent acoustic model on L1 corpora to extract bottleneck features that represent the linguistic content of utterances. Then, we develop a speaker encoder and an accent encoder to generate embedding vectors for the desired voice identity (L2 speaker's) and accent (L1 accent), respectively. Lastly, we use a sequence-to-sequence model to transform bottleneck-features to Mel-spectrograms, conditioned on the L2 speaker embedding and the L1 accent embedding. We conducted experiments on the L2-ARCTIC corpus under two testing conditions: the *standard* FAC setting where test L2 speakers were seen during training, and a *zero-shot* FAC setting where test L2 speakers were unseen during training. Accentron achieves over 27% relative improvement in accentedness ratings compared to two state-of-the-art FAC systems in the standard

*This work was supported by NSF awards 1619212 and 1623750.

*Corresponding author

Email addresses: shjd@tamu.edu (Shaojin Ding), gzhao@tamu.edu (Guanlong Zhao), rgutier@cse.tamu.edu (Ricardo Gutierrez-Osuna)

[1]Zhao is now with Google. This work was done solely with TAMU resources.

IPR2026-00272
Krisp Technologies EX1010 Page 97

Journal Pre-proof

FAC setting. More importantly, our results show that Accentron generalizes to the zero-shot FAC setting with no performance loss. Therefore, in practical use scenarios (e.g., computer-assisted pronunciation training software), Accentron can effectively avoid the need to adapt or retrain the model, which significantly reduces computations and the users' waiting time.

Keywords: Foreign accent conversion,, Speech synthesis, Computer-assisted pronunciation training

1. Introduction

Foreign accent conversion (FAC) [1] aims to create a new voice that has the voice identity of a given L2 speaker and the accent of an L1 speaker. In pronunciation training, FAC can serve as a "golden speaker" for the L2 speaker to practice with: their own voice, but with a native accent [1, 2, 3, 4]. FAC also finds applications in movie dubbing [5], personalized text-to-speech (TTS) synthesis [6, 7], and improving speech recognition performance [8]. A variety of techniques have been proposed to perform FAC, including voice morphing [1, 9, 10], frame pairing [11, 12], articulatory synthesis [13, 14], and sequence-to-sequence (seq2seq) modeling [15, 16]. However, previous FAC approaches have two major limitations. First, they operate in a one-to-one fashion, i.e., they require training a separate model for each pair of L1 and L2 speakers. Second, they need a considerable amount of speech data (∼1,000 utterances) for each L2 speaker. Thus, when using these conventional FAC methods in real-world applications such as pronunciation training, L2 learners need to record a large number of utterances and then wait for a dedicated model to be trained, which can be tedious and demotivating.

To address this issue, we propose Accentron, a zero-shot learning [17] approach to FAC that can synthesize speech for arbitrary L2 speakers who were unseen during training. Accentron consists of four independent models: (1) a speaker-independent acoustic model that captures the linguistic content of an L1 utterance as a sequence of *bottleneck feature vectors*, (2) a speaker encoder

IPR2026-00272
Krisp Technologies EX1010 Page 98

Journal Pre-proof

that captures the voice identity of the L2 speaker, denoted as a *speaker embedding*, (3) an accent encoder that captures the desired L1 accent, denoted as an *accent embedding*, and (4) a seq2seq model that generates a Mel-spectrogram from the sequence of bottleneck features, conditioned on the desired speaker and accent embeddings. These components can be trained independently, at which point the system can generate accent conversions to arbitrary L2 speakers given a few seconds of audio (i.e., enough speech to compute a speaker embedding), without the need to have any model re-training or adaptation process.

To our knowledge, ours is the first work to apply zero-shot learning for the task of FAC. Though zero-shot learning has been used for voice conversion [18, 19, 20, 21] and voice cloning [22, 23], previous studies [18, 19, 20, 21, 22, 23] have focused exclusively on manipulating voice identity, ignoring the speaker's accent, which holds important cues to speaker recognition [24] and speech perception [25, 26, 27, 28]. Incorporating accent into the conversion process requires changes to the conventional encoder-decoder structure of sequence-to-sequence (seq2seq) models for voice conversion. Our encoder takes a sequence of L1 bottleneck feature vectors as the input, and produces a hidden representation sequence. In a conventional voice conversion system [18, 19, 29, 30, 31, 32, 33, 34], this hidden representation sequence is then concatenated with the speaker embedding of the target speaker. In our case, however, the system also concatenates the accent embedding, which is treated as an additional independent and controllable factor during synthesis. The combined bottleneck/speaker/accent embedding is consumed by a decoder coupled with a location-sensitive attention mechanism [35]. During each decoding step, the decoder autoregressively predicts a Mel-spectrogram frame based on the output from the previous decoding step and a context vector produced by the attention mechanism. Finally, the output Mel-spectrogram is converted back into a waveform through either the Griffin-Lim algorithm [36] or a separately trained vocoder (e.g., WaveNet [37], WaveRNN [38]).

We thoroughly evaluated Accentron on the L2-ARCTIC corpus [39]. First, we visualized the speaker and accent embedding distributions for the accent-

IPR2026-00272
Krisp Technologies EX1010 Page 99

Journal Pre-proof

converted speech and natural speech, and the results show that Accentron synthesizes can successfully capture the L2 voice identity along with an L1 accent. Second, we conducted a series of listening tests under two different settings: (1) a *standard* FAC setting, where the test L2 speakers were available during training, and (2) a *zero-shot* FAC setting, which assumes that the test L2 speakers were not available during training. Accentron achieves 27% relative improvement in accentedness while retaining the acoustic quality and voice identity, compared to two state-of-the-art FAC systems in standard FAC settings. In addition, Accentron showed no performance degradation when tested under zero-shot FAC setting.

The manuscript is organized as follows. Section 2 reviews prior approaches to foreign accent conversion, many-to-many voice conversion, and sequence-to-sequence models. Section 3 describes the architecture of Accentron in detail. Section 4 provides the experimental setup, including the corpora and implementation details. Section 5 presents the objective and subjective evaluations of Accentron and analyzes the results. We discuss the implications of the results in Section 6. Lastly, we conclude the findings of this work and point out potential future directions in Section 7.

2. Related work

2.1. Foreign accent conversion

The problem of foreign accent conversion was first formulated by Felps *et al.* [1] as the means to provide implicit feedback in computer assisted pronunciation training. Early approaches [14, 40, 41, 42] involved building an articulatory synthesizer for the L2 speaker. The articulatory synthesizer was trained to map the L2 speaker's articulatory trajectories (e.g., tongue and lip movements) into his or her acoustics features (e.g., Mel Cepstra) using GMMs [14], unit-selection models [40], and DNNs [41]. Once the synthesizer was built, it could be driven with articulatory trajectories from an L1 speaker to synthesize FAC speech. However, these approaches were impractical for pronunciation training since col-

IPR2026-00272
Krisp Technologies EX1010 Page 100

lecting articulatory data is expensive and requires specialized equipment[2]. As a result, later work on FAC has focused on acoustic methods, since they only require recording speech with a microphone. Previous acoustic methods can be grouped into two categories: frame-pairing methods [11, 12] and seq2seq methods [15, 16]. As the name suggests, frame-pairing methods builds a lookup table where L1 and L2 speech frames are paired based on their similarity, and then use a statistical model (e.g., a GMM) to convert from L1 frames to their corresponding L2 frames in the lookup table. Aryal and Gutierrez-Osuna [11] first proposed a technique to pair L1-L2 frames based on their acoustic similarity (in MFCC space), after applying vocal tract length normalization to reduce global differences between the L1 and L2 spectra. Following this, Zhao *et al.* [12] argued that the L1 and L2 frames should be paired based on their linguistic content, and consequently, they used Phonetic-PosteriorGram (PPG) similarity instead of MFCC similarity. More recently, methods based on seq2seq models have been shown to significantly improve synthesis quality. In a previous study [15], we proposed a seq2seq PPG-to-Mel synthesizer for FAC. During training, the system learns a seq2seq model to convert PPGs to Mel-spectra extracted from utterances of an L2 speaker. During inference, the model is driven by PPGs extracted from a reference L1 utterance, which then produces FAC synthesis. In related work, Liu *et al.* [16] proposed a novel recognizer-synthesizer framework to remove the need for a reference L1 utterance. Their system trained a speaker recognizer, a multi-speaker text-to-speech (TTS) model, and an accent-sensitive automatic speech recognition (ASR) system. During inference, they feed L2 Mel-spectra to the ASR system with the corresponding accent, and then feed the output of the ASR system and the L2 speaker embedding to the multi-speaker TTS model to generate accent-converted utterances. These seq2seq model based FAC approaches can convert segmental and prosody features simultaneously, producing syntheses with higher speech naturalness and acoustic

[2]Articulatory measurements can be performed via electromagnetic articulography [40], ultrasound imaging [43], palatography [44], and more recently real-time MRI [45].

IPR2026-00272
Krisp Technologies EX1010 Page 101

Journal Pre-proof

quality.

2.2. Many-to-many voice conversion

Foreign accent conversion is related to the more general problem of voice conversion (VC) [46, 47], which aims to synthesize a voice that has the linguistic content of an utterance from a source speaker and the voice identity of a target speaker. Traditional VC approaches use GMMs [48, 49], sparse representations [50, 51], and DNNs [52, 53, 54, 55, 56, 57, 21, 58] to transform the spectra from a source speaker to that of the target speaker. These methods require training a separate model for each pair of source-target speakers. More recently, several studies have proposed many-to-many VC approaches based on Variational Autoencoders (VAE) [19, 29, 30, 31, 20, 59] and the PPG-to-speech synthesizer [18, 32, 33, 60, 61]. Hsu *et al.* [19, 62] first proposed to use a VAE for many-to-many VC. During training, the encoder learns a speaker-independent latent embedding from input speech signals, and the decoder reconstructs the input speech signals given the latent embedding and the corresponding speaker embedding. During inference, the speaker embedding is replaced with that of a target speaker to produce a VC synthesis. A number of subsequent studies have been conducted to improve performance through various techniques, such as auxiliary classifiers [20], WaveNet vocoder adaption [59], and discrete latent space [29, 63]. Other studies [18, 32, 33] have used a PPG-to-speech synthesizer approach to perform many-to-many VC. The PPG-to-speech synthesizer is a neural network that takes PPGs as an input, and predict spectra conditioned on the speaker embedding of the target speaker. Early many-to-many VC models used one-hot vectors as the speaker embedding due to its simplicity, but recent studies [30, 18, 32, 33] have used learned speaker embeddings (e.g., i-vector [64], d-vector [65]) to generalize to unseen speakers, which make it possible to perform VC in a zero-shot fashion.

2.3. Seq2seq models

The seq2seq model was originally proposed by Sutskever *et al.* [66] for machine translation. The seq2seq model usually has an encoder-decoder archi-

IPR2026-00272
Krisp Technologies EX1010 Page 102

Journal Pre-proof

tecture. The encoder learns a hidden representation sequence from an input sequence, and the decoder learns to autoregressively generate the output sequence given the hidden representation. To capture contextual information and handle length mismatches between the input and output sequences, an attention mechanism [67] is added between the encoder and the decoder. In recent years, there has been growing interest in applying seq2seq model to speech synthesis. Wang *et al.* [68] first proposed a seq2seq based TTS synthesizer (Tacotron), which significantly improved the acoustic quality of the syntheses over previous methods. Following this, Shen *et al.* [69] proposed Tacotron2, which further improved the acoustic quality of Tacotron by using a novel model architecture and a WaveNet vocoder. Jia *et al.* [22] extended Tacotron2 to voice cloning by conditioning a speaker embedding on the decoder. Besides, Biadsy *et al.* [8] and Jia *et al.* [70] also explored the use of seq2seq model in end-to-end speech-to-speech transformation, which avoids the need of multi-stage recognizer-synthesizer framework in related tasks, such as hearing-impaired speech synthesis and speech-to-speech translation. Seq2seq model has also been applied to voice conversion [56, 60, 61] and foreign accent conversion [15, 16], which significantly improved the performance on these tasks compared to conventional approaches.

3. Methods

Accentron consists of four modules: (1) a speaker-independent acoustic model that generates a linguistic representation of an utterance, (2) a speaker encoder that captures the voice identity of the desired speaker, (3) an accent encoder that captures the desired accent, and (4) a seq2seq model that consumes the previous three representations to synthesize Mel-spectrogram for an arbitrary L2 speaker.

The workflow for training Accentron is shown in Figure 1(a). The acoustic model, speaker encoder, and accent encoder are trained separately, and then are used as feature extractors for the seq2seq model. The seq2seq model is trained on a parallel corpus with multiple L1 and L2 speakers, capturing the

IPR2026-00272
Krisp Technologies EX1010 Page 103

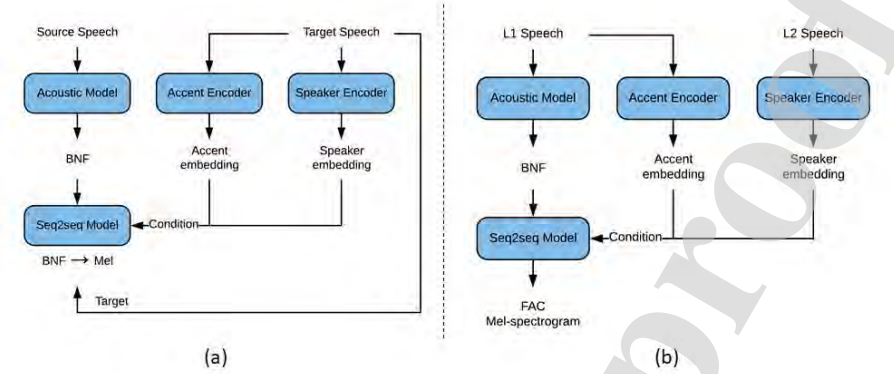


Figure 1: (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

voice characteristics of different speakers and accents. In what follows, we define a *"source"* speaker to be a selected reference L1 speaker, and a *"target"* speaker to be any L1/L2 speaker. To train the seq2seq model, we pair the source speaker with each target speaker. Then, for each pair of speakers, we feed source utterances to the speaker-independent acoustic model to extract bottleneck features (BNFs), which we assume capture only the linguistic content. Next, we feed an utterance from the target speaker to the speaker encoder and the accent encoder, which extract their speaker embedding and accent embedding, respectively. Finally, we train the seq2seq model to convert the source BNFs to the target Mel-spectrogram, conditioning on the target speaker's speaker and accent embeddings.

The workflow during inference is illustrated in Figure 1(b). Accentron requires a source utterance from an L1 speaker and an utterance from the L2 speaker. First, we extract BNFs and accent embedding from the L1 utterance, which encode the desired linguistic content and native accent, and a speaker embedding from the L2 utterance, which encodes the desired voice identity of the L2 speaker. Then, we pass the L1 BNFs, L1 accent embedding, and L2 speaker

IPR2026-00272
Krisp Technologies EX1010 Page 104

Journal Pre-proof

embedding to the seq2seq model, which generates the accent-converted Mel-spectrogram. Finally, the Mel-spectrogram is converted back to a waveform, in our case using a separately trained WaveRNN [38], though other vocoders or the Griffin-Lim algorithm[36] can also be used.

3.1. Speaker-independent acoustic model

To capture the linguistic content of an utterance, we use the output of the last hidden layer of a speaker-independent acoustic model (AM) as BNFs, rather than the output of the final layer of the AM, which represent the PPGs (i.e., the probabilities of each senone/tri-phone). BNFs contain similar linguistic information as PPGs but have much lower dimensionality (e.g., Senone-PPG: 6,024 dimensions; BNF: 256 dimensions), which avoids the need to perform dimensionality reduction in the seq2seq model.

Our AM is based on a Factorized Time Delayed Neural Network (TDNN-F) [71, 72], a feed-forward neural network acting as a sequential classifier. Given an input acoustic feature vector (i.e., 40-dimensional MFCCs), the TDNN-F produces the probabilities of the vector belonging to each senone/triphone (6,024 senones). The TDNN-F takes time-delayed input frames as side inputs to its hidden layers to model long-term temporal dependencies, concatenated with a 100-dimensional i-vector [64] of the corresponding speaker[3]. Additionally, the TDNN-F uses factorized layers with semi-orthogonal constraints as hidden layers and dilated connections between hidden layers, which are more efficient during training and inference than recurrent layers due to their feed-forward nature [71]. The TDNN-F model is composed of five hidden layers. Each of the first four hidden layers has 1,280 neurons, followed by ReLU activation and batch normalization [73], whereas the last hidden layer has 256 neurons, corresponding to the dimensionality of the BNFs. We train the model through a supervised 6,024-way senone classification task. To promote that the AM

[3]As noted by Peddinti *et al.* [72], this allows the model to capture both speaker and environment specific information, which is useful for neural network adaption.

IPR2026-00272
Krisp Technologies EX1010 Page 105

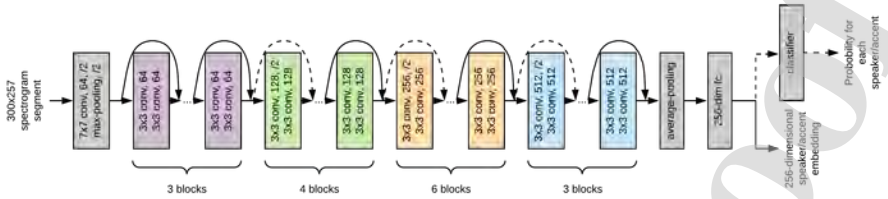


Figure 2: Speaker/accent encoder model architecture. The model is based on ResNet-34 [75]. Each convolution block is illustrated as the kernel size and channel numbers. "/2" means the layer divides the spatial resolution by 2.

produces speaker-independent BNFs, we train the model on speech data from several thousands of speakers (Librispeech corpus [74], 2,484 native English speakers; see Section 4.1).

3.2. Speaker and accent encoders

We use two independent encoders to compute the desired voice identity and accent. The speaker encoder is built as a speaker recognition model, which is trained to determine the identity of a speaker from an input utterance, whereas the accent encoder is based on an accent recognition model, which is trained to recognize accent/dialect patterns (e.g., pronunciation and prosody). For this work, we use a convolutional neural network (CNN) based on ResNet-34 [75] for both the speaker encoder and the accent encoder, following a previous speaker-recognition study [76]. We use the same CNN architecture for both models, so here we provide a detailed workflow only for the speaker encoder; the training and inference workflows of the accent encoder can be derived similarly.

The architecture of the speaker encoder is shown in Figure 2. The model takes 300×257 in time×frequency magnitude spectrogram segments as inputs. The inputs are first fed to a convolution layer containing 64 7×7 kernels with 2×2 stride, followed by a 2×2 max-pooling layer. These layers decrease the spatial resolution of the feature maps, reducing model complexity and improving training speed. On top of them, there are 16 convolution residual blocks,

IPR2026-00272
Krisp Technologies EX1010 Page 106

Journal Pre-proof

which extract more abstract features. Each convolution block consists of two convolution layers with 3×3 kernels. The first convolution layer in each block has 2×2 stride to further decrease the spatial resolution of the feature maps. More importantly, each block has a skip connection as an alternative path to avoid gradient vanishing in a very deep model. The 16 convolution blocks have different numbers of kernels, as highlighted in different colors in Figure 2 (Purple: 64 kernels; Green: 128 kernels; Orange: 256 kernels; Blue: 512 kernels). Next, there is an average pooling layer that produces a 512-dimensional vector, followed by a 256-dimensional fully-connected layer. All the layers are followed by ReLU activations and batch-normalization [73].

The model is trained through a supervised speaker-classification task. During training, a classifier on top of the 256-dimensional fully-connected layer produces the probabilities that the segment belongs to each speaker. The network is then optimized by minimizing the cross-entropy loss between the prediction and the target speaker label. During inference, we discard the final classifier layer and directly use the 256-dimensional bottleneck feature as the segment-wise speaker embedding. To obtain utterance-level speaker embeddings for a speaker that does not appear during training, we divide each test utterance into 300-frame segments with a 150-frame overlap using a sliding window, and then we compute the average of these segment-wise embeddings as the utterance-level speaker embedding (i.e., d-vectors [65]).

3.3. Seq2seq foreign accent conversion model

Our seq2seq model is inspired by the text-to-speech Tacotron2 model [69]. As shown in Figure 3, the seq2seq model has an encoder-decoder architecture. During training, inputs to the network consist of a triplet: (1) a sequence of BNFs from the source (L1) speaker, $\mathbf{x} \in \mathbb{R}^{T_i \times D_{BNF}}$, (2) a speaker embedding of the target (L1 or L2) speaker $\mathbf{s} \in \mathbb{R}^{D_{speaker}}$ extracted from the speaker encoder, and (3) the accent embedding of the target speaker $\mathbf{a} \in \mathbb{R}^{D_{accent}}$ extracted from the accent encoder. T_i is the length of the sequence $\mathbf{x}$. D_{BNF} is the dimensionality of the BNFs (e.g., 256 in this work). $D_{speaker}$ and D_{accent} are the

IPR2026-00272
Krisp Technologies EX1010 Page 107

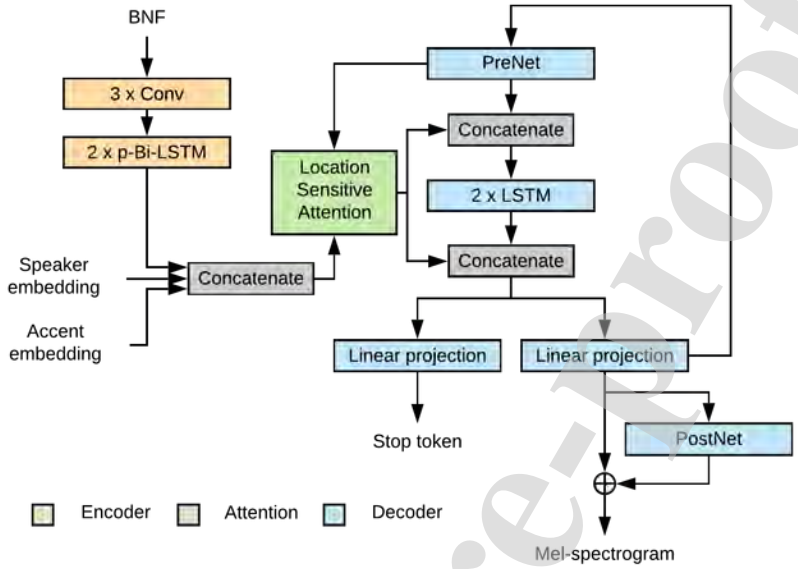


Figure 3: The seq2seq model in Accentron.

dimensionalities of the speaker embedding $\mathbf{s}$ and accent embedding $\mathbf{a}$, respectively (both of them are 256 in this work). The ground-truth target of the model is a sequence of Mel-spectrogram frames $\mathbf{y} \in \mathbb{R}^{T_o \times D_{Mel}}$, where T_o is the length of the sequence and D_{Mel} is the number of Mel-filterbanks (e.g., 80 in this work). First, the encoder accepts a BNF sequence $\mathbf{x}$ and produces a hidden representation $\mathbf{h}$:

$$\mathbf{h} = \text{Encoder}(\mathbf{x}) \tag{1}$$

Then, to condition the decoder on the voice identity and the accent of the target speaker, we concatenate the target speaker's speaker embedding and accent embedding to the hidden representation:

$$\mathbf{h_{concat}} = [\mathbf{h}, \mathbf{s}, \mathbf{a}] \tag{2}$$

Finally, the decoder autoregressively predicts the Mel-spectrogram of the target speech using the attention context computed based on the concatenated hidden

IPR2026-00272
Krisp Technologies EX1010 Page 108

Journal Pre-proof

Table 1: Hyperparameters of the seq2seq foreign accent conversion model.

Block	Component	Parameters
Inputs	BNF	256-dimensional
	Speaker d-vector	256-dimensional
	Accent d-vector	256-dimensional
Encoder	3 × Conv layers	512 5×1 kernel with 1×1 stride; ReLU; batch norm
	2 × p-Bi-LSTM	256 cells per direction;
Attention	Location-sensitive attention	128-dim attention context; 32 31×1 attention conv kernels
Decoder	PreNet	2 × Fully-connected layer ; 256 neurons; ReLU
	2 × LSTM	1024 cells
	Linear (Mel)	1 × Fully-connect layer; 80 units; no activation
	Linear (stop token)	1 × Fully-connect layer; 1 unit; no activation
	PostNet	5 × Conv layers; 512 5×1 kernel with 1×1 stride; tanh; batch norm
Outputs	Mel-spectrogram	80-dimensional
	Stop token	2-dimensional

representation:

$$\hat{\mathbf{y}}^t = \text{Decoder}(\hat{\mathbf{y}}^{t-1}, \mathbf{h_{concat}}) \tag{3}$$

where $\hat{\mathbf{y}}^t$ is the t-th frame of the predicted Mel-spectrogram.

During inference, the inputs to the network are also a triplet: (1) a sequence of BNFs from an L1 speaker, $\mathbf{x}$, (2) a speaker embedding of an L2 speaker, $\mathbf{s_{L2}}$, and (3) an accent embedding of an L1 speaker, $\mathbf{a_{L1}}$. The network produces a hidden representation $\mathbf{h}$, concatenates it with $\mathbf{s_{L2}}$ and $\mathbf{a_{L1}}$, and feeds it to the decoder to produce the predicted FAC Mel-spectrogram. We describe each component in the following subsections. The hyper-parameters of each component are summarized in Table 1.

IPR2026-00272
Krisp Technologies EX1010 Page 109

Journal Pre-proof

3.3.1. Encoder

The encoder converts a BNF sequence to a hidden representation sequence. The original text-to-speech Tacotron2 encoder contains three 1-dimensional convolution layers and one Bidirectional Long Short-Term Memory (Bi-LSTM) layer. However, in our case, the inputs of the seq2seq model are BNF sequences instead of text embeddings, which are usually significantly longer. To capture the high-level phonetic and contextual information in an input BNF sequence, we replace the LSTM layer in the encoder with two pyramidal Bidirectional LSTM (p-Bi-LSTM) layers [77]. Each p-Bi-LSTM reduces the time resolution by a factor of two, and therefore our encoder produces four times shorter hidden representation sequences compared with the input sequences. A convolution layer has 512 kernels with 5×1 shape in time×frequency and 1×1 stride, followed by ReLU activation and batch normalization. Each convolution kernel spans five BNF frames, which models the local context information. A p-Bi-LSTM layer has 256 cells in each direction, followed by ReLU activation and batch normalization, producing a 512-dimensional hidden representation sequence.

3.3.2. Decoder

The decoder is an autoregressive recurrent neural network coupled with a local sensitive attention mechanism [35]. The decoder accepts the concatenated hidden representation sequences as inputs, and produces an 80-dimensional Mel-spectrogram as the prediction of the L2 speech. During each decoding step, the predicted Mel-spectrogram frame from the previous step is first passed into a pre-net that has two 256-dimensional fully-connected layers with ReLU activations:

$$\mathbf{q}^t = \text{PreNet}(\hat{\mathbf{y}}^{t-1}) \tag{4}$$

The pre-net acts as an information bottleneck, which is essential for learning attentions [69]. Next, the location-sensitive attention mechanism computes a 128-dimensional attention context vector $\mathbf{c}^t$ based on the pre-net output, the

IPR2026-00272
Krisp Technologies EX1010 Page 110

Journal Pre-proof

concatenated hidden representations, and the attention context from the previous step:

$$\mathbf{c}^t = \text{Attention}(\mathbf{q}^t, \mathbf{h}_{\mathbf{concat}}, \mathbf{c}^{t-1}) \tag{5}$$

Following this, the pre-net output is concatenated with the context vector and fed to two unidirectional LSTM layers with 256 cells. Then, the output of the second LSTM layer is concatenated again with the context vector $\mathbf{c}^t$ and passed through an 80-unit linear layer to make a prediction of the 80-dimensional L2 Mel-spectrogram frame:

$$\hat{\mathbf{y}}^t_{\mathbf{pre}} = \text{Linear}(\text{LSTM}(\mathbf{q}^t, \mathbf{c}^t), \mathbf{c}^t) \tag{6}$$

More importantly, the network also predicts if the generating process should stop at the current decoding step at the same time, i.e., a stop token $\hat{\mathbf{t}} \in \mathbb{R}^{T_o}$. Finally, to incorporate the spectral residual and improve synthesis quality, the predicted Mel-spectrogram is passed through a post-net consisting of five convolution layers to predict the residual. Each of these layers has 512 kernels with 5×1 shape and 1×1 stride, followed by *tanh* activation and batch normalization. The residual is added back to the original prediction to form the final prediction:

$$\hat{\mathbf{y}}^t = \hat{\mathbf{y}}^t_{\mathbf{pre}} + \text{PostNet}(\hat{\mathbf{y}}^t_{\mathbf{pre}}) \tag{7}$$

The model is optimized by minimizing the Euclidean distance between the target Mel-spectrogram and the prediction before/after the post-net. We also jointly minimize an extra cross-entropy loss to learn the stop token for model inference.

$$L = ||\hat{\mathbf{y}}^t_{\mathbf{pre}} - \mathbf{y}||_2^2 + ||\hat{\mathbf{y}}^t - \mathbf{y}||_2^2 + \lambda \text{CrossEntropy}(\hat{\mathbf{t}}, \mathbf{t}) \tag{8}$$

where $||\cdot||_2^2$ is the Euclidean distance; $\hat{\mathbf{t}}$ is the sequence of predicted stop tokens, and $\mathbf{t}$ is the sequence of target stop tokens; λ is the weight controlling the relative importance of the cross-entropy loss. Additionally, we use the teacher-forcing procedure during training by feeding in the correct output instead of

IPR2026-00272
Krisp Technologies EX1010 Page 111

Journal Pre-proof

the predicted output on the decoder side, which has been shown to improve the efficiency of the model training [78].

4. Experimental setup

4.1. Acoustic model

We trained the TDNN-F acoustic model using the Librispeech corpus [74], which consists of 960 hours of 16 kHz audiobook speech data produced by 2,484 native English speakers, the majority being American English. The training set consists of two "clean" subsets and a "noisy" subset[4]. We used both sets in training to ensure that the BNF was speaker-independent. In addition, we used a subset (200 hours) of the training set to train the i-vector extractor. We implemented the training following the official "tdnn_1d" recipe of the TDNN-F model in Kaldi[5]. The trained model achieves 3.76% word error rate (WER) on Librispeech's test-clean subset and 8.92% WER on the test-other subset.

4.2. Speaker encoder

We trained the speaker encoder using the VoxCeleb1 corpus [79], which contains 153,516 utterances of 16 kHz speech produced by 1,251 speakers. Specifically, we used the training set from the official identification split, which is comprised of 138,316 utterances (~300 hours) from these speakers. We extracted 257-dimensional magnitude spectrograms with a 25ms window and 10ms shift. We trained the model on a single NVIDIA Tesla V100 GPU with a batch size of 128. We used Adam Optimizer with an initial learning rate of 10^{-2}, which was annealed down to zero following a cosine schedule [80]. The trained model achieves 81.34% Top-1 accuracy and 94.49% Top-5 accuracy on the official VoxCeleb1 identification testing set.

[4] We use the term "noisy" subset to refer to the test-other test subset in the LibriSpeech corpus. LibriSpeech has two test subsets: test-clean and test-other. The recordings in test-other subset have significantly higher background noise level than the test-clean subset.

[5] https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/local/chain/tuning/run_tdnn_1d.sh

IPR2026-00272
Krisp Technologies EX1010 Page 112

Journal Pre-proof

4.3. Accent encoder

We trained the accent encoder using the Speech Accent Archive dataset [81], which consists of recordings of the "Please call Stella" paragraph [81] produced by speakers in 386 native and non-native English accents. For most of the accents, however, the number of speakers is limited, which may degrade the performance of the accent encoder. To address this issue, we selected a subset of accents where each accent has at least 30 speakers. The resulting subset has 18 accents[6], with an average of 107 speakers in each accent. The total length of the selected subset is around 16 hours. We randomly selected 90% utterances from each accent as the training set and used the remaining 10% utterances as the testing set. The audio waveforms in the original dataset have 8 kHz sampling rate. To match it with other modules, we resample them to 16 kHz. Other configurations were the same as that for speaker recognition. Our trained model achieves 79.36% Top-1 accuracy and 95.42% Top-5 accuracy on the testing set.

4.4. Seq2seq foreign accent conversion model

To evaluate the proposed approach, we conducted experiments with the ARCTIC [82] and L2-ARCTIC corpora [39]. We used four native English speakers from ARCTIC (BDL, RMS, SLT, CLB) and all 24 non-native English speakers from L2-ARCTIC. For each speaker, we divided their utterances into three subsets: a training set of 1,032 utterances (∼1 hour of speech), a validation set of 50 utterances, and a testing set of 50 utterances. During training, we set BDL as the source speaker and paired it with all 28 speakers, including himself. During inference, we used both BDL (male) and CLB (female; used as an unseen L1 speaker for the zero-shot FAC setting) as the native reference speakers, and we performed FAC on four L2 speakers whose first languages were different:

[6] These 18 accents were Arabic, Cantonese, Dutch, English, Farsi, French, German, Hindi, Italian, Japanese, Korean, Mandarin, Polish, Portuguese, Russian, Spanish, Turkish, and Vietnamese.

IPR2026-00272
Krisp Technologies EX1010 Page 113

Journal Pre-proof

NJS (Spanish, female), TXHC (Mandarin, male), YKWK (Korean, male), and ZHAA (Arabic, female).

The original L2-ARCTIC audio waveforms have a 44.1 kHz sampling rate, so we resampled them to 16 kHz to match the ARCTIC recordings. We extracted 80-dimensional Mel-spectrogram with a 25ms window and 10ms shift. Following the same frame shift, we extracted BNFs for each utterance using the acoustic model (Section 3.1). In addition, we extracted utterance-level speaker and accent d-vectors from the speaker encoder and accent encoder, respectively. We implemented the model using TensorFlow [83] and trained it on a single NVIDIA Tesla V100 GPU. The hyperparameter λ (eq. 8) in the loss function was set to 0.005 empirically. We set the batch size to 48, and we used an Adam Optimizer with an initial learning rate of 10^{-3}, which was then annealed down to 10^{-5} following exponential scheduling. The model converged after 200,000 steps, and the entire training time was around 100 hours.[7] During model inference, we used a separately trained speaker-independent WaveRNN vocoder to invert the Mel-spectrogram back to the time-domain waveform. We trained the WaveRNN model on the Librispeech dataset.

5. Results

We have validated Accentron through a series of objective and subjective evaluations. For the objective evaluation, we used t-distributed stochastic neighbor embedding (t-SNE) [84] to analyze the distribution of speaker embeddings and accent embeddings for the original speech and the accent-converted speech. This allowed us to examine whether the two encoder networks can decouple speaker identity and accent. For the first of two sets of subjective evaluations, we tested the system when the test L1 and L2 speakers were seen during training (*standard* FAC setting) and compared it against two state-of-the-art FAC systems [15, 16]. We also tested whether our system could be used in the reverse

[7] Audio samples from this work can be found at `https://shaojinding.github.io/samples/accentron/`. We intend to open-source our code after this work has been peer-reviewed.

IPR2026-00272
Krisp Technologies EX1010 Page 114

Journal Pre-proof

direction, i.e., to impart a non-native accent to a native speaker's voice. In the second set of subjective evaluation, we explored the effectiveness of Accentron when the test L1 and L2 speakers were unseen during training (*zero-shot* FAC setting). Further, we characterized the performance of Accentron as a function of the number of available L2 test utterances during inference (i.e., which are used to extract the L2 speaker's voice identity footprint), and we compared it against a system that uses these utterances to fine-tune a pre-trained FAC system to understand the tradeoff between using L2 test utterances to compute a speaker embedding (zero-shot learning) and using them to refine a pre-trained model (fine-tuning).

5.1. Objective evaluation: speaker and accent embedding spaces

We used t-SNE [84] to visualize the speaker and accent embedding spaces[8], which helped provide a qualitative and intuitive explanation of how our proposed system operates. First, we visualized the speaker and accent embeddings of 20 FAC utterances for TXHC, a male Mandarin speaker. We used the system in the zero-shot condition when both L1 and L2 speakers were unseen (see Section 5.3.1) to generate the syntheses, since it acts as a performance lower-bound for all our systems, and it can also provide insights for zero-shot FAC. We also plotted the embeddings of natural speech from ten L1 and L2 speakers as references (20 utterances for each speaker). Results are shown in Figure 4, where we use colors and shapes to represent speaker and accent, respectively (e.g., orange denotes speaker BDL, and diamond denotes L1 accent). As shown, the speaker embeddings of utterances from the same speaker form a cluster, and the boundary between different clusters are clear. Similarly, accent embeddings from speakers with the same accent form a cluster. These two results indicate that the speaker and accent encoders are operating as expected. In terms of

[8]The axes in a t-SNE plot have no actual meaning. t-SNE projects high dimensional vectors into a lower dimensional plane (e.g., 2-D), such that similar points in the projection are also similar in the original space, and vice versa.

IPR2026-00272
Krisp Technologies EX1010 Page 115

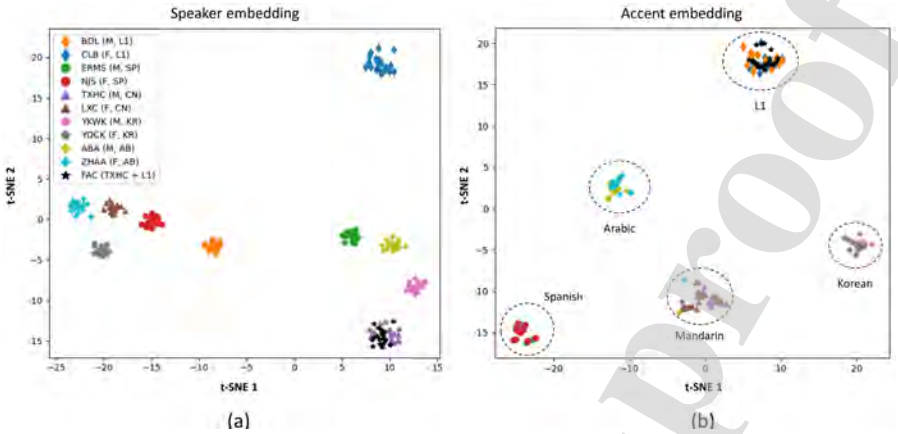


Figure 4: Speaker and accent embedding visualization of FAC syntheses for TXHC using t-SNE. (a): speaker embedding; (b): accent embedding. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent.

the FAC utterances, their speaker embeddings are distributed in the cluster of TXHC, which indicate that Accentron can generate speech that matches the voice identity of the target speaker, and the accent embeddings lie in a cluster of L1 speakers (BDL and CLB), which also indicate the Accentron can generate the desired accent. Together, these visualizations indicate that Accentron can successfully generate speech with the same voice identity as TXHC but with a native accent.

We also conducted t-SNE visualizations on a "reverse" FAC task [12, 85], where the goal was to synthesize speech with the voice identity of a given L1 speaker but with an L2 accent. This is a straightforward process in Accentron, since we only need to change the inputs of the seq2seq model to use an L2 accent embedding and an L1 speaker embedding during inference. For these visualizations, we synthesized a voice that had the voice identity of CLB but with an L2 (Mandarin) accent. As shown in Figure 5, the speaker embeddings of the reverse FAC syntheses lie in the same cluster as CLB utterances, whereas the

IPR2026-00272
Krisp Technologies EX1010 Page 116

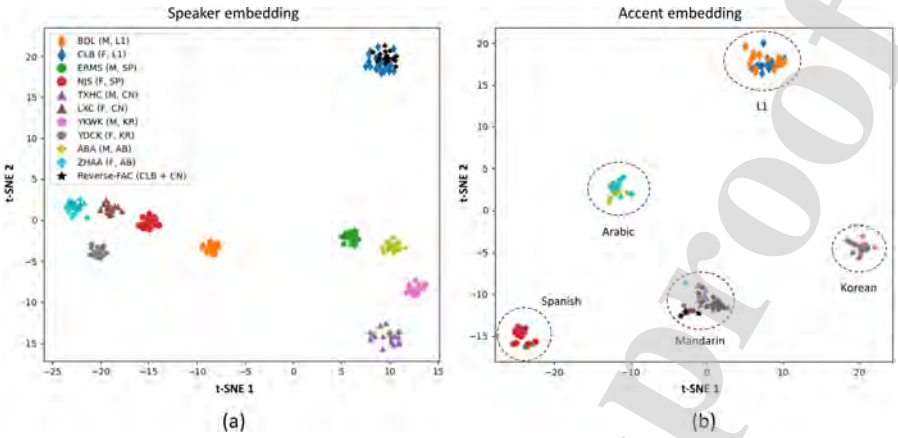


Figure 5: Speaker and accent embedding visualization of *reverse* FAC syntheses (CLB with a Mandarin accent) using t-SNE. (a): speaker embedding; (b): accent embedding. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent.

accent embeddings lie in the same cluster as the two Mandarin speakers (TXHC and LXC), indicating that the reverse FAC syntheses have a voice identity of CLB and a Mandarin accent.

Another interesting question that we would like to investigate here is whether the speaker embedding also carries accent cues, as these two aspects are closely related in the recognition of speakers [25, 26, 69, 28]. If the speaker embedding is entangled with accent information, then there is the risk that Accentron would generate speech with an incorrect accent that is introduced by speaker embeddings. To examine this potential issue, we generated t-SNE visualizations of speaker embeddings for natural utterances from 16 speakers with 4 different accents in L2-ARCTIC corpus (See Figure 6 for details of the speakers and accents). Results in Figure 6 show no adjacency patterns among speakers sharing the same accent. Instead, t-SNE shows a distinct separation among speakers based on gender. These results suggest that speaker embedding mainly encodes information such as voice quality/timbre and pitch, and no evidence of potential

IPR2026-00272
Krisp Technologies EX1010 Page 117

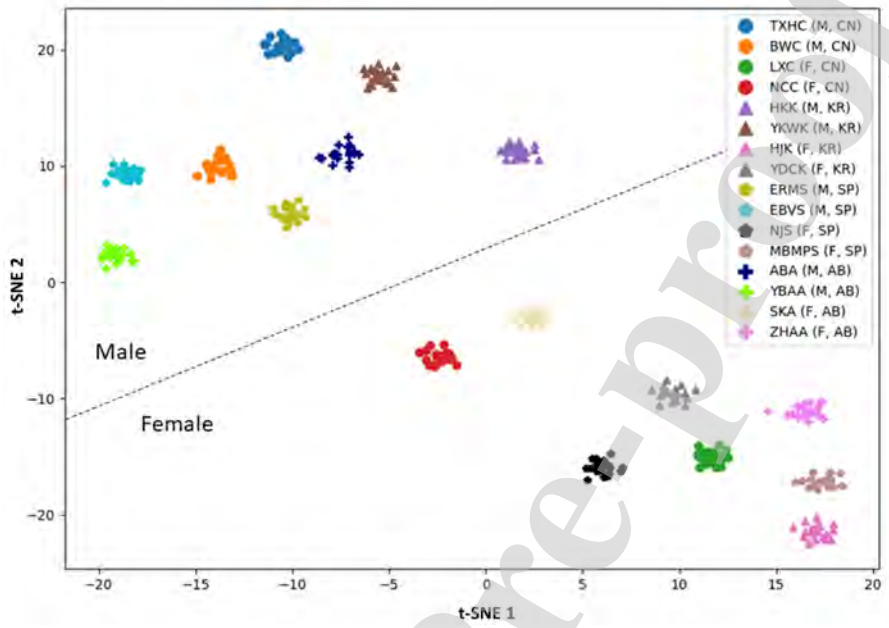


Figure 6: t-SNE visualization of the speaker embeddings from 16 speakers with 4 accents. Colors and shapes represent speaker and accent, respectively. . Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent.

entanglement between speaker and accent cues.

5.2. Subjective evaluations under standard FAC setting

We also evaluated Accentron through two sets of perceptual experiments. In the first set (this sub-section), we evaluated the system under the standard FAC setting, i.e., the test L1 and L2 speakers were seen during training. In the second set (Section 5.3), we evaluated the system under the zero-shot FAC setting, where the L1 speaker and/or the L2 speaker were unseen during training.

For the standard FAC setting, we used the union of the training sets of all 28 speakers to train the system. During inference, we used BDL as the L1 speaker, who then had been "seen" during training. First, we compared the proposed approach against two state-of-the-art FAC approaches:

IPR2026-00272
Krisp Technologies EX1010 Page 118

Journal Pre-proof

- **Baseline1:** the FAC system proposed by Zhao *et al.* [15], a one-to-one FAC approach based on seq2seq model. This baseline system trains a seq2seq PPG-to-speech synthesizer for each L2 speaker, and drives the synthesizer with PPGs extracted from an L1 speaker. As such, the baseline system requires training separate models for each L2 speaker.

- **Baseline2:** the system proposed by Liu *et al.* [16], a reference-free many-to-many FAC approach based on a novel recognizer-synthesizer architecture. The system is trained on 105 speakers from CSTR VCTK dataset [86]. Audio samples were produced by feeding the test utterances through their system, which is provided as a courtesy by Liu *et al.* Due to the implementation differences between the systems, we conducted two post-processing steps to ensure a fair comparison. First, as the accent conversion model of Liu *et al.* was trained on VCTK speakers, the stop-token predictions on L2-ARCTIC test utterances are not robust, occasionally resulting in a few seconds of white noise at the end of speech in accent conversion syntheses. To solve this issue, we manually removed the trailing white noises in these test utterances. Second, we resampled the syntheses of Liu system from 22.05 kHz to 16 kHz to make the sampling rate be consistent with other systems.

We conducted listening tests through Amazon Mechanical Turk[9] to rate three perceptual attributes of the synthesized speech:

- **Accentedness:** : The test asked participants to rate the degree of foreign accentedness of each utterance in a 9-point scale (1-no foreign accent; 9-very strong foreign accent), which is commonly used in the pronunciation literature [87]. Participants were told that the native accent in this task was General American.

- **Acoustic quality:** The test asked participants to rate the acoustic qual-

[9] https://www.mturk.com

IPR2026-00272
Krisp Technologies EX1010 Page 119

Journal Pre-proof

ity of each utterance through a standard 5-point Mean Opinion Score (MOS; 1-bad, 5-excellent).

- **Voice identity:** The test asked participants to rate the voice similarity between the FAC syntheses and the original L2 speech through a 14-point Voice Similarity Score (VSS) [88]. For each FAC-L2 utterances pair, participants were required to decide whether the two utterances were from the same speaker and then rate their confidence in the decision on a 7-point scale (1: not confident at all; 3: somewhat confident; 5: quite a bit confident; 7: extremely confident). The VSS was computed by collapsing the above two fields into a 14-point scale: -7 (definitely different speakers) to +7 (definitely the same speaker). To minimize the influence of accent, the two utterances had different linguistic content and were played in *reverse*, following [1].

Instructions were given in each test to help participants focus on the target speech attribute. For example, in the accentedness test, participants were asked to *"Try to ignore the audio quality (noise, distortions). Please focus only on the speaker's accent, for example, their pronunciation, rhythm, and fluency"*. Test utterances were randomly selected from our test set, and the presentation order was counter-balanced. Additionally, in each listening test, we included five calibration utterances to detect if participants were cheating. We excluded ratings of the calibration utterances from the data analysis [89]. We recruited 18 participants for each listening test. All participants resided in the United States, and they passed a qualification test that asked them to identify different regional accents in the United States.

5.2.1. Comparison to the baseline system

Accentedness. Participants rated 20 utterances per system (5 utterances for each test L2 speaker). These utterances shared the same linguistic content across all the systems to ensure a fair comparison. Additionally, participants also rated the same set of sentences from the original L1 and L2 speakers as a

IPR2026-00272
Krisp Technologies EX1010 Page 120

Journal Pre-proof

Table 2: Accentedness (1-no foreign accent, 9-very strong foreign accent) results and acoustic quality (1-bad, 5-excellent) results under standard FAC setting. All the results are shown as average ± 95% confidence intervals.

System	Accentedness	Acoustic quality
Original L2	7.11 ± 0.21	3.67 ± 0.28
Original L1	1.06 ± 0.12	4.90 ± 0.10
Baseline1	4.63 ± 0.10	3.47 ± 0.14
Baseline2	6.25 ± 0.39	3.12 ± 0.13
Proposed	**3.39 ± 0.14**	**3.51 ± 0.15**

reference. Results are shown in the first column of Table 2. Accentron received significantly lower ratings of foreign accentedness (3.39) than the original L2 utterance (7.11), though not as low as those of the original L1 utterance (1.06). These results suggest that our proposed seq2seq FAC model can effectively reduce foreign accentedness from the L2 speech. Accentron also outperformed the two baseline systems (Baseline1: 4.63, 27% relative improvement, $p \ll 0.001$; Baseline2: 6.25, 46% relative improvement, $p \ll 0.001$).

Acoustic quality. As shown in the second column of Table 2, the proposed method achieved an MOS of 3.51, which is comparable to Baseline1 (3.47, $p > 0.5$) but significantly higher than Baseline2 (3.12, 13% relative improvement, $p \ll 0.001$). The original L1 speech received the highest MOS (4.90), followed by the original L2 speech (3.67). Note that the MOS ratings of Accentron are closer to those of the original L2 speech than to the original L1 speech, possibly due to native listeners confounding acoustic quality with intelligibility [1], and therefore, they may be influenced by the intelligibility and provide lower ratings for non-native speech. Thus, the proposed system achieves similar acoustic quality as the baseline systems (or better), but unlike them does not require training a separate model for each new test L2 speaker.

Voice identity. Participants rated 20 pairs of utterances per system (5 pairs of utterances for each test L2 speaker). Each pair consisted of a FAC utterance and an utterance randomly selected from the L2 speaker. Voice identity

IPR2026-00272
Krisp Technologies EX1010 Page 121

Journal Pre-proof

Table 3: Voice identity results under standard FAC setting. Voice Similarity Score ranges from -7 (definitely different speakers) to +7 (definitely the same speaker). All the results are shown as average ± 95% confidence intervals.

System	Voice Similarity
Baseline1	5.05 ± 0.28
Baseline2	3.81 ± 0.29
Proposed (All pairs)	**5.05 ± 0.31**
Proposed (Intra-gender)	5.29 ± 0.30
Proposed (Inter-gender)	4.80 ± 0.35

results are shown in Table 3. Accentron achieved a 5.05 VSS, indicating that participants were "quite confident" that the FAC syntheses and the L2 speech were produced by the same speaker. These ratings are comparable to those of Baseline1 (5.05 VSS, $p > 0.5$) and significantly higher than those of Baseline2 (3.81 VSS, 33% relative improvement, $p \ll 0.001$). It is worth noting that the L1 speaker in Baseline1 had the same gender as the L2 speaker, whereas Accentron used the same L1 speaker for all L2 speakers. As a result, syntheses from Accentron included both intra (same)-gender FAC pairs and inter (different)-gender FAC pairs, the latter being more challenging due to the differences in prosody and pitch range. Although the VSS on inter-gender pairs (4.80) was lower than that on intra-gender pairs (5.29) and Baseline1, the difference was not significant ($p = 0.14$). These results suggest that the proposed system can generate FAC syntheses that greatly resemble the voice identity of L2 speakers of any gender, using a canonical reference L1 speaker.

5.2.2. Performance on reverse FAC

To evaluate Accentron on the reverse FAC task, we synthesized testing utterances using the accent embeddings from NJS, TXHC, YKWK, and ZHAA, and the speaker embedding from BDL. Table 4 shows the accentedness, acoustic quality, and voice identity results of the reverse FAC evaluation. Accentron received a 5.58 accentedness rating, much closer to that of the original L2 speech

IPR2026-00272
Krisp Technologies EX1010 Page 122

Journal Pre-proof

Table 4: Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (-7-definitely different speakers, +7-definitely the same speaker) of *reverse* foreign accent conversion under standard condition. All the results are shown as average ± 95% confidence intervals.

System	Accentedness	Acoustic quality	Voice Similarity (All pairs)	Voice Similarity (Intra-gender)	Voice Similarity (Inter-gender)
Proposed	5.58 ± 0.35	3.24 ± 0.17	4.91 ± 0.34	5.11 ± 0.35	4.71 ± 0.32

(7.11) than to the original L1 speech (1.06), indicating that our approach was able to impart an L2 accent to utterances from an L1 speaker. Accentron also received a 3.24 MOS, significantly lower ($p = 0.02$) than the MOS of the "direct" FAC syntheses (3.51), a result that is likely due to the correlation between acoustic quality and intelligibility –see Section 5.2.1. Finally, Accentron received a 4.91 VSS, indicating that raters were "quite confident" that the reverse FAC syntheses and the L1 speech were produced by the same speaker; we found no significant differences between the voice identity ratings of reverse and direct FAC syntheses. Thus, we can conclude that Accentron can also operate in the reverse direction, generating non-native utterances with the voice identity of a native speaker.

5.3. Subjective evaluations under zero-shot FAC setting

In the second set of subjective evaluations, we evaluated the proposed system under the zero-shot FAC setting, where the L1 speaker and/or the L2 speaker were unseen during training. The zero-shot FAC setting is appealing for real-world applications since it requires minimal data from the target speaker. First, we compared the performance of Accentron when using seen/unseen L1 or L2 speakers during inference. Then, we characterize its performance as a function of the number of available L2 utterances.

5.3.1. Comparing different conditions in zero-shot foreign accent conversion

We considered four different conditions in this experiment, as summarized in Table 5. In condition SS, the L1 speaker and the L2 speaker were both

IPR2026-00272
Krisp Technologies EX1010 Page 123

Journal Pre-proof

Table 5: The four conditions in zero-shot FAC experiment.

		L1 speaker	
		Seen	**Unseen**
L2 speaker	**Seen**	Condition SS	Condition US
	Unseen	Condition SU	Condition UU

seen during training. Note this condition is the same as the system evaluated in Section 5.2.1, so it serves as a best-case scenario. In condition US, the L1 speaker was unseen during training, and the L2 speaker was seen during training. In condition SU, the L1 speaker was seen during training, and the L2 speaker was unseen during training. Finally, in condition UU, the L1 speaker and the L2 speaker were both unseen during training. Thus, condition UU was the most challenging of the four.

To ensure that the test speakers were unseen during training, we trained four models using different training sets. In Condition SS, we used the same model as in the standard FAC condition. In Condition US, we excluded CLB from the training set and used it as the test L1 speaker. In Condition SU, we excluded the four test L2 speakers from the training set. In Condition UU, we excluded the four test L2 speakers and CLB from the training set, and we also used CLB as the test L1 speaker. For unseen L1/L2 speakers, we used the 50 utterances from the test set to generate the accent/speaker embedding. As before, we conducted three types of listening tests through Amazon Mechanical Turk to rate the accentedness, acoustic quality, and voice similarity of the synthesized speech. In addition, we kept the participants the same as those in the first experiment, so that the results are comparable between different experiments (e.g., participants in the accentedness test for the two experiments were the same).

Results from the accentedness, acoustic quality, and voice identity tests are shown in Table 6. We found no statistically significant differences between condition SS (best-case scenario) and the three more challenging conditions

IPR2026-00272
Krisp Technologies EX1010 Page 124

Journal Pre-proof

Table 6: Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (-7-definitely different speakers, +7-definitely the same speaker) under zero-shot FAC condition. All the results are shown as average ± 95% confidence intervals.

System	Accentedness	Acoustic quality	Voice Similarity
Condition SS	3.39 ± 0.14	3.51 ± 0.15	5.05 ± 0.28
Condition US	3.33 ± 0.26	3.47 ± 0.13	4.99 ± 0.30
Condition SU	3.35 ± 0.25	3.50 ± 0.12	4.92 ± 0.28
Condition UU	3.30 ± 0.26	3.43 ± 0.12	4.59 ± 0.34

(US, SU, UU); $p > 0.5$ in all cases. This result suggests that Accentron has no trouble generalizing to unseen L1 or (and) L2 speakers during inference without any degradation in accentedness, acoustic quality, and voice identity.

5.3.2. Influence of the number of available L2 utterances

For practical FAC applications, it is important to understand the minimum amount of data needed from a target speaker. Requiring L2 learners to record a large amount of speech before they can hear their "golden speaker" voice can be tedious and demotivating. On the other hand, training the system with insufficient speech data might significantly degrade synthesis quality. To characterize the data requirements of Accentron under zero-shot FAC condition, we measured its performance of the UU codition (unseen L1 speaker and unseen L2 speaker) as a function of the number of available L2 utterances. We used the UU condition since it is the most flexible for real-world applications, and also the most challenging, which provides a lower bound of performance. For these experiments, we used 50 test L2 utterances to produce the speaker embedding during inference, and reduced the number from 50 to 1 (N = 50, 20, 10, 5, 1) and re-evaluated system performance. Results are shown in Table 7. Reducing the number of utterances from 50 to 1 has no impact on any of the three perceptual measures ($p > 0.5$ in all cases). These results indicate that as little as a single utterance (∼3 seconds of speech) is sufficient to generate accent conversions for a new unseen L2 speaker, with no impact on performance.

IPR2026-00272
Krisp Technologies EX1010 Page 125

Journal Pre-proof

Table 7: Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (-7-definitely different speakers, +7-definitely the same speaker) with different numbers of available L2 (non-native) utterances during inference. All the results are shown as average ± 95% confidence intervals.

#L2 utterances	Accentedness		Acoustic quality		Voice Similarity	
	Proposed	**Fine-tuned**	**Proposed**	**Fine-tuned**	**Proposed**	**Fine-tuned**
50	3.30 ± 0.26	3.03 ± 0.24	3.43 ± 0.12	3.54 ± 0.11	4.59 ± 0.34	4.97 ± 0.27
20	3.30 ± 0.22	3.47 ± 0.18	3.45 ± 0.11	3.48 ± 0.11	4.68 ± 0.30	4.65 ± 0.23
10	3.34 ± 0.26	3.84 ± 0.18	3.44 ± 0.12	3.46 ± 0.11	4.59 ± 0.29	4.06 ±0.26
5	3.32 ± 0.23	4.58 ± 0.10	3.43 ± 0.11	3.38 ± 0.10	4.42 ± 0.33	3.49 ± 0.34
1	3.31 ± 0.25	4.72 ± 0.08	3.43 ± 0.12	3.24 ± 0.13	4.57 ± 0.29	3.73 ± 0.35

To some extent, the above result is to be expected since test utterances are only used to compute the speaker embedding. Thus, we also examined whether test utterances could instead be more beneficial if they were used to fine-tune a pre-trained FAC system. Starting with a pre-trained UU model, we fine-tuned the model on each unseen L1-L2 speaker pair (i.e., CLB-NJS, CLB-TXHC, CLB-YKWK, CLB-ZHAA) with N = 50, 20, 10, 5, 1 test utterances, resulting in 20 fine-tuned models (4 speakers; 5 models with different number of training utterances for each speaker) for the unseen L2 speakers. Results are also shown in Table 7. We observe performance degradations in all three measurements when reducing the number of utterances from 50 to 1. When there are 50 test utterances, the fine-tuned system shows a marginal improvement compared to the zero-shot model (i.e., without fine-tuning), though the differences are not statistically significant (Accentedness: 3.03 vs. 3.30, $p = 0.03$; Acoustic quality: 3.54 vs. 3.43, $p = 0.18$; Voice identity: 4.97 vs. 4.59, $p = 0.25$). When decreasing the number from 50 to 20, the fine-tuned system achieves comparable performance as the zero-shot system ($p > 0.5$). Surprisingly, however, fine-tuning the systems with fewer than 20 utterances degrades performance compared to the zero-shot model. In the extreme case (with only 1 utterance), the zero-shot model significantly outperforms the fine-tuned model in all three

IPR2026-00272
Krisp Technologies EX1010 Page 126

Journal Pre-proof

measurements (Accentedness: 4.72 vs. 3.31, $p \ll 0.001$; Acoustic quality: 3.24 vs. 3.43, $p = 0.01$; Voice identity: 3.73 vs. 4.57, $p \ll 0.001$). These results further speak of the robustness of Accentron in the zero-shot condition, and they also illustrate the tradeoff between zero-shot learning models and fine-tuning models in FAC.

6. Discussion

We have proposed Accentron, a zero-shot many-to-many speech synthesizer that can convert utterances from a source speaker to appear as if someone else, and with a different accent, had produced it. We thoroughly evaluated the system through a series of objective and perceptual listening experiments. Visualizations through t-SNE show that Accentron captures the target voice identity and accent, and that the speaker and accent embeddings are independent of each other and effectively summarize the speaker and accent characteristic of an utterance.

We evaluated Accentron in a standard FAC setting and compared it against two state-of-the-art baseline FAC systems. Compared to baseline 1 [15], Accentron achieves significantly better (i.e., lower) ratings of foreign accentedness, and similar acoustic quality and voice identity ratings. This is an important finding since baseline 1 builds a dedicated model for each pair of L1-L2 speakers, which one would expect would help capture voice identity more faithfully than Accentron's many-to-many mapping. Although baseline 1 and Accentron use the same backbone architecture for the seq2seq model, Accentron achieves significantly better (lower) ratings of accentedness. A possible explanation for this result is that baseline 1 is trained exclusively on L2 utterances. Thus, if the L2 speaker has systematic substitution or deletion errors (e.g., Mandarin speakers from certain areas systematically substitute /SH/ with /S/), the correct pronunciations will be missing in their utterances. Thus, when the baseline 1 model is driven by L1 BNFs during inference, it has to interpolate these missing pronunciations, which leads to noticeable segmental errors. In con-

IPR2026-00272
Krisp Technologies EX1010 Page 127

trast, Accentron avoids this potential issue since it is trained using both L1 and L2 speech. Compared to baseline 2 [16], Accentron achieves better ratings in all three measurements (accentedness, acoustic quality, and voice identity). Though Accentron and baseline 2 both use a seq2seq architecture, Accentron has two additional components: an acoustic model and an accent encoder to extract the linguistic content and accent embeddings, respectively, from an L1 reference during inference. These two embeddings capture the L1 segmental and prosodic patterns, respectively, which are shown to be essential to achieve advanced FAC performance.

We also evaluated Accentron on a "reverse FAC" task, where the goal was to impart an L2 accent to a native utterance. Results on this task corroborates the t-SNE visualizations, and suggest that Accentron can also preserve an L2 accent and implant it into an L1 speaker's utterance. Combined, results from the direct and reverse FAC tasks indicate that the proposed system can disentangle linguistic content, voice identity, and accent in speech signals, instead of merely memorizing mappings between different speaker pairs.

Finally, we also evaluated Accentron in a zero-shot FAC setting. First, we compared all four combinations in which the L1 speaker and L2 speakers could have been seen/unseen during training (i.e., SS, SU, US, UU). And we found no significant differences among the four conditions in terms of accentedness, acoustic quality, or voice identity. Thus, Accentron performs equally well under a standard FAC setting (condition SS) and a zero-shot FAC setting (condition UU), which indicates that it generalizes to unseen L1 and L2 speakers without the need to re-train or fine-tune the model.

Can Accentron generalize to unseen accents? In principle, the Accentron architecture can naturally generalize to unseen accents –in the same way that it is able to model unseen speakers. However, this requires access to sufficient training data on a large number of accents. The Speech Accent Archive has hundreds of accents, but unfortunately each speaker only produces less than one minute of speech, which is insufficient for training the FAC model (empirically, at least one hour per speaker is needed). Its counterpart, the L2-ARCTIC corpus,

IPR2026-00272
Krisp Technologies EX1010 Page 128

Journal Pre-proof

has substantially more speech data per subject (around one hour), but only has six accents, which is insufficient to achieve reasonable transferability to an unseen accent. Thus, while we believe that Accentron can be used to generate unseen accents, this ability can only be truly assessed when more accented speech corpora become available.

Our results under both standard and zero-shot FAC setting indicate that Accentron can generate FAC synthesis with high-quality, and can achieve it with limited data from new L1 and L2 speakers (one utterance, or around three seconds of speech). This capability can dramatically simplify the deployment of pronunciation-training tools (our envisioned target application). When deploying conventional FAC models, one needs to design and implement the model training pipeline, as it requires training a dedicated model for each target L2 speaker. Running the training pipeline is usually resource- and time-consuming, and therefore, one has to include specific modules to manage the server's computational resources (e.g., an asynchronous queue was used in [3]). Such systems are hard to scale to an increasing number of users, due to the heavy resource demands. Instead, Accentron only requires running the model inference pipeline when producing FAC synthesis, and the inference can be accomplished in real time, thus essentially reducing the resource demands and simplifying the application design and deployment. In addition, in previous FAC studies (e.g., [11, 12, 15]), inter-gender conversion usually achieved inferior performance than intra-gender conversion, due to the mismatch in pitch ranges. As a result, when deploying these methods into practical applications, they have to use a reference L1 speaker that has the same gender as the target L2 speaker. In contrast, Accentron achieves similar voice identity ratings for intra-gender pairs and inter-gender pairs, which gives more flexibility when choosing the reference L1 speakers in the pronunciation-training tool. Finally, Accentron significantly reduces the data required for each new L2 speaker from hours to seconds. As a result, the L2 learners only need to go through a simple speaker enrollment process (recording several seconds of their speech), before practicing with their "golden speaker" voices, which makes Accentron an ideal system for performing

IPR2026-00272
Krisp Technologies EX1010 Page 129

Journal Pre-proof

pronunciation-training studies at scale.

7. Conclusion and future work

In this paper, we have proposed Accentron, a zero-shot learning system that can generate accent conversion for any L2 speaker (seen or unseen). Our proposed approach is in contrast to most of existing FAC approaches, which require building a separate model for each L2 speaker. The proposed approach first trains separate models to extract L1 bottleneck features, L1 accent embeddings, and L2 speaker embeddings. Then it uses a seq2seq model to transform L1 bottleneck features to accent-converted Mel-spectrogram, conditioned on an L1 accent embedding and L2 speaker embedding. Our results suggest that the system can successfully transform L1 speech to match the voice identity of an L2 speaker while using a small amount of data from the L2 speaker.

One possible future direction of this work is to improve its robustness in generating long utterances. Currently, our system uses a location-sensitive attention mechanism [35] in the seq2seq model, which can fail when the utterances are too long [90] (e.g., longer than 10 seconds). To solve this problem, an alternative attention mechanism could be used, such as Gaussian mixture attention mechanism [91], which has been shown to be more robust in generating long utterances [90]. An additional potential improvement would be to add an auxiliary decoder to perform phoneme recognition (during training) [8, 70]. Such auxiliary decoder would guide the hidden representation produced by the encoder to preserve phonetic information, enforcing the synthesized speech to be phonetically reasonable and improving synthesis quality [8, 70].

References

[1] D. Felps, H. Bortfeld, R. Gutierrez-Osuna, Foreign accent conversion in computer assisted pronunciation training, Speech communication 51 (10) (2009) 920–932.

IPR2026-00272
Krisp Technologies EX1010 Page 130

Journal Pre-proof

[2] K. Probst, Y. Ke, M. Eskenazi, Enhancing foreign language tutors–in search of the golden speaker, Speech Communication 37 (3-4) (2002) 161–173.

[3] S. Ding, C. Liberatore, S. Sonsaat, I. Lučić, A. Silpachai, G. Zhao, E. Chukharev-Hudilainen, J. Levis, R. Gutierrez-Osuna, Golden speaker builder–an interactive tool for pronunciation training, Speech Communication 115 (2019) 51–66.

[4] R. Wang, J. Lu, Investigation of golden speakers for second language learners from imitation preference perspective by voice modification, Speech Communication 53 (2) (2011) 175–184.

[5] O. Turk, L. M. Arslan, Subband based voice conversion, in: International Conference on Spoken Language Processing, 2002.

[6] L. Sun, H. Wang, S. Kang, K. Li, H. M. Meng, Personalized, cross-lingual TTS using phonetic posteriorgrams., in: INTERSPEECH, 2016, pp. 322–326.

[7] Y. Oshima, S. Takamichi, T. Toda, G. Neubig, S. Sakti, S. Nakamura, Non-native speech synthesis preserving speaker individuality based on partial correction of prosodic and phonetic characteristics, in: INTERSPEECH, 2015.

[8] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, Y. Jia, Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation, in: INTERSPEECH, 2019, pp. 4115–4119.

[9] S. Aryal, D. Felps, R. Gutierrez-Osuna, Foreign accent conversion through voice morphing., in: INTERSPEECH, 2013, pp. 3077–3081.

[10] M. Huckvale, K. Yanagisawa, Spoken language conversion with accent morphing, in: ISCA Workshop on Speech Synthesis, 2007, pp. 64–70.

IPR2026-00272
Krisp Technologies EX1010 Page 131

Journal Pre-proof

[11] S. Aryal, R. Gutierrez-Osuna, Can voice conversion be used to reduce non-native accents?, in: ICASSP, IEEE, 2014, pp. 7879–7883.

[12] G. Zhao, R. Gutierrez-Osuna, Using phonetic posteriorgram based frame pairing for segmental accent conversion, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (10) (2019) 1649–1660.

[13] S. Aryal, R. Gutierrez-Osuna, Articulatory-based conversion of foreign accents with deep neural networks, in: INTERSPEECH, 2015.

[14] S. Aryal, R. Gutierrez-Osuna, Reduction of non-native accents through statistical parametric articulatory synthesis, The Journal of the Acoustical Society of America 137 (1) (2015) 433–446.

[15] G. Zhao, S. Ding, R. Gutierrez-Osuna, Foreign accent conversion by synthesizing speech from phonetic posteriorgrams., in: INTERSPEECH, 2019, pp. 2843–2847.

[16] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, et al., End-to-end accent conversion without using native utterances, in: ICASSP, IEEE, 2020, pp. 6289–6293.

[17] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 951–958.

[18] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, H. Meng, Voice conversion across arbitrary speakers based on a single target-speaker utterance., in: INTERSPEECH, 2018, pp. 496–500.

[19] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, H.-M. Wang, Voice conversion from non-parallel corpora using variational auto-encoder, in: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, 2016, pp. 1–6.

IPR2026-00272
Krisp Technologies EX1010 Page 132

Journal Pre-proof

[20] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder, arXiv preprint arXiv:1808.05092.

[21] L. Sun, K. Li, H. Wang, S. Kang, H. Meng, Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, in: IEEE International Conference on Multimedia and Expo, IEEE, 2016, pp. 1–6.

[22] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, in: Advances in Neural Information Processing Systems, 2018, pp. 4480–4490.

[23] S. Ö. Arık, J. Chen, K. Peng, W. Ping, Y. Zhou, Neural voice cloning with a few samples, in: International Conference on Neural Information Processing Systems, 2018, pp. 10040–10050.

[24] A. Das, G. Zhao, J. Levis, E. Chukharev-Hudilainen, R. Gutierrez-Osuna, Understanding the effect of voice quality and accent on talker similarity, in: INTERSPEECH, 2020, pp. 1763–1767.

[25] M. Leikin, R. Ibrahim, Z. Eviatar, S. Sapir, Listening with an accent: Speech perception in a second language by late bilinguals, Journal of psycholinguistic research 38 (5) (2009) 447.

[26] R. C. Major, S. F. Fitzmaurice, F. Bunta, C. Balasubramanian, The effects of nonnative accents on listening comprehension: Implications for esl assessment, TESOL quarterly 36 (2) (2002) 173–190.

[27] M. Van Heugten, C. Bergmann, A. Cristia, The effects of talker voice and accent on young children's speech perception, in: Individual differences in speech production and perception, Peter Lang, 2015, pp. 57–88.

[28] A. Cristia, A. Seidl, C. Vaughn, R. Schmale, A. Bradlow, C. Floccia, Linguistic processing of accented speech across the lifespan, Frontiers in psychology 3 (2012) 479.

IPR2026-00272
Krisp Technologies EX1010 Page 133

Journal Pre-proof

[29] S. Ding, R. Gutierrez-Osuna, Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion., in: INTERSPEECH, 2019, pp. 724–728.

[30] Y. Saito, Y. Ijima, K. Nishida, S. Takamichi, Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors, in: ICASSP, IEEE, 2018, pp. 5274–5278.

[31] W.-N. Hsu, Y. Zhang, J. Glass, Unsupervised learning of disentangled and interpretable representations from sequential data, in: Advances in Neural Information Processing Systems, 2017, pp. 1878–1889.

[32] S. H. Mohammadi, T. Kim, One-shot voice conversion with disentangled representations by leveraging phonetic posteriorgrams., in: INTERSPEECH, 2019, pp. 704–708.

[33] H. Lu, Z. Wu, D. Dai, R. Li, S. Kang, J. Jia, H. Meng, One-shot voice conversion with global speaker embeddings., in: INTERSPEECH, 2019, pp. 669–673.

[34] S. Ding, G. Zhao, R. Gutierrez-Osuna, Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition, in: INTERSPEECH, 2020, pp. 776–780.

[35] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in: Advances in Neural Information Processing Systems, 2015, pp. 577–585.

[36] D. Griffin, J. Lim, Signal estimation from modified short-time fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (2) (1984) 236–243.

[37] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, T. Toda, Speaker-dependent wavenet vocoder., in: INTERSPEECH, Vol. 2017, 2017, pp. 1118–1122.

IPR2026-00272
Krisp Technologies EX1010 Page 134

Journal Pre-proof

[38] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, K. Kavukcuoglu, Efficient neural audio synthesis, in: International Conference on Machine Learning, 2018.

[39] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, R. Gutierrez-Osuna, L2-ARCTIC: A non-native English speech corpus, in: INTERSPEECH, 2018, pp. 2783–2787.

[40] D. Felps, C. Geng, R. Gutierrez-Osuna, Foreign accent conversion through concatenative synthesis in the articulatory domain, IEEE Transactions on Audio, Speech, and Language Processing 20 (8) (2012) 2301–2312.

[41] S. Aryal, R. Gutierrez-Osuna, Data driven articulatory synthesis with deep neural networks, Computer Speech & Language 36 (2016) 260–273.

[42] M. Brand, Voice puppetry, in: Annual conference on Computer graphics and interactive techniques, 1999, pp. 21–28.

[43] B. Denby, M. Stone, Speech synthesis from real time ultrasound images of the tongue, in: ICASSP, Vol. 1, IEEE, 2004, pp. I–685.

[44] R. Mumtaz, S. Preuß, C. Neuschaefer-Rube, C. Hey, R. Sader, P. Birkholz, Tongue contour reconstruction from optical and electrical palatography, IEEE Signal Processing Letters 21 (6) (2014) 658–662.

[45] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, S. S. Narayanan, Articulatory synthesis based on real-time magnetic resonance imaging data., in: INTERSPEECH, 2016, pp. 1492–1496.

[46] S. H. Mohammadi, A. Kain, An overview of voice conversion systems, Speech Communication 88 (2017) 65–82.

[47] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning, arXiv preprint arXiv:2008.03648.

IPR2026-00272
Krisp Technologies EX1010 Page 135

Journal Pre-proof

[48] T. Toda, A. W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, IEEE Transactions on Audio, Speech, and Language Processing 15 (8) (2007) 2222–2235.

[49] D. Erro, A. Moreno, A. Bonafonte, INCA algorithm for training voice conversion systems from nonparallel corpora, IEEE Transactions on Audio, Speech, and Language Processing 18 (5) (2009) 944–953.

[50] R. Takashima, T. Takiguchi, Y. Ariki, Exemplar-based voice conversion in noisy environment, in: 2012 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2012, pp. 313–317.

[51] S. Ding, G. Zhao, C. Liberatore, R. Gutierrez-Osuna, Learning structured sparse representations for voice conversion, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2019) 343–354.

[52] L.-H. Chen, Z.-H. Ling, L.-J. Liu, L.-R. Dai, Voice conversion using deep neural networks with layer-wise generative training, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (12) (2014) 1859–1872.

[53] S. Desai, A. W. Black, B. Yegnanarayana, K. Prahallad, Spectral mapping using artificial neural networks for voice conversion, IEEE Transactions on Audio, Speech, and Language Processing 18 (5) (2010) 954–964.

[54] T. Nakashika, T. Takiguchi, Y. Minami, Non-parallel training in voice conversion using an adaptive restricted boltzmann machine, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (11) (2016) 2032–2045.

[55] F.-L. Xie, F. K. Soong, H. Li, A KL divergence and DNN-based approach to voice conversion without parallel training sentences., in: INTERSPEECH, 2016, pp. 287–291.

[56] H. Miyoshi, Y. Saito, S. Takamichi, H. Saruwatari, Voice conversion using sequence-to-sequence learning of context posterior probabilities, in: INTERSPEECH, 2017, pp. 1268–1272.

IPR2026-00272
Krisp Technologies EX1010 Page 136

Journal Pre-proof

[57] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion, in: ICASSP, IEEE, 2019, pp. 6820–6824.

[58] Y. Zhou, X. Tian, H. Xu, R. K. Das, H. Li, Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling, in: ICASSP, IEEE, 2019, pp. 6790–6794.

[59] W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, H.-M. Wang, Refined WaveNet vocoder for variational autoencoder based voice conversion, in: European Signal Processing Conference, IEEE, 2019, pp. 1–5.

[60] T. Kaneko, H. Kameoka, K. Hiramatsu, K. Kashino, Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks., in: INTERSPEECH, Vol. 2017, 2017, pp. 1283–1287.

[61] J.-X. Zhang, Z.-H. Ling, L.-R. Dai, Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2019) 540–552.

[62] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, H.-M. Wang, Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks, in: INTERSPEECH, Vol. 2017, 2017.

[63] A. Van Den Oord, O. Vinyals, et al., Neural discrete representation learning, in: Advances in Neural Information Processing Systems, 2017, pp. 6306–6315.

[64] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 19 (4) (2010) 788–798.

IPR2026-00272
Krisp Technologies EX1010 Page 137

Journal Pre-proof

[65] E. Variani, X. Lei, E. McDermott, I. L. Moreno, J. Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, in: ICASSP, IEEE, 2014, pp. 4052–4056.

[66] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[67] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015.

[68] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards end-to-end speech synthesis, in: INTERSPEECH, 2017, pp. 4006–4010.

[69] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions, in: ICASSP, IEEE, 2018, pp. 4779–4783.

[70] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, Y. Wu, Direct speech-to-speech translation with a sequence-to-sequence model, in: INTERSPEECH, 2019.

[71] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, S. Khudanpur, Semi-orthogonal low-rank matrix factorization for deep neural networks., in: INTERSPEECH, 2018, pp. 3743–3747.

[72] V. Peddinti, D. Povey, S. Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts, in: INTERSPEECH, 2015.

[73] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.

IPR2026-00272
Krisp Technologies EX1010 Page 138

Journal Pre-proof

[74] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books, in: ICASSP, IEEE, 2015, pp. 5206–5210.

[75] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[76] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, Z. Zhu, Deep speaker: an end-to-end neural speaker embedding system, arXiv preprint arXiv:1705.02304.

[77] W. Chan, N. Jaitly, Q. Le, O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in: ICASSP, IEEE, 2016, pp. 4960–4964.

[78] R. J. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, Neural Computation 1 (2) (1989) 270–280.

[79] A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: A large-scale speaker identification dataset, in: INTERSPEECH, 2017, pp. 2616–2620.

[80] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations, 2017.

[81] S. Weinberger, Speech accent archive, George Mason University.

[82] J. Kominek, A. W. Black, The CMU ARCTIC speech databases, in: Fifth ISCA workshop on speech synthesis, 2004.

[83] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467.

[84] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11).

IPR2026-00272
Krisp Technologies EX1010 Page 139

Journal Pre-proof

[85] G. E. Henter, J. Lorenzo-Trueba, X. Wang, M. Kondo, J. Yamagishi, Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody, in: ICASSP, IEEE, 2018, pp. 4799–4803.

[86] C. Veaux, J. Yamagishi, K. MacDonald, et al., Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

[87] M. J. Munro, T. M. Derwing, Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, Language Learning 49 (s1) (2002) 285–310.

[88] D. Felps, R. Gutierrez-Osuna, Developing objective measures of foreign-accent conversion, IEEE Transactions on Audio, Speech, and Language Processing 18 (5) (2010) 1030–1040.

[89] S. Buchholz, J. Latorre, Crowdsourcing preference tests, and how to detect cheating., in: INTERSPEECH, 2011, pp. 3053–3056.

[90] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, R. A. Saurous, Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron, arXiv preprint arXiv:1803.09047.

[91] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850.

IPR2026-00272
Krisp Technologies EX1010 Page 140

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof

Appendix DING08

http://web.archive.org/web/20211001000000*/https://psi.engr.tamu.edu/portfolio/accent-conversion/ October 28, 2025

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE

INTERNET ARCHIVE WayBackMachine

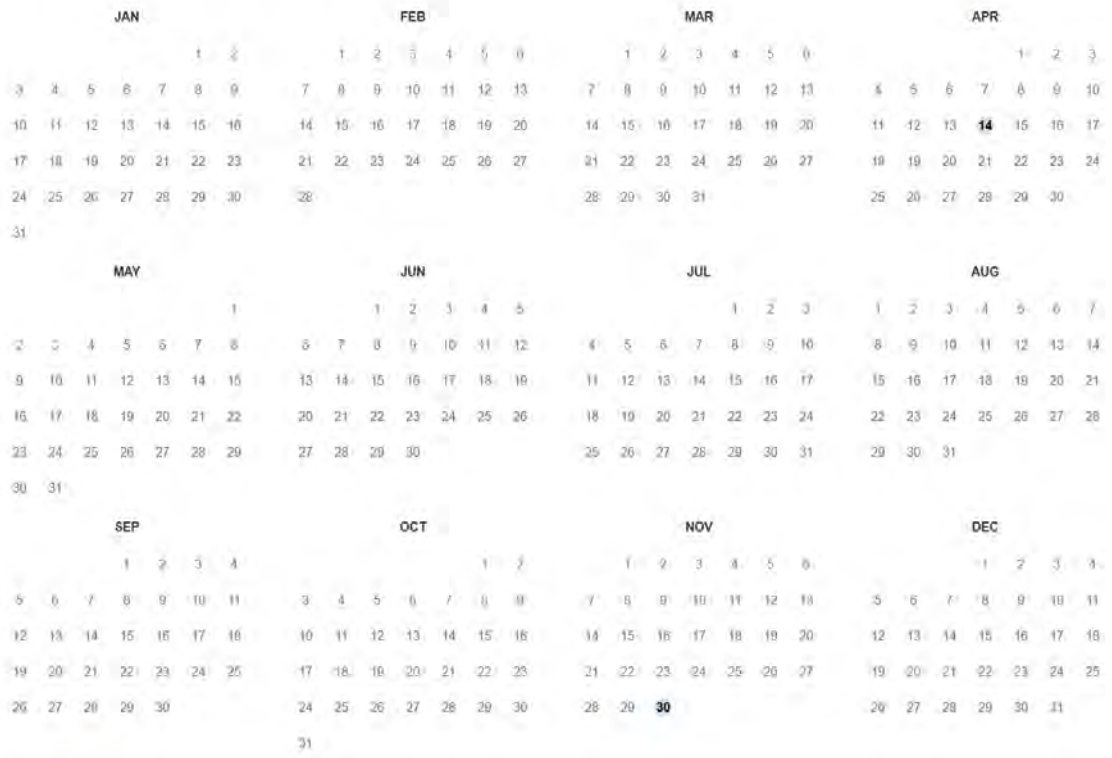
Explore more than 1 trillion web pages saved over time

https://psi.engr.tamu.edu/portfolio/accent-conversion/

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **16 times** between April 14, 2021 and August 3, 2025.





Note

This calendar view maps the number of times **https://psi.engr.tamu.edu/portfolio/accent-conversion/** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

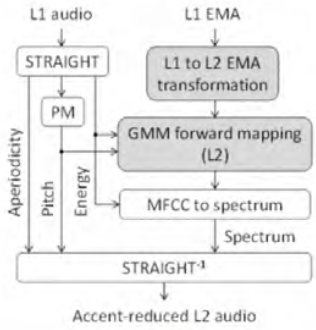
http://web.archive.org/web/20211130062820/https://psi.engr.tamu.edu/portfolio/accent-conversion/ October 28, 2025
INTERNET ARCHIVE Wayback Machine https://psi.engr.tamu.edu/portfolio/accent-conversion/ Go APR NOV 30 MAY
16 captures 14 Apr 2021 - 3 Aug 2025 2020 2021 2023 About this capture



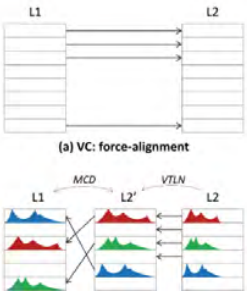
Accent conversion

Learners of a second language practice their pronunciation by listening to and imitating utterances from native speakers. Recent research has shown that choosing a well-matched native speaker to imitate can have a positive impact on pronunciation training. Towards this goal, we are developing speech-modification techniques that can generate utterances with the vocal properties of the learner and the accent of a native speaker. This is accomplished by altering both prosodic and segmental characteristics of speech.

Our articulatory-based accent conversion is a two-step process. In the first stage, we build an articulatory synthesizer for the nonnative learner. In the next step, we drive the synthesizer with articulatory gestures recorded from a native speaker.



We have also developed an accent conversion method that relies exclusively on acoustic information. The technique is based on the standard voice conversion model but uses a different pairing of source-target frames. Unlike conventional voice conversion, where the source-target mapping is trained on **time-aligned** source and target spectral vectors from parallel utterances, in our approach the mapping is trained on pairs selected based on their **acoustic similarity** following vocal tract length normalization.





(b) AC: acoustic similarity

(a) Conventional approach to voice conversion; source and target utterances are paired based on their ordering in a forced-aligned parallel corpus. (b) Our approach to accent conversion: source and target utterances are paired based on their acoustic similarity following vocal-tract-length normalization (VTLN). MCD: Mel Cepstral Distortion

2021

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning Journal Article Forthcoming

In: Computer Speech & Language, Forthcoming.

Links | BibTeX

https://www.sciencedirect.com/science/article/pii/S0885230821001029

https://psi.engr.tamu.edu/wp-content/uploads/2021/10/1-s2.0-S0885230821001029-mai[...]

Close

Silpachai, A.; Rehman, I.; Barmuso, T. A.; Levis, J.; Chukharev-Hudilainen, E.; Zhao, G.; Gutierrez-Osuna, R.

Effects Of Voice Type And Task On L2 Learners' Awareness Of Pronunciation Errors Inproceedings

In: Proc. Interspeech, 2021.

BibTeX

Liberatore, C.; Gutierrez-Osuna, R.

An Exemplar Selection Algorithm For Native-Nonnative Voice Conversion Inproceedings

In: Proc. Interspeech, 2021.

BibTeX

Zhao, G.; Ding, S.; Gutierrez-Osuna, R.

Converting Foreign Accent Speech Without a Reference Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29 , pp. 2367, 2021.

Links | BibTeX

2020

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition Inproceedings

In: Proc. Interspeech, 2020.

Links | BibTeX

Das, A.; Zhao, G.; Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.

Understanding the Effect of Voice Quality and Accent on Talker Similarity Inproceedings

In: Proc. Interspeech, 2020.

Links | BibTeX

Lučić, I.; Silpachai, A.; Levis, J.; Zhao, G; Gutierrez-Osuna, R.

The English Pronunciation of Arabic Speakers - A Data-Driven Approach to Segmental Error Identification Journal Article

In: Language Teaching Research, 2020.

Links | BibTeX

2019

Ding, S.; Zhao, G; Liberatore, C.; Gutierrez-Osuna, R.

Learning Structured Sparse Representations for Voice Conversion Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, 28 , pp. 343-354, 2019.

Links | BibTeX

Ding, S.; Liberatore, C.; Sonsaat, S.; Lučić, I.; Silpachai, A.; Zhao, G; Chukharev-Hudilainen, E.; Levis, J.; Gutierrez-Osuna, R.

Golden speaker builder – An interactive tool for pronunciation training Journal Article

In: Speech Communication, 115., pp. 51-66, 2019.

Links | BibTeX

Zhao, G; Ding, S.; Gutierrez-Osuna, Ricardo

Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams Inproceedings

In: Proc. Interspeech, 2019.

Links | BibTeX

Ding, S.; Gutierrez-Osuna, Ricardo

Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion Inproceedings

In: Proc. Interspeech, 2019.

Links | BibTeX

Zhao, G; Gutierrez-Osuna, R

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (10), pp. 1649-1660, 2019, ISSN: 2329-9290.

Abstract | Links | BibTeX

2018

Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.; Lucic, I.; Silpachai, A.; Sonsaat, S.

Golden Speaker: Learner Experience with Computer-assisted Pronunciation Practice Inproceedings

In: Proc. Pronunciation in Second Language Learning and Teaching Conference, 2018.

BibTeX

Ding, S.; Liberatore, C.; Gutierrez-Osuna, R.

Learning Structured Dictionaries for Exemplar-based Voice Conversion Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX

Zhao, G; Sonsaat, S; Silpachai, A; Lucic, I; Chukharev-Hudilainen, E; Levis, J; Gutierrez-Osuna, R

L2-ARCTIC: A Non-Native English Speech Corpus Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX

Ding, S.; Zhao, G; Liberatore, C.; Gutierrez-Osuna, R.

Improving Sparse Representations in Exemplar-Based Voice Conversion with a Phoneme-Selective Objective Function Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX

Zhao, G; Sonsaat, S; Levis, J; Chukharev-Hudilainen, E; Gutierrez-Osuna, R

Accent conversion using phonetic posteriorgrams Inproceedings

In: Proc. ICASSP, 2018.

Links | BibTeX

Liberatore, C; Zhao, G; Gutierrez-Osuna, R

Voice Conversion through Residual Warping in a Sparse, Anchor-Based Representation of Speech Inproceedings

In: Proc. ICASSP, 2018.

Abstract | Links | BibTeX

2016

Aryal, S; Gutierrez-Osuna, R

Comparing Articulatory and Acoustic Strategies for Reducing Non-Native Accents Inproceedings

In: Proc. Interspeech, 2016.

Links | BibTeX

Aryal, S; Gutierrez-Osuna, R

Data driven articulatory synthesis with deep neural networks Journal Article

In: Computer Speech and Language, 36 , pp. 260-273, 2016.

Links | BibTeX

2015

Aryal, S; Gutierrez-Osuna, R

Articulatory-based conversion of foreign accents with deep neural networks Inproceedings

In: Proc. Interspeech, pp. 3385-3389, 2015.

Links | BibTeX

Liberatore, C; Aryal, S; Wang, Z; Polsley, S; Gutierrez-Osuna, R

SABR: Sparse, Anchor-Based Representation of the Speech Signal Inproceedings

In: Proc. Interspeech 2015, pp. 608-612, 2015.

Abstract | Links | BibTeX

Liberatore, C; Gutierrez-Osuna, R

Joint Optimization of Anatomical and Gestural Parameters in a Physical Vocal Tract Model Inproceedings

In: ICASSP, IEEE, 2015.

Links | BibTeX

Aryal, S; Gutierrez-Osuna, R

Reduction of non-native accents through statistical parametric articulatory synthesis Journal Article

In: Journal of the Acoustical Society of America, 137 (1), pp. 433-446, 2015

Links | BibTeX

2014

Aryal, S; Gutierrez-Osuna, R

Accent conversion through cross-speaker articulatory synthesis Inproceedings

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7744-7748, 2014.

Links | BibTeX

Aryal, S; Gutierrez-Osuna, R

Can voice conversion be used to reduce non-native accents Inproceedings

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7929-7933, 2014.

Links | BibTeX

Felps, D; Aryal, S; Gutierrez-Osuna, R

Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis Inproceedings

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3051-3055, 2014.

Links | BibTeX

2013

Aryal, S; Felps, D; Gutierrez-Osuna, R

Foreign Accent Conversion through Voice Morphing Inproceedings

In: Interspeech, pp. 3077-3081, 2013.

Links | BibTeX

2012

Felps, D; Geng, C; Gutierrez-Osuna, R

Foreign accent conversion through concatenative synthesis in the articulatory domain Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, 2012

Abstract | Links | BibTeX

2010

Gutierrez-Osuna, R; Felps, D

Foreign Accent Conversion through Voice Morphing Technical Report

2010

Abstract | Links | BibTeX

Felps, D; Gutierrez-Osuna, R

Developing objective measures of foreign-accent conversion Journal Article

In: Audio, Speech, and Language Processing, IEEE Transactions on, 18 (5), pp. 1030-1040, 2010.

Abstract | Links | BibTeX

Felps, D; Geng, C; Berger, M; Richmond, K; Gutierrez-Osuna, R

Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database Conference

Interspeech, 2010.

Abstract | Links | BibTeX

2009

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Foreign accent conversion in computer assisted pronunciation training Journal Article

In: Speech communication, 51 (10), pp. 920-932, 2009

Abstract | Links | BibTeX

2008

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Prosodic and segmental factors in foreign-accent conversion Technical Report

2008

Abstract | Links | BibTeX

Physical: 309 Peterson Building, Texas A&M University
Mailing: PSI Lab, 3112 TAMU, College Station, TX 77843

http://web.archive.org/web/20210615000000*/https://psi.engr.tamu.edu/publications/?tgid=18 October 28, 2025

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE

INTERNET ARCHIVE WayBackMachine

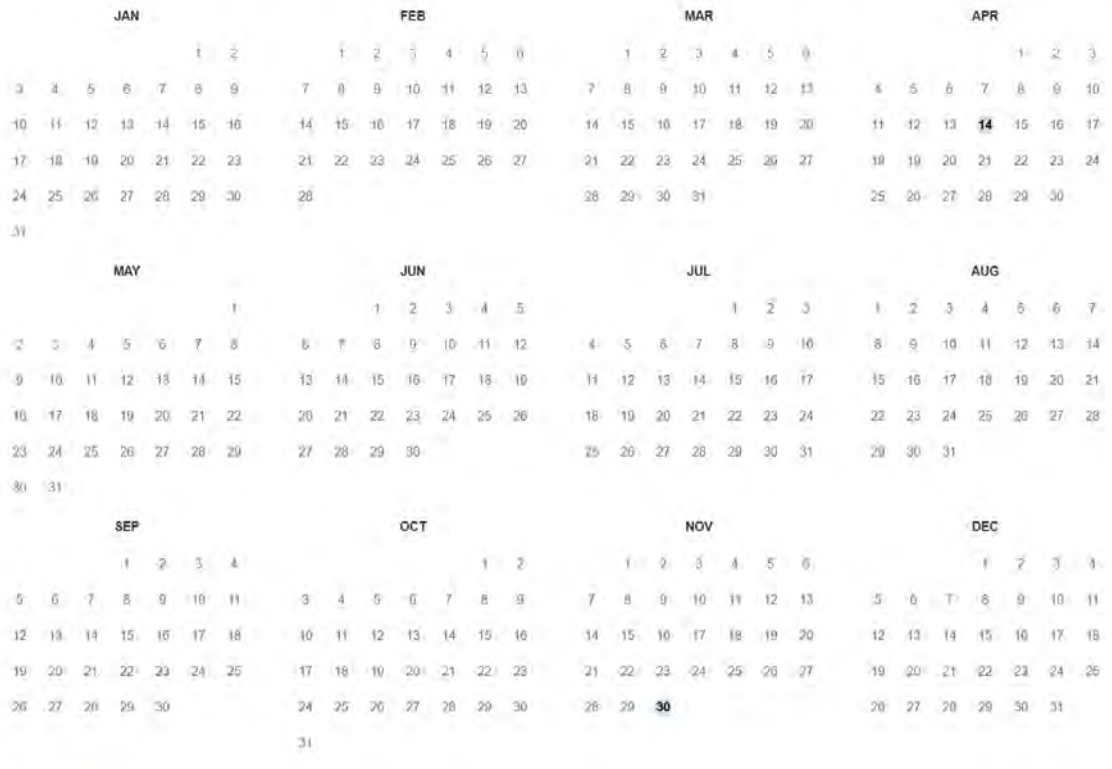
Explore more than 1 trillion web pages saved over time

https://psi.engr.tamu.edu/publications/?tgid=18

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **4 times** between April 14, 2021 and July 25, 2024.





Note

This calendar view maps the number of times **https://psi.engr.tamu.edu/publications/?tgid=18** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

Appendix DING09

http://web.archive.org/web/20211130052733/https://psi.engr.tamu.edu/publications/?tgid=18 October 28, 2025

INTERNET ARCHIVE Wayback Machine https://psi.engr.tamu.edu/publications/?tgid=18 Go APR NOV MAY
4 captures 14 Apr 2021 - 25 Jul 2024 2020 30 2021 2024 About this capture

Publications

Accent conversion Active sensing Articulatory inversion Articulatory synthesis Automatic Speech Recognition Chemical sensors Childhood apraxia of speech Computer vision Continuous glucose monitors Electrodermal activity Electronic nose Face perception Facial animation Games Gestures Heart rate variability Infrared spectroscopy Machine olfaction Metal-oxide sensors Mobile computing Neuromorphic models OCR Olfaction Other Pattern recognition Robotics Speech Stress Temperature modulation Wearable sensors

All years | All types | All authors | All users

Show all

2021

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning Journal Article Forthcoming

In: Computer Speech & Language, Forthcoming.

Links | BibTeX | Tags: Accent conversion, Speech

```
@article{shaojin2021accentron,
title = {Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning},
author = {S. Ding and G. Zhao and R. Gutierrez-Osuna},
url = {https://www.sciencedirect.com/science/article/pii/S0885230821001029
https://psi.engr.tamu.edu/wp-content/uploads/2021/10/1-s2.0-S0885230821001029-main.pdf},
year = {2021},
date = {2021-10-14},
journal = {Computer Speech & Language},
keywords = {Accent conversion, Speech},
pubstate = {forthcoming},
tppubtype = {article}
}
```

Close

Sitpachai, A.; Rehman, I.; Barriuso, T. A.; Levis, J.; Chukharev-Hudilainen, E.; Zhao, G.; Gutierrez-Osuna, R.

Effects Of Voice Type And Task On L2 Learners' Awareness Of Pronunciation Errors Inproceedings

In: Proc. Interspeech, 2021.

BibTeX | Tags: Accent conversion, Speech

Liberatore, C.; Gutierrez-Osuna, R.

An Exemplar Selection Algorithm For Native-Nonnative Voice Conversion Inproceedings

In: Proc. Interspeech, 2021.

BibTeX | Tags: Accent conversion, Speech

Zhao, G.; Ding, S.; Gutierrez-Osuna, R.

Converting Foreign Accent Speech Without a Reference Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp. 2367, 2021.

Links | BibTeX | Tags: Accent conversion, Speech

2020

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition Inproceedings

In: Proc. Interspeech, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

Das, A.; Zhao, G.; Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.

Understanding the Effect of Voice Quality and Accent on Talker Similarity Inproceedings

In: Proc. Interspeech, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

Lučić, I., Silpachai, A., Levis, J., Zhao, G., Gutierrez-Osuna, R.

The English Pronunciation of Arabic Speakers - A Data-Driven Approach to Segmental Error Identification Journal Article

In: Language Teaching Research, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

2019

Ding, S., Zhao, G, Liberatore, C., Gutierrez-Osuna, R.

Learning Structured Sparse Representations for Voice Conversion Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, 28 , pp. 343-354, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G, Chukharev-Hudilainen, E., Levis, J., Gutierrez-Osuna, R.

Golden speaker builder – An interactive tool for pronunciation training Journal Article

In: Speech Communication, 115 , pp. 51-66, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G, Ding, S., Gutierrez-Osuna, Ricardo

Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams Inproceedings

In: Proc. Interspeech, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Ding, S., Gutierrez-Osuna, Ricardo

Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion Inproceedings

In: Proc. Interspeech, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G, Gutierrez-Osuna, R.

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (10), pp. 1649-1660, 2019, ISSN: 2329-9290.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2018

Levis, J., Chukharev-Hudilainen, E., Gutierrez-Osuna, R., Lucic, I., Silpachai, A., Sonsaat, S.

Golden Speaker: Learner Experience with Computer-assisted Pronunciation Practice Inproceedings

In: Proc. Pronunciation in Second Language Learning and Teaching Conference, 2018.

BibTeX | Tags: Accent conversion, Speech

Ding, S., Liberatore, C., Gutierrez-Osuna, R.

Learning Structured Dictionaries for Exemplar-based Voice Conversion Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G, Sonsaat, S, Silpachai, A, Lucic, I, Chukharev-Hudilainen, E, Levis, J, Gutierrez-Osuna, R.

L2-ARCTIC: A Non-Native English Speech Corpus Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Ding, S., Zhao, G, Liberatore, C., Gutierrez-Osuna, R.

Improving Sparse Representations in Exemplar-Based Voice Conversion with a Phoneme-Selective Objective Function Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G, Sonsaat, S, Levis, J, Chukharev-Hudilainen, E, Gutierrez-Osuna, R.

Accent conversion using phonetic posteriorgrams Inproceedings

In: Proc. ICASSP, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Liberatore, C, Zhao, G, Gutierrez-Osuna, R.

Voice Conversion through Residual Warping in a Sparse, Anchor-Based Representation of Speech Inproceedings

In: Proc. ICASSP, 2018.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2016

Aryal, S, Gutierrez-Osuna, R.

Comparing Articulatory and Acoustic Strategies for Reducing Non-Native Accents **Inproceedings**

In: Proc. Interspeech, 2016.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

Aryal, S; Gutierrez-Osuna, R.

Data driven articulatory synthesis with deep neural networks **Journal Article**

In: Computer Speech and Language, 36 , pp. 260-273, 2016.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

2015

Aryal, S; Gutierrez-Osuna, R.

Articulatory-based conversion of foreign accents with deep neural networks **Inproceedings**

In: Proc. Interspeech, pp. 3385-3389, 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

Liberatore, C; Aryal, S; Wang, Z; Polsley, S; Gutierrez-Osuna, R.

SABR: Sparse, Anchor-Based Representation of the Speech Signal **Inproceedings**

In: Proc. Interspeech 2015, pp. 608-612, 2015.

Abstract | Links | BibTeX | Tags: Accent conversion, Articulatory synthesis

Liberatore, C; Gutierrez-Osuna, R.

Joint Optimization of Anatomical and Gestural Parameters in a Physical Vocal Tract Model **Inproceedings**

In: ICASSP, IEEE 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory inversion, Articulatory synthesis

Aryal, S; Gutierrez-Osuna, R.

Reduction of non-native accents through statistical parametric articulatory synthesis **Journal Article**

In: Journal of the Acoustical Society of America, 137 (1), pp. 433-446, 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

2014

Aryal, S; Gutierrez-Osuna, R.

Accent conversion through cross-speaker articulatory synthesis **Inproceedings**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7744-7748, 2014.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

Aryal, S; Gutierrez-Osuna, R.

Can voice conversion be used to reduce non-native accents **Inproceedings**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7929-7933, 2014.

Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Aryal, S; Gutierrez-Osuna, R.

Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis **Inproceedings**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3051-3055, 2014.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

2013

Aryal, S; Felps, D; Gutierrez-Osuna, R.

Foreign Accent Conversion through Voice Morphing **Inproceedings**

In: Interspeech, pp. 3077-3081, 2013.

Links | BibTeX | Tags: Accent conversion, Speech

2012

Felps, D; Geng, C; Gutierrez-Osuna, R.

Foreign accent conversion through concatenative synthesis in the articulatory domain **Journal Article**

In: IEEE Transactions on Audio, Speech and Language Processing, 2012.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2010

Gutierrez-Osuna, R; Felps, D.

Foreign Accent Conversion through Voice Morphing **Technical Report**

2010.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Gutierrez-Osuna, R

Developing objective measures of foreign-accent conversion Journal Article

In: Audio, Speech, and Language Processing, IEEE Transactions on, 18 (5), pp. 1030–1040, 2010

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Geng, C; Berger, M; Richmond, K; Gutierrez-Osuna, R

Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database Conference

Interspeech, 2010

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2009

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Foreign accent conversion in computer assisted pronunciation training Journal Article

In: Speech communication, 51 (10), pp. 920–932, 2009.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2008

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Prosodic and segmental factors in foreign-accent conversion Technical Report

2008

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Physical: 309 Peterson Building, Texas A&M University
Mailing: PSI Lab, 3112 TAMU, College Station, TX 77843

Appendix DING10

http://web.archive.org/web/20210501000000*/https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=&auth= November 14, 2025

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE

INTERNET ARCHIVE WayBackMachine

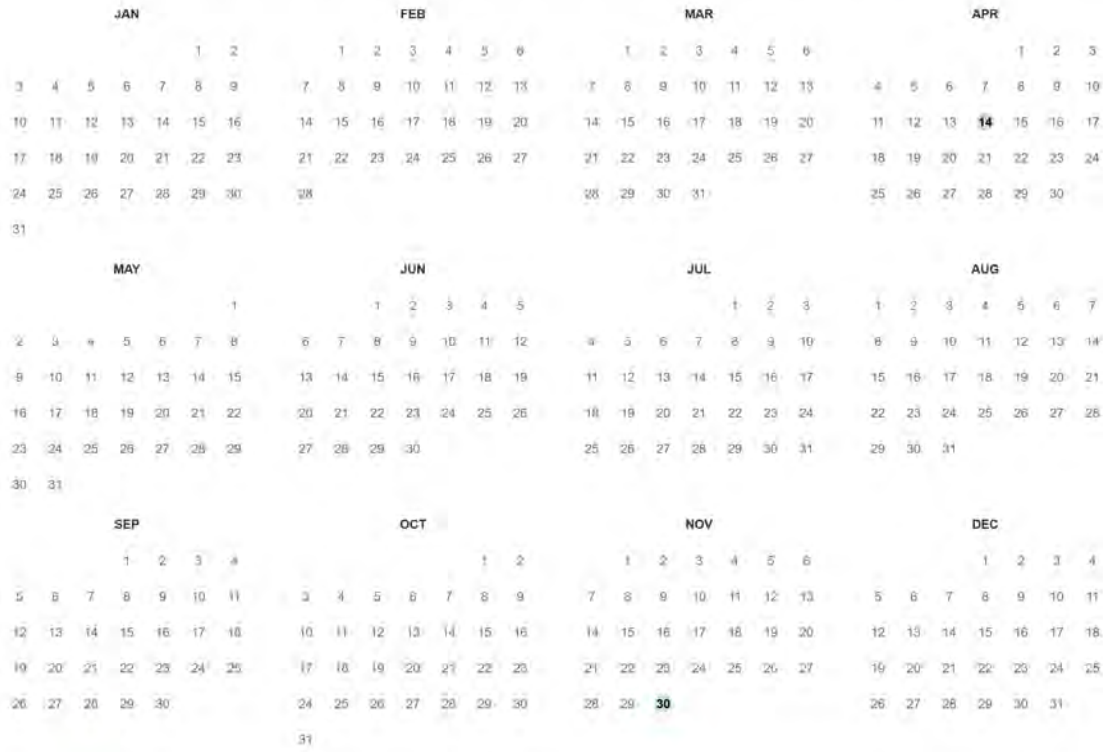
Explore more than 1 trillion web pages saved over time

https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **4 times** between April 14, 2021 and March 31, 2025.





Note

This calendar view maps the number of times **https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=&auth=** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use

http://web.archive.org/web/20211130053613/https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=&auth= November 14, 2025

INTERNET ARCHIVE WaybackMachine https://psi.engr.tamu.edu/publications/?tgid=11&yr=&type=&usr=&auth= Go APR NOV JUL
4 captures 14 Apr 2021 - 31 Mar 2025 2020 30 2021 2024 About this capture

Publications

Accent conversion Active sensing Articulatory inversion Articulatory synthesis Automatic Speech Recognition Chemical sensors Childhood apraxia of speech Computer vision Continuous glucose monitors Electrodermal activity Electronic nose Face perception Facial animation Games Gestures Heart rate variability Infrared spectroscopy Machine olfaction Metal-oxide sensors Mobile computing Neuromorphic models OCR Olfaction Other Pattern recognition Robotics Speech Stress Temperature modulation Wearable sensors

All years | All types | All authors | All users

Show all

2021

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning Journal Article Forthcoming

In: Computer Speech & Language, Forthcoming.

Links | BibTeX | Tags: Accent conversion, Speech

```
@article{shaojin2021accentron,
title = {Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning},
author = {S. Ding and G. Zhao and R. Gutierrez-Osuna},
url = {https://www.sciencedirect.com/science/article/pii/S0885230821001029
https://psi.engr.tamu.edu/wp-content/uploads/2021/10/1-s2.0-S0885230821001029-main.pdf},
year = {2021},
date = {2021-10-14},
journal = {Computer Speech & Language},
keywords = {Accent conversion, Speech},
pubstate = {forthcoming},
tppubtype = {article}
}
```

Close

Hair, A.; Zhao, G.; Ahmed, B.; Ballard, K. J.; Gutierrez-Osuna, R.

Assessing Posterior-Based Mispronunciation Detection on Field-Collected Recordings from Child Speech Therapy Sessions Inproceedings

In: Proc. Interspeech, 2021.

Links | BibTeX | Tags: Automatic Speech Recognition, Childhood apraxia of speech, Speech

Silpachai, A.; Rehman, I.; Barruso, T. A.; Levis, J.; Chukharev-Hudilainen, E.; Zhao, G.; Gutierrez-Osuna, R.

Effects Of Voice Type And Task On L2 Learners' Awareness Of Pronunciation Errors Inproceedings

In: Proc. Interspeech, 2021.

BibTeX | Tags: Accent conversion, Speech

Liberatore, C.; Gutierrez-Osuna, R.

An Exemplar Selection Algorithm For Native-Nonnative Voice Conversion Inproceedings

In: Proc. Interspeech, 2021.

BibTeX | Tags: Accent conversion, Speech

Zhao, G.; Ding, S.; Gutierrez-Osuna, R.

Converting Foreign Accent Speech Without a Reference Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29 , pp. 2367, 2021.

Links | BibTeX | Tags: Accent conversion, Speech

Hair, A.; Ballard, K. J.; Markoulli, C.; Monroe, P.; McKechnie, J.; Ahmed, B.; Gutierrez-Osuna, R.

A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World Journal Article

In: ACM Transactions On Accessible Computing, 14 (1), 2021.

Abstract | Links | BibTeX | Tags: Games, Speech

2020

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition Inproceedings

In: Proc. Interspeech, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

Das, A.; Zhao, G.; Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.

Understanding the Effect of Voice Quality and Accent on Talker Similarity Inproceedings

In: Proc. Interspeech, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

Lučić, I.; Silpachai, A.; Levis, J.; Zhao, G; Gutierrez-Osuna, R.

The English Pronunciation of Arabic Speakers - A Data-Driven Approach to Segmental Error Identification Journal Article

In: Language Teaching Research, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

McKechnie, J.; Ahmed, B.; Gutierrez-Osuna, R.; Murray, E.; McCabe, P.; Ballard, K.

The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech Journal Article

In: Journal of Communication Disorders, 2020.

Links | BibTeX | Tags: Speech

Hair, A; Markoulli, C; Monroe, P; McKechnie, J; Ballard, K J; Ahmed, B; Gutierrez-Osuna, R

Preliminary Results From a Longitudinal Study of a Tablet-Based Speech Therapy Game Inproceedings

In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing, ACM, 2020, ISBN: 978-1-4503-6819-3/20/04.

Abstract | Links | BibTeX | Tags: Automatic Speech Recognition, Childhood apraxia of speech, Speech

2019

Ding, S.; Zhao, G; Liberatore, C.; Gutierrez-Osuna, R.

Learning Structured Sparse Representations for Voice Conversion Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, 28., pp. 343-354, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Ding, S.; Liberatore, C.; Sonsaat, S.; Lučić, I.; Silpachai, A.; Zhao, G; Chukharev-Hudilainen, E.; Levis, J.; Gutierrez-Osuna, R.

Golden speaker builder – An interactive tool for pronunciation training Journal Article

In: Speech Communication, 115 , pp. 51-66, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Hair, A; Ballard, K J; Ahmed, B; Gutierrez-Osuna, R

Evaluating Automatic Speech Recognition for Child Speech Therapy Applications Inproceedings

In: ACM SIGACCESS Conference on Computers and Accessibility, ACM 2019, ISBN: 978-1-4503-6676-2/19/10.

Abstract | Links | BibTeX | Tags: Automatic Speech Recognition, Childhood apraxia of speech, Speech

---

Zhao, G; Ding, S.; Gutierrez-Osuna, Ricardo

Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams Inproceedings

In: Proc. Interspeech, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Ding, S.; Gutierrez-Osuna, Ricardo

Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion Inproceedings

In: Proc. Interspeech, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G; Gutierrez-Osuna, R

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (10), pp. 1649-1660, 2019, ISSN: 2329-9290.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Monteiro, C. D. D.; Shipman, F. M.; III, S. Duggina; Gutierrez-Osuna, R.

Tradeoffs in the Efficient Detection of Sign Language Content in Video Sharing Sites Journal Article

In: ACM Transactions on Accessible Computing, 12 (2), pp. 1-16, 2019.

Links | BibTeX | Tags: Computer vision, Speech

2018

Ahmed, B; Monroe, P; Hair, A; Tan, C-T; Gutierrez-Osuna, R; Ballard, K J

Speech-driven mobile games for speech therapy: User experiences and feasibility Journal Article

In: International Journal of Speech-Language Pathology , 20 (6), pp. 644-658, 2018.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Speech

Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.; Lucic, I.; Silpachai, A.; Sonsaat, S.

Golden Speaker: Learner Experience with Computer-assisted Pronunciation Practice Inproceedings

In: Proc. Pronunciation in Second Language Learning and Teaching Conference, 2018.

BibTeX | Tags: Accent conversion, Speech

Ding, S.; Liberatore, C.; Gutierrez-Osuna, R.

Learning Structured Dictionaries for Exemplar-based Voice Conversion Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G; Sonsaat, S; Silpachai, A; Lucic, I; Chukharev-Hudilainen, E; Levis, J; Gutierrez-Osuna, R

L2-ARCTIC: A Non-Native English Speech Corpus Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

McKechnie, J.; Ahmed, B.; Gutierrez-Osuna, R.; Monroe, P.; McCabe, P.; Ballard, K. J.

Automated speech analysis tools for children's speech production: A systematic literature review Journal Article

In: International Journal of Speech-Language Pathology, 20 (6), pp. 583–598, 2018.

Links | BibTeX | Tags: Childhood apraxia of speech, Speech

Ding, S.; Zhao, G; Liberatore, C.; Gutierrez-Osuna, R.

Improving Sparse Representations in Exemplar-Based Voice Conversion with a Phoneme-Selective Objective Function Inproceedings

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Hair, A; Monroe, P; Ahmed, B; Ballard, K J; Gutierrez-Osuna, R

Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders Inproceedings

In: Proceedings of the 2018 Conference on Interaction Design and Children, ACM, 2018, ISBN: 978-1-4503-5152-2/18/06.

Abstract | Links | BibTeX | Tags: Childhood apraxia of speech, Mobile computing, Speech

---

Zhao, G; Sonsaat, S; Levis, J; Chukharev-Hudilainen, E; Gutierrez-Osuna, R

Accent conversion using phonetic posteriorgrams Inproceedings

In: Proc. ICASSP, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Liberatore, C; Zhao, G; Gutierrez-Osuna, R

Voice Conversion through Residual Warping in a Sparse, Anchor-Based Representation of Speech Inproceedings

In: Proc. ICASSP, 2018.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2017

Shipman, F; Duggina, S; Monteiro, C; Gutierrez-Osuna, R

Speed-Accuracy Tradeoffs for Detecting Sign Language Content in Video Sharing Sites Inproceedings

In: Proceedings of ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2017), pp. 185–189, 2017.

Links | BibTeX | Tags: Computer vision, Gestures, Speech

Zhao, G; Gutierrez-Osuna, R

Exemplar selection methods in voice conversion Inproceedings

In: Proc. 42nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5525-5529, 2017.

Links | BibTeX | Tags: Speech, Voice conversion

2016

Aryal, S; Gutierrez-Osuna, R

Comparing Articulatory and Acoustic Strategies for Reducing Non-Native Accents Inproceedings

In: Proc. Interspeech, 2016.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

Liberatore, C; Gutierrez-Osuna, R

Generating Gestural Scores from Acoustics Through a Sparse Anchor-Based Representation of Speech Inproceedings

In: Proc. Interspeech, 2016.

Links | BibTeX | Tags: Articulatory synthesis, Speech

Shahin, M; Gutierrez-Osuna, R; Ahmed, B

Classification of bisyllabic lexical stress patterns in disordered speech using deep learning Inproceedings

In: Proc. International Conference on Acoustics, Speech, and Signal Processing, 2016.

Links | BibTeX | Tags: Speech

McKechnie, J; Ballard, K J; McCabe, P; Murray, E; Lan, T; Gutierrez-Osuna, R; Ahmed, B

Influence of type of feedback on effect of tablet-based delivery of intensive speech therapy in children with Childhood Apraxia of Speech Inproceedings

In: Proceedings of the Motor Speech Conference, 2016.

BibTeX | Tags: Childhood apraxia of speech, Games, Mobile computing, Speech

---

Aryal, S; Gutierrez-Osuna, R

Data driven articulatory synthesis with deep neural networks Journal Article

In: Computer Speech and Language, 36 , pp. 260-273, 2016

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

2015

Parnandi, A; Karappa, V; Lan, T; Shahin, M; McKechnie, J; Ballard, K; Ahmed, B; Gutierrez-Osuna, R

Development of a remote therapy tool for childhood apraxia of speech Journal Article

In: ACM Transactions on Accessible Computing, 7 (3), pp. 10:1-10:23, 2015.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Mobile computing, Speech

---

Aryal, S; Gutierrez-Osuna, R

Articulatory-based conversion of foreign accents with deep neural networks Inproceedings

In: Proc. Interspeech, pp. 3385-3389, 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

Shahin, M; Ahmed, B; Parnandi, A; Karappa, V; McKechnie, J; Ballard, K; Gutierrez-Osuna, R

Tabby Talks: an automated tool for the assessment of childhood apraxia of speech Journal Article

In: Speech Communication, in press , 2015.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Mobile computing, Speech

---

Aryal, S; Gutierrez-Osuna, R

Reduction of non-native accents through statistical parametric articulatory synthesis Journal Article

In: Journal of the Acoustical Society of America, 137 (1), pp. 433-446, 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

2014

Lan, T; Aryal, S; Ahmed, B; Ballard, K; Gutierrez-Osuna, R

Flappy Voice: An Interactive Game for Childhood Apraxia of Speech Therapy Inproceedings

In: Proc. CHI-PLAY, 2014.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Speech

---

Shahin, M; Ahmed, B; McKechnie, J; Ballard, K; Gutierrez-Osuna, R

A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech Inproceedings

In: Proc. Interspeech, 2014.

Links | BibTeX | Tags: Childhood apraxia of speech, Speech

McKechnie, J; Ballard, K; McCabe, P; Gutierrez-Osuna, R; Karappa, V; Parnandi, A; Shahin, M; Murray, E; Ahmed, B

Tablet-based delivery of intensive speech therapy in children with Childhood Apraxia of Speech - Pilot Phase Inproceedings

In: Speech Pathology Australia National Conference, 2014.

BibTeX | Tags: Games, Speech

Aryal, S; Gutierrez-Osuna, R

Accent conversion through cross-speaker articulatory synthesis Inproceedings

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7744-7748, 2014.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

Aryal, S; Gutierrez-Osuna, R

Can voice conversion be used to reduce non-native accents Inproceedings

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7929-7933, 2014.

Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Aryal, S; Gutierrez-Osuna, R

Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis Inproceedings

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3051-3055, 2014.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

2013

Parnandi, A; Karappa, V; Sen, Y; Shahin, M; McKechnie, J; Ballard, K; Ahmed, B; Gutierrez-Osuna, R

Architecture of an automated therapy tool for childhood apraxia of speech Conference

The 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2013.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Mobile computing, Speech

---

Aryal, S; Felps, D; Gutierrez-Osuna, R

Foreign Accent Conversion through Voice Morphing Inproceedings

In: Interspeech, pp. 3077-3081, 2013.

Links | BibTeX | Tags: Accent conversion, Speech

Aryal, S; Gutierrez-Osuna, R

Articulatory inversion and synthesis: towards articulatory-based modification of speech Inproceedings

In: 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7952-7956, 2013.

Links | BibTeX | Tags: Articulatory inversion, Articulatory synthesis, Speech

---

2012

Aryal, S; Gutierrez-Osuna, R

Articulatory inversion and Synthesis: Towards Articulatory-Based Modification of Speech Technical Report

2012.

Abstract | Links | BibTeX | Tags: Articulatory inversion, Articulatory synthesis, Speech

---

Parnandi, A; Son, Y; Shahin, M; Ahmed, B; Gutierrez-Osuna, R

Architecture of an Automated Therapy Tool for Childhood Apraxia of Speech Technical Report

2012.

Abstract | Links | BibTeX | Tags: Games, Mobile computing, Speech

---

Felps, D; Geng, C; Gutierrez-Osuna, R

Foreign accent conversion through concatenative synthesis in the articulatory domain Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, 2012.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2010

Felps, D; Gutierrez-Osuna, R

Normalization of Articulatory Data through Procrustes Transformations and Analysis-by-synthesis Technical Report

2010.

Abstract | Links | BibTeX | Tags: Articulatory synthesis, Speech

---

Gutierrez-Osuna, R; Felps, D

Foreign Accent Conversion through Voice Morphing Technical Report

2010.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Gutierrez-Osuna, R

Developing objective measures of foreign-accent conversion Journal Article

In: Audio, Speech, and Language Processing, IEEE Transactions on, 18 (5), pp. 1030–1040, 2010.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Geng, C; Berger, M; Richmond, K; Gutierrez-Osuna, R

Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database Conference

Interspeech, 2010.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2009

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Foreign accent conversion in computer assisted pronunciation training Journal Article

In: Speech communication, 51 (10), pp. 920–932, 2009.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Pazarloglou, A; Stoleru, R; Gutierrez-Osuna, R

High-resolution speech signal reconstruction in wireless sensor networks Conference

Consumer Communications and Networking Conference, IEEE 2009.

Abstract | Links | BibTeX | Tags: Speech

2008

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Prosodic and segmental factors in foreign-accent conversion Technical Report

2008.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Choi, H; Gutierrez-Osuna, R; Choi, S; Choe, Y

Kernel oriented discriminant analysis for speaker-independent phoneme spaces Conference

International Conference on Pattern Recognition, IEEE 2008.

Abstract | Links | BibTeX | Tags: Speech

2005

Gutierrez-Osuna, R; Kakumanu, P; Esposito, A; Garcia, ON; Bojorquez, A; Castillo, JL; Rudomin, I

Speech-driven facial animation with realistic dynamics Journal Article

In: Multimedia, IEEE Transactions on, 7 (1), pp. 33–42, 2005.

Abstract | Links | BibTeX | Tags: Facial animation, Speech

2001

Kakumanu, P; Gutierrez-Osuna, R; Esposito, A; Bryll, R; Goshtasby, A; Garcia, ON

Speech driven facial animation Conference

Proceedings of the 2001 workshop on Perceptive user interfaces, ACM 2001.

Abstract | Links | BibTeX | Tags: Facial animation, Speech

Physical: 309 Peterson Building, Texas A&M University
Mailing: PSI Lab, 3112 TAMU, College Station, TX 77843

Appendix DING11



Publications

Accent conversion Active sensing Articulatory synthesis Chemical sensors Childhood apraxia of speech Computer vision Continuous glucose monitors Deep learning Diabetes Electrodermal activity Electronic nose Face perception Facial animation Games Gestures Health Heart rate variability Infrared spectroscopy Machine olfaction Metal-oxide sensors Mobile computing Neuromorphic models Olfaction Other Pattern recognition Robotics Speech Stress Temperature modulation Wearable sensors

All years | All types | All authors | All users

2025

Das, A.; Keri, D.; Olantz, N.; Bevier, W.; Santiago, R.; Gutierrez-Osuna, R.; Mortazavi, B. J.

A pilot scientific dataset for personalized nutrition and diet monitoring Journal Article

In: Scientific Data, 2025

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes, Health

---

Guamer, W.; Gutierrez-Osuna, R.

Disentangling Segmental and Prosodic Factors to Non-Native Speech Comprehensibility Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, vol. 33, 2025

Links | BibTeX | Tags: Accent conversion, Speech

K. Vyas, M.-R. Tseng, A. Das, Gutierrez-Osuna, R.

Hypoglycemia prediction in type 1 diabetes with electrocardiography beat ensembles Journal Article

In: Journal of Diabetes Science and Technology, 2025.

Links | BibTeX | Tags: Continuous glucose monitors, Deep learning, Diabetes, Health, Wearable sensors

2024

A. Das, W. Quamer, R. Gutierrez-Osuna

Speech synthesis and pronunciation teaching Book Chapter

In: J. Levis, C. A. Chapelle, M. Munro, Huensch, A. (Ed.), 2024.

Links | BibTeX | Tags: Accent conversion, Speech

Quamer, W., Gutierrez-Osuna, R.

End-to-end streaming model for low-latency speech anonymization Proceedings Article

In: Proc. IEEE Spoken Language Technology Workshop (SLT 2024), 2024.

Links | BibTeX | Tags: Accent conversion, Speech

E. Do, A. Das, N. Glanz, Gutierrez-Osuna, R.

Macronutrient constraints and priors improve carbohydrate predictions from continuous glucose monitors Proceedings Article

In: IEEE EMBS International Conference on Body Sensor Networks (BSN'24), 2024.

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes

Das, A., R. Gutierrez-Osuna

Improving mispronunciation detection using speech reconstruction Journal Article Forthcoming

In: IEEE/ACM Transactions on Audio, Speech and Language Processing, Forthcoming.

Links | BibTeX | Tags: Accent conversion, Speech

```
@article{nokey,
title = {Improving mispronunciation detection using speech reconstruction},
author = {A. Das and R. Gutierrez-Osuna,},
url = {https://psi.engr.tamu.edu/wp-content/uploads/2024/07/anurag-taslp-2024.pdf},
year = {2024},
date = {2024-07-19},
journal = {IEEE/ACM Transactions on Audio, Speech and Language Processing},
keywords = {Accent conversion, Speech},
pubstate = {forthcoming},
tppubtype = {article}
}
```

Close

Dave, D., Vyas, K., Braman, K., McKay, S., DeSalvo, D. J., Gutierrez-Osuna, R., Cole, G. L., Erraguntla, M.

Detection of hypoglycemia and hyperglycemia using ECG and accelerometer based non-invasive wearable sensors Journal Article

In: Journal of Diabetes Science and Technology, vol. 18, iss. 2, 2024.

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes, Wearable sensors

da C. Silva, D. R., Gutierrez-Osuna, R.

Does gamified breath-biofeedback promote adherence, relaxation, and skill transfer in the wild? Journal Article Forthcoming

In: IEEE Transactions on Affective Computing, Forthcoming.

Links | BibTeX | Tags: Games, Stress

```
@article{nokey,
title = {Does gamified breath-biofeedback promote adherence, relaxation, and skill transfer in the wild?},
author = {D. R. da C. Silva and R. Gutierrez-Osuna},
url = {https://psi.engr.tamu.edu/wp-content/uploads/2024/07/Dennis-TAC-GBF-2024.pdf},
year = {2024},
date = {2024-07-15},
journal = {IEEE Transactions on Affective Computing},
keywords = {Games, Stress},
pubstate = {forthcoming},
tppubtype = {article}
}
```

Close

R. Neriz, A. Silpachai, M. Novotny, Chukharev, E.

Corrective feedback accuracy and pronunciation improvement: Feedback that is 'good enough' Journal Article

In: Language Learning & Technology, vol. 28, iss. 1, pp. 1-16, 2024.

Links | BibTeX | Tags: Speech

J. McMurray K. L. Branan; R. Jennings II

Physiological sensing on the upper arm with a wireless multi-modal wearable Proceedings Article

In: Optical Diagnostics and Sensing XXIV: Toward Point-of-Care Diagnostics, 2024

Links | BibTeX | Tags: Diabetes, Wearable sensors

2023

Sicong Huang Lida Zhang, Anurag Das

Joint embedding of food photographs and blood blucose for improved calorie estimation Proceedings Article

In: IEEE-EMBS Int. Conf. Body Sensor Networks (IEEE BSN) , 2023.

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes

Anurag Das Edmund Do, Namino Glanz

Modeling the effect of non-exercise activity thermogenesis on peak post-prandial glucose Proceedings Article

In: IEEE-EMBS International Conference on Body Sensor Networks: Sensors and Systems for Digital Health (IEEE BSN) 2023 , 2023.

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes

Carolina Villegas Kathan Vyas, Elizabeth Kubota-Mishra, Gutierrez-Osuna, Ricardo

Detection of glycemic excursions using morphological and time-domain ECG features Proceedings Article

In: IEEE-EMBS Intl. Conf. Body Sensor Networks(IEEE BSN) 2023 , 2023

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes

Anurag Das Waris Quamer, Ricardo Gutierrez-Osuna

Decoupling segmental and prosodic cues of non-native speech through vector quantization Proceedings Article

In: Proc. Interspeech, 2023

Links | BibTeX | Tags: Accent conversion, Speech, Voice conversion

da C. Silva, D. R.; Wang, Z., Gutierrez-Osuna, R.

Towards Participant-Independent Stress Detection Using Instrumented Peripherals Journal Article

In: IEEE Transactions on Affective Computing, 2023.

Links | BibTeX | Tags: Health, Stress, Wearable sensors

Mortazavi, B. J.; Gutierrez-Osuna, R.

A review of digital innovations for diet monitoring and precision nutrition Journal Article

In: Journal of Diabetes Science and Technology, 2023.

Links | BibTeX | Tags: Chemical sensors, Diabetes, Wearable sensors

2022

RaVuri, Vinesh; Gutierrez-Osuna, Ricardo; Chaspari, Theodora

Preserving Mental Health Information in Speech Anonymization Proceedings Article

In: 10th International Conference on Affective Computing & Intelligent Interaction (ACII2022) , 2022.

Links | BibTeX | Tags: Mental health

Hussein, D. Al; Lin, P. T.; Sukhishvili, S. A., Zhou, J.; Li, J.; Lin, P. T; Gutierrez-Osuna, R.; Cote, G. L.

Silane-modified mesoporous silica nanocoatings for selective mid-infrared waveguide-based gas sensing Journal Article Forthcoming

In: Advanced Materials Interfaces, Forthcoming.

BibTeX | Tags: Chemical sensors

Quamer, W.; Das, A., Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.

Zero-Shot Foreign Accent Conversion without a Native Reference Proceedings Article Forthcoming

In: Proc. Interspeech, Forthcoming.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

Zhou, J.; Hussein, D. Al; Li, J.; Lin, Z; Sukhishvili, S.; Cote, G. L.; Gutierrez-Osuna, R.; Lin, P. T.

Mid-Infrared Serial Microring Resonator Array for Real-time Detection of Vapor Phase Volatile Organic Compounds Journal Article Forthcoming

In: Analytical Chemistry, Forthcoming.

BibTeX | Tags: Chemical sensors, Infrared spectroscopy

Das, A.; Mortazavi, B.; Sajjadi, S.; Chaspari, T., Ruebush, L. E.; Deutz, N. E.; Cote, G. L.; Gutierrez-Osuna, R.

Predicting the macronutrient composition of mixed meals from dietary biomarkers in blood Journal Article

In: IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 6, pp. 2726-2736, 2022.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

Liberatore, C., Gutierrez-Osuna, R.

Minimizing residuals for native-nonnative voice conversion in a sparse, anchor-based representation of speech Proceedings Article

In: Proc. ICASSP, 2022.

BibTeX | Tags: Accent conversion, Speech

Yang, M.; Dave, D.; Erraguntla, M.; Cote, G.; Gutierrez-Osuna, R.

Joint hypoglycemia prediction and glucose forecasting via deep multi-task learning Proceedings Article

In: Proc. ICASSP, 2022.

Links | BibTeX | Tags: Continuous glucose monitors, Deep learning, Health, Wearable sensors

Paul, Sudip; Sharma, Rohit; Tathireddy, Prashant; Gutierrez-Osuna, Ricardo

On-line drift-compensation for continuous monitoring with arrays of cross-sensitive chemical sensors Journal Article Forthcoming

In: Sensors and Actuators B: Chemical, Forthcoming.

BibTeX | Tags: Chemical sensors

Zhou, Junchao; Husseini, Diana Al; Li, Junyan; Lin, Zhihai; Sukhishvili, Svetlana; Cotè, Gerard L.; Gutierrez-Osuna, Ricardo; Lin, Pao Tai

Detection of Volatile Organic Compounds Using Mid-Infrared Silicon Nitride Waveguide Sensors Journal Article

In: Scientific Reports, vol. 12, no. 5572 , 2022.

Links | BibTeX | Tags: Chemical sensors

Branan, Kimberly L.; Reyes, Gilberto O. Flores; Abel, John A.; Erraguntla, Madhav; Gutierrez-Osuna, Ricardo; Cotè, Gerard L.

Multi-modal physiological sensing on the upper arm Proceedings Article

In: Proc. SPIE, 2022.

Links | BibTeX | Tags: Continuous glucose monitors, Health, Wearable sensors

2021

Chakravarthy, N. V.; Silva, D. R. Da Cunha; Gutierrez-Osuna, R.

Evaluating the role of breathing guidance on game-based interventions for relaxation training Journal Article

In: Frontiers in Digital Health, 2021.

Links | BibTeX | Tags: Games, Health, Stress, Wearable sensors

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning Journal Article

In: Computer Speech & Language, 2021.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

Hair, A.; Zhao, G.; Ahmed, B.; Ballard, K. J.; Gutierrez-Osuna, R.

Assessing Posterior-Based Mispronunciation Detection on Field-Collected Recordings from Child Speech Therapy Sessions Proceedings Article

In: Proc. Interspeech, 2021.

Links | BibTeX | Tags: Automatic Speech Recognition, Childhood apraxia of speech, Speech

Sitpachai, A.; Rehman, I.; Barriuso, T. A.; Levis, J.; Chukharev-Hudilainen, E.; Zhao, G.; Gutierrez-Osuna, R.

Effects Of Voice Type And Task On L2 Learners' Awareness Of Pronunciation Errors Proceedings Article

In: Proc. Interspeech, 2021.

BibTeX | Tags: Accent conversion, Speech

Liberatore, C.; Gutierrez-Osuna, R.

An Exemplar Selection Algorithm For Native-Nonnative Voice Conversion Proceedings Article

In: Proc. Interspeech, 2021.

Links | BibTeX | Tags: Accent conversion, Speech

Yang, M.; Paromita, P.; Chaspari, T.; Das, A.; Sajjadi, S.; Mortazavi, B. J.; Gutierrez-Osuna, R.

A Metric Learning Approach for Personalized Meal Macronutrient Estimation from Postprandial Glucose Response Signals Proceedings Article

In: Proc. IEEE/EMBS Intl. Conf. Biomedical And Health Informatics (BHI 2021), 2021.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Deep learning, Health, Wearable sensors

Hagve, M.; Simbo, S. Y.; Ruebush, L. E.; Engelen, M. P. K. J.; Gutierrez-Osuna, R.; Mortazavi, B. J.; Cote, G. L.; Deutz, N. E. P.

Postprandial concentration of circulating branched chain amino acids are able to predict the carbohydrate content of the ingested mixed meal Journal Article

In: Clinical Nutrition, 2021.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

Zhao, G.; Ding, S.; Gutierrez-Osuna, R.

Converting Foreign Accent Speech Without a Reference Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2367, 2021.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

Das, A.; Mortazavi, B.; Chaspari, T.; Sajjadi, S.; Paromita, P.; Ruebush, L.; Deutz, N.; Gutierrez-Osuna, R.

A sparse coding approach to automatic diet monitoring with continuous glucose monitors Proceedings Article

In: In Proc. ICASSP, 2021.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

Sajjadi, S.; Mortazavi, B. J.; Das, A.; Chaspari, T.; Paromita, P.; Ruebush, L. E.; Deutz, N. E.; Gutierrez-Osuna, R.

Towards the development of subject-independent inverse metabolic models Proceedings Article

In: In Proc. ICASSP. 2021.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

---

Paromita, P.; Chaspari, T.; Sajjadi, S.; Das, A.; Mortazavi, B. J.; Gutierrez-Osuna, R.

Personalized Meal Classification Using Continuous Glucose Monitors Proceedings Article

In: In Proc. IUI HEALTHI Workshop, 2021.

BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Deep learning, Health, Wearable sensors

---

Hair, A.; Ballard, K. J.; Markoulli, C.; Monroe, P.; McKechnie, J.; Ahmed, B.; Gutierrez-Osuna, R.

A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World Journal Article

In: ACM Transactions On Accessible Computing, vol. 14, no. 1, 2021.

Abstract | Links | BibTeX | Tags: Games, Health, Speech

Parnandi, A.; Gutierrez-Osuna, R.

Partial reinforcement in game biofeedback for relaxation training Journal Article

In: IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 141-153, 2021.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Wearable sensors

---

Hussein, D. Al; Karanth, Y.; Zhou, J.; Wilhelm, D.; Qian, X.; Gutierrez-Osuna, R.; Cotè, G. L.; and P. Tai Lin; Sukhishvili, S. A.

Surface Functionalization Utilizing Mesoporous Silica Nanoparticles for Enhanced Evanescent-Field Mid-Infrared Waveguide Gas Sensing Journal Article

In: Coatings, vol. 11, no. 118, pp. 1-12, 2021.

Links | BibTeX | Tags: Chemical sensors, Infrared spectroscopy

---

2020

Ahmed, B.; Zafar, M.; Ribawi, R.; Gutierrez-Osuna, R.

Gaming away stress: Using biofeedback games to learn paced breathing Journal Article

In: IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 519-531, 2020.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Stress, Wearable sensors

---

Zaman, S.; Wesley, A.; Silva, D. Rodrigo Da Cunha; Buddharaju, P.; Akbar, F.; Gao, G.; Mark, G.; Gutierrez-Osuna, R.; Pavlidis, I.

Stress and productivity patterns of interrupted, synergistic, and antagonistic office activities Journal Article

In: Scientific Data, vol. 6, no. 1, 2020.

Links | BibTeX | Tags: Electrodermal activity, Health, Heart rate variability, Stress, Wearable sensors

---

Ding, S.; Zhao, G.; Gutierrez-Osuna, R.

Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition Proceedings Article

In: Proc. Interspeech, 2020.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

---

Das, A.; Zhao, G.; Levis, J.; Chukharev-Hudilainen, E.; Gutierrez-Osuna, R.

Understanding the Effect of Voice Quality and Accent on Talker Similarity Proceedings Article

In: Proc. Interspeech, 2020.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

---

Hussein, D. Al; Zhou, J.; Wilhelm, D.; Hastings, T.; Day, G. S.; Zhou, H. -C.; Cotè, G. L.; Qian, X.; R. Gutierrez-Osuna, P. T. Lin; Sukhishvili, S.

All-Nanoparticle Layer-by-layer Coatings for Mid-IR On-Chip Gas Sensing Journal Article

In: Chemical Communications, 2020.

Links | BibTeX | Tags: Chemical sensors

Lučić, I.; Silpachai, A.; Levis, J.; Zhao, G. Gutierrez-Osuna, R.

The English Pronunciation of Arabic Speakers - A Data-Driven Approach to Segmental Error Identification Journal Article

In: Language Teaching Research, 2020.

Links | BibTeX | Tags: Accent conversion, Speech

McKechnie, J.; Ahmed, B.; Gutierrez-Osuna, R.; Murray, E.; McCabe, P.; Ballard, K.

The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech Journal Article

In: Journal of Communication Disorders, 2020.

Links | BibTeX | Tags: Health, Speech

Blank, C.; Zaman, S.; Wesley, A.; Tsiamyrtzis, P.; Silva, D. R. Da Cunha; Gutierrez-Osuna, R.; Mark, G.; Pavlidis, I.

Emotional Footprints of Email Interruptions Proceedings Article

In: Proc. CHI, 2020.

Links | BibTeX | Tags: Electrodermal activity, Health, Stress, Wearable sensors

---

Hair, A.; Markoulli, C.; Monroe, P.; McKechnie, J.; Ballard, K. J.; Ahmed, B.; Gutierrez-Osuna, R.

Preliminary Results From a Longitudinal Study of a Tablet-Based Speech Therapy Game Proceedings Article

In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing. ACM, 2020. ISBN: 978-1-4503-6819-3/20/04.

Abstract | Links | BibTeX | Tags: Automatic Speech Recognition, Childhood apraxia of speech, Health, Speech

2019

Ding, S., Zhao, G. Liberatore, C. Gutierrez-Osuna, R.

Learning Structured Sparse Representations for Voice Conversion Journal Article

In: IEEE Transactions on Audio, Speech and Language Processing, vol. 28, pp. 343-354, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Ding, S., Liberatore, C., Sonsaat, S., Lucic, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis, J., Gutierrez-Osuna, R.

Golden speaker builder – An interactive tool for pronunciation training Journal Article

In: Speech Communication, vol. 115, pp. 51–66, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Hair, A., Ballard, K. J., Ahmed, B., Gutierrez-Osuna, R.

Evaluating Automatic Speech Recognition for Child Speech Therapy Applications Proceedings Article

In: ACM SIGACCESS Conference on Computers and Accessibility, ACM, 2019, ISBN: 978-1-4503-6676-2/19/10.

Abstract | Links | BibTeX | Tags: Automatic Speech Recognition, Childhood apraxia of speech, Health, Speech

---

Zhao, G., Ding, S., Gutierrez-Osuna, Ricardo

Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams Proceedings Article

In: Proc. Interspeech, 2019.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

---

Ding, S., Gutierrez-Osuna, Ricardo

Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion Proceedings Article

In: Proc. Interspeech, 2019.

Links | BibTeX | Tags: Accent conversion, Speech

Akbar, F., Mark, G., Pavlidis, I., Gutierrez-Osuna, R.

An Empirical Study Comparing Unobtrusive Physiological Sensors for Stress Detection in Computer Work Journal Article

In: Sensors, vol. 19, 2019.

Links | BibTeX | Tags: Electrodermal activity, Health, Heart rate variability, Stress, Wearable sensors

---

Zhao, G., Gutierrez-Osuna, R.

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion Journal Article

In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1649-1660, 2019, ISSN: 2329-9290.

Abstract | Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

---

Monteiro, C. D. D., Shipman, F. M. III, S. Duggirir, Gutierrez-Osuna, R.

Tradeoffs in the Efficient Detection of Sign Language Content in Video Sharing Sites Journal Article

In: ACM Transactions on Accessible Computing, vol. 12, no. 2, pp. 1-16, 2019.

Links | BibTeX | Tags: Computer vision, Speech

Huo, Z., Mortazavi, B. J., Chaspari, T., Deutz, N., Ruebush, L., Gutierrez-Osuna, R.

Predicting the meal macronutrient composition from continuous glucose monitors Proceedings Article

In: Proc. IEEE Conf. on Biomedical and Health Informatics, 2019.

BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

---

Parnandi, A., Gutierrez-Osuna, R.

Visual Biofeedback and Game Adaptation in Relaxation Skill Transfer Journal Article

In: IEEE Transactions on Affective Computing, vol. 10, no. 2, pp. 276 - 289, 2019.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Stress, Wearable sensors

---

Akbar, F., Bayraktaroglu, A. E., Gao, G., Grover, T., Mark, G., Storer, K., Silva, D. R. Da Cunha, Gutierrez-Osuna, R., Wang, Z., Buddharaju, P., Jones, N., Cooper, Pavlidis, I., Wesley, A., S. Zaman

Email Makes You Sweat: Examining Email Interruptions and Stress with Thermal Imaging Proceedings Article

In: Proc. CHI, 2019.

Links | BibTeX | Tags: Electrodermal activity, Health, Heart rate variability, Stress, Wearable sensors

2018

Ahmed, B., Monroe, P., Hair, A., Tan, C.-T., Gutierrez-Osuna, R., Ballard, K. J.

Speech-driven mobile games for speech therapy: User experiences and feasibility Journal Article

In: International Journal of Speech-Language Pathology, vol. 20, no. 6, pp. 644-658, 2018.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Health, Speech

---

Levis, J., Chukharev-Hudilainen, E., Gutierrez-Osuna, R., Lucic, I., Silpachai, A., Sonsaat, S.

Golden Speaker: Learner Experience with Computer-assisted Pronunciation Practice Proceedings Article

In: Proc. Pronunciation in Second Language Learning and Teaching Conference, 2018.

BibTeX | Tags: Accent conversion, Speech

Ding, S., Liberatore, C., Gutierrez-Osuna, R.

Learning Structured Dictionaries for Exemplar-based Voice Conversion **Proceedings Article**

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Zhao, G; Sonsaat, S; Silpachai, A; Lucic, I; Chukharev-Hudilainen, E; Levis, J; Gutierrez-Osuna, R

L2-ARCTIC: A Non-Native English Speech Corpus **Proceedings Article**

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

McKechnie, J; Ahmed, B; Gutierrez-Osuna, R; Monroe, P; McCabe, P; Ballard, K J

Automated speech analysis tools for children's speech production: A systematic literature review **Journal Article**

In: International Journal of Speech-Language Pathology, vol. 20, no. 6, pp. 583–598, 2018.

Links | BibTeX | Tags: Childhood apraxia of speech, Health, Speech

---

Ding, S; Zhao, G; Liberatore, C; Gutierrez-Osuna, R

Improving Sparse Representations in Exemplar-Based Voice Conversion with a Phoneme-Selective Objective Function **Proceedings Article**

In: Proc. Interspeech, 2018.

Links | BibTeX | Tags: Accent conversion, Speech

Goel, N; Chaspari, T; Mortazavi, B J; Prioleau, T; A Sabharwal; Gutierrez-Osuna, R

Knowledge-driven dictionaries for sparse representation of continuous glucose monitoring signals **Proceedings Article**

In: Proc. EMBC, 2018.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

---

Hair, A; Monroe, P; Ahmed, B; Ballard, K J; Gutierrez-Osuna, R

Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders **Proceedings Article**

In: Proceedings of the 2018 Conference on Interaction Design and Children, ACM, 2018, ISBN: 978-1-4503-5152-2/18/06.

Abstract | Links | BibTeX | Tags: Childhood apraxia of speech, Health, Mobile computing, Speech

---

Zhao, G; Sonsaat, S; Levis, J; Chukharev-Hudilainen, E; Gutierrez-Osuna, R

Accent conversion using phonetic posteriorgrams **Proceedings Article**

In: Proc. ICASSP 2018.

Links | BibTeX | Tags: Accent conversion, Deep learning, Speech

---

Liberatore, C; Zhao, G; Gutierrez-Osuna, R

Voice Conversion through Residual Warping in a Sparse, Anchor-Based Representation of Speech **Proceedings Article**

In: Proc. ICASSP 2018.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Angelio, G; Zhao, G; Maram, A B; Gutierrez-Osuna, R

Training Behavior of Successful Tacton-Phoneme Learners **Proceedings Article**

In: IEEE Haptics Symposium, 2018.

Links | BibTeX | Tags: Other, Speech

Jin, Tianlong; Zhou, Juntao; Wang, Zetun; Gutierrez-Osuna, Ricardo; Ahn, Charles; Hwang, Wonjun; Park, Ren-Lin; Pao-Tai

Real-time Gas Mixture Analysis Using Mid-infrared Membrane Microcavities **Journal Article**

In: Analytical Chemistry, vol. 90, no. 7, pp. 4348–4353, 2018.

Links | BibTeX | Tags: Chemical sensors, Infrared spectroscopy

---

Chaspari, T; Mortazavi, B; Prioleau, T; Sabharwal, A; Gutierrez-Osuna, R

Sparse representation models of continuous glucose monitoring time-series **Proceedings Article**

In: Proc. 15th International Conference on Wearable and Implantable Body Sensor Networks, 2018.

Links | BibTeX | Tags: Chemical sensors, Continuous glucose monitors, Health, Wearable sensors

---

Wang, Z; Pamardi, A; Gutierrez-Osuna, R

BioPad: Leveraging off-the-Shelf Video Games for Stress Self-Regulation **Journal Article**

In: IEEE Journal of Biomedical and Health Informatics, in press, 2018.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Stress, Wearable sensors

---

2017

Shipman, F; Duggins, S; Monteiro, C; Gutierrez-Osuna, R

Speed-Accuracy Tradeoffs for Detecting Sign Language Content in Video Sharing Sites **Proceedings Article**

In: Proceedings of ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2017), pp. 185–189, 2017.

Links | BibTeX | Tags: Computer vision, Gestures, Speech

Hair, A; Gutierrez-Osuna, R

Deep Breaths: An Internally- and Externally-Paced Deep Breathing Guide **Workshop**

Proc. 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017.

Abstract | Links | BibTeX | Tags: Health, Heart rate variability, Stress, Wearable sensors

Nguyen, K. N., Liu, X., Korniggartseva, O., Gutierrez-Osuna, R., Choe, Y.

Explanation of the Perceptual Oblique Effect Based on the Fidelity of Oculomotor Control During Saccades Proceedings Article

In: 2017 Joint IEEE International Conference on Development and Learning, pp. 18-23, 2017.

Links | BibTeX | Tags: Computer vision

Chesty, M., Gupta, A., Grumbach, E., Mandal, L., Furuta, R., Gutierrez-Osuna, R.

Mass Digitization of Early Modern Texts with Optical Character Recognition Journal Article

In: ACM Journal on Computing and Cultural Heritage, vol. In press, 2017.

Links | BibTeX | Tags: OCR

Kaneamkai, P., Gutierrez-Osuna, R.

Optical Computation of Chemometrics Projections using a Digital Micromirror Device Proceedings Article

In: Proc. International Symposium on Olfaction and Electronic Nose (ISOEN), 2017.

Links | BibTeX | Tags: Active sensing, Infrared spectroscopy

Wang, Z., Gutierrez-Osuna, R.

Mixture quantification in the presence of unknown interferences Proceedings Article

In: Proc. International Symposium on Olfaction and Electronic Nose (ISOEN), 2017.

Links | BibTeX | Tags: Active sensing, Chemical sensors, Infrared spectroscopy

---

Zhao, G., Gutierrez-Osuna, R.

Exemplar selection methods in voice conversion Proceedings Article

In: Proc. 42nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5625-5629, 2017.

Links | BibTeX | Tags: Speech, Voice conversion

Parnandi, A., Gutierrez-Osuna, R.

Physiological modalities for relaxation skill transfer in biofeedback games Journal Article

In: Journal of Biomedical and Health Informatics, vol. In press, 2017.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Stress, Wearable sensors

---

2016

Monteiro, C., Mathew, C., Shipman, F., Gutierrez-Osuna, R.

Detecting and Identifying Sign Languages through Visual Features Proceedings Article

In: 2016 IEEE International Symposium on Multimedia (ISM), 2016.

Links | BibTeX | Tags: Computer vision, Gestures, Pattern recognition

---

Aryal, S., Gutierrez-Osuna, R.

Comparing Articulatory and Acoustic Strategies for Reducing Non-Native Accents Proceedings Article

In: Proc. Interspeech, 2016.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

Liberatore, C., Gutierrez-Osuna, R.

Generating Gestural Scores from Acoustics Through a Sparse Anchor-Based Representation of Speech Proceedings Article

In: Proc. Interspeech, 2016.

Links | BibTeX | Tags: Articulatory synthesis, Speech

Huang, J., Gutierrez-Osuna, R.

Active wavelength selection for mixture identification with tunable mid-infrared detectors Journal Article

In: Analytical Chimica Acta, vol. In press, 2016.

Links | BibTeX | Tags: Active sensing, Chemical sensors, Infrared spectroscopy

---

Li, J., Gutierrez-Osuna, R., Hodges, R. D., Luckey, G., Crowell, J., Schiffman, S. S., Nagle, H. T.

Using Field Asymmetric Ion Mobility Spectrometry for Odor Assessment of Automobile Interior Components Journal Article

In: IEEE Sensors Journal, vol. In press, 2016.

Links | BibTeX | Tags: Chemical sensors, Electronic nose, Machine olfaction

---

Shahin, M., Gutierrez-Osuna, R., Ahmed, B.

Classification of bisyllabic lexical stress patterns in disordered speech using deep learning Proceedings Article

In: Proc. International Conference on Acoustics, Speech, and Signal Processing, 2016.

Links | BibTeX | Tags: Speech

McKechnie, J., Ballard, K. J., McCabe, P., Murray, E., Lan, T., Gutierrez-Osuna, R., Ahmed, B.

Influence of type of feedback on effect of tablet-based delivery of intensive speech therapy in children with Childhood Apraxia of Speech Proceedings Article

In: Proceedings of the Motor Speech Conference, 2016.

BibTeX | Tags: Childhood apraxia of speech, Games, Health, Mobile computing, Speech

---

Aryal, S., Gutierrez-Osuna, R.

Data driven articulatory synthesis with deep neural networks Journal Article

In: Computer Speech and Language, vol. 36, pp. 260-273, 2016.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Deep learning, Speech

2015

Li, J; Gutierrez-Osuna, R; Hodges, R D; Luckey, G; Crowell, J; Schiffman, S S; Nagle, H T.

Odor Assessment of Automobile Interior Components Using Ion Mobility Spectrometry Proceedings Article

In: IEEE Sensors Conference, 2015.

Links | BibTeX | Tags: Chemical sensors, Infrared spectroscopy, Olfaction

Parnandi, A; Karappa, V; Lan, T; Shahin, M; McKechnie, J; Ballard, K; Ahmed, B; Gutierrez-Osuna, R.

Development of a remote therapy tool for childhood apraxia of speech Journal Article

In: ACM Transactions on Accessible Computing, vol. 7, no. 3, pp. 10:1-10:23, 2015.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Health, Mobile computing, Speech

Aryal, S; Gutierrez-Osuna, R.

Articulatory-based conversion of foreign accents with deep neural networks Proceedings Article

In: Proc. Interspeech, pp. 3385-3389, 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Deep learning, Speech

Liberatore, C; Aryal, S; Wang, Z; Polsley, S; Gutierrez-Osuna, R.

SABR: Sparse, Anchor-Based Representation of the Speech Signal Proceedings Article

In: Proc. Interspeech 2015, pp. 608-612, 2015.

Abstract | Links | BibTeX | Tags: Accent conversion, Articulatory synthesis

Ahmed, B; Ali, H; Choi, J; Gutierrez-Osuna, R.

ReBreathe: A calibration protocol that improves stress/relax classification by relabeling deep breathing relaxation exercises Journal Article

In: IEEE Transactions on Affective Computing, vol. in press, 2015.

Links | BibTeX | Tags: Health, Stress, Wearable sensors

Huang, J; Gutierrez-Osuna, R.

Detection of weak chemicals in strong backgrounds with a tunable infrared sensor Proceedings Article

In: International Symposium on Olfaction and Electronic Nose, 2015.

BibTeX | Tags: Active sensing, Chemical sensors

Nagle, H T; Gutierrez-Osuna, R; Sudick, K S; Persaud, K; Covington, J; Hodges, R D; Luckey, G; Crowell, J; Schiffman, S S.

Augmenting human odor assessments of cabin air quality of automobiles by instrumental measurements Proceedings Article

In: International Symposium on Olfaction and Electronic Nose, 2015.

Links | BibTeX | Tags: Chemical sensors, Machine olfaction

Bhandari, R; Parnandi, A; Shipp, E; Ahmed, B; Gutierrez-Osuna, R.

Music-based respiratory biofeedback in visually-demanding tasks Proceedings Article

In: 15th International Conference on New Interfaces for Musical Expression (NIME), 2015.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Stress, Wearable sensors

Liberatore, C; Gutierrez-Osuna, R.

Joint Optimization of Anatomical and Gestural Parameters in a Physical Vocal Tract Model Proceedings Article

In: ICASSP, IEEE 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory inversion, Articulatory synthesis

Shahin, M; Ahmed, B; Parnandi, A; Karappa, V; McKechnie, J; Ballard, K; Gutierrez-Osuna, R.

Tabby Talks: an automated tool for the assessment of childhood apraxia of speech Journal Article

In: Speech Communication, vol. in press, 2015.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Health, Mobile computing, Speech

Shipman, F; Gutierrez-Osuna, R; Shipman, T; Monteiro, C; Karappa, V.

Towards a Distributed Digital Library for Sign Language Content Proceedings Article

In: Proc. ACM/IEEE Joint Conference on Digital Libraries, 2015.

Links | BibTeX | Tags: Computer vision, Gestures, Pattern recognition

Gupta, A; Gutierrez-Osuna, R; Christy, M; Capitanu, B; Auvil, L; Grumbach, L; Furuta, R; Mandell, L.

Automatic assessment of OCR quality in historical documents Proceedings Article

In: Proc. AAAI, 2015.

Links | BibTeX | Tags: OCR, Other, Pattern recognition

Aryal, S; Gutierrez-Osuna, R.

Reduction of non-native accents through statistical parametric articulatory synthesis Journal Article

In: Journal of the Acoustical Society of America, vol. 137, no. 1, pp. 433-446, 2015.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

Wang, Z; Wang, J; Zhang, S; Gong, Y.

Visual Tracking based on Online Sparse Feature Learning **Journal Article**

In: Image and Vision Computing, vol. 38, no. 24--32, 2015.

Abstract | Links | BibTeX | Tags:

Huang, J; Gutierrez-Osuna, R

Active wavelength selection for mixture analysis with tunable infrared detectors **Journal Article**

In: Sensors and Actuators B: Chemical, vol. 209, pp. 245--257, 2015.

Links | BibTeX | Tags: Active sensing, Chemical sensors, Infrared spectroscopy

---

2014

Li, J; Hodges, R D; Gutierrez-Osuna, R; Luckey, G; Crowell, J; Schiffman, S S; Nagle, H T

Odor Assessment of Automobile Cabin Air by Machine Olfaction **Proceedings Article**

In: Proc. IEEE Sensors Conference, 2014.

Links | BibTeX | Tags: Chemical sensors, Electronic nose, Olfaction

---

Lam, T; Aryal, S; Ahmed, B; Ballard, K; Gutierrez-Osuna, R

Flappy Voice: An Interactive Game for Childhood Apraxia of Speech Therapy **Proceedings Article**

In: Proc. CHI-PLAY, 2014.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Health, Speech

---

Al-Rihawi, R; Ahmed, B; Gutierrez-Osuna, R

Dodging Stress With A Personalized Biofeedback Game **Proceedings Article**

In: Proc. CHI-PLAY, 2014.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Stress, Wearable sensors

---

Shahin, M; Ahmed, B; McKechnie, J; Ballard, K; Gutierrez-Osuna, R

A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech **Proceedings Article**

In: Proc. Interspeech, 2014.

Links | BibTeX | Tags: Childhood apraxia of speech, Health, Speech

---

Christy, M; Auvil, L; Gutierrez-Osuna, R; Capitanu, B; Gupta, A; Grumbach, E

Diagnosing Page Image Problems with Post-OCR Triage for eMOP **Proceedings Article**

In: Proc. Digital Humanities Conference, 2014.

BibTeX | Tags: OCR, Other, Pattern recognition

Avrett, M S; Phillips, D; Bhattacharya, R; Gutierrez-Osuna, R; Malak, R; Mayer, R; Nteimo, L; Quadrifoglio, L; Tsvetkov, P

Toward an Interdisciplinary Program in Systems Engineering at Texas A&M University **Proceedings Article**

In: Proc. 4th International Engineering Systems Symposium (CESUN 2014), 2014.

BibTeX | Tags: Other

McKechnie, J; Ballard, K; McCabe, P; Gutierrez-Osuna, R; Karappa, V; Parnandi, A; Shahin, M; Murray, E; Ahmed, B

Tablet-based delivery of intensive speech therapy in children with Childhood Apraxia of Speech - Pilot Phase **Proceedings Article**

In: Speech Pathology Australia National Conference, 2014.

BibTeX | Tags: Games, Health, Speech

Karappa, V; Montairo, C; Shipman, F; Gutierrez-Osuna, R

Detection of sign-language content in video through polar motion profiles **Proceedings Article**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1290-1303, 2014.

Links | BibTeX | Tags: Computer vision, Gestures

Aryal, S; Gutierrez-Osuna, R

Accent conversion through cross-speaker articulatory synthesis **Proceedings Article**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7744-7748, 2014.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

Aryal, S; Gutierrez-Osuna, R

Can voice conversion be used to reduce non-native accents **Proceedings Article**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7929-7933, 2014.

Links | BibTeX | Tags: Accent conversion, Speech

Felps, D; Aryal, S; Gutierrez-Osuna, R

Normalization of articulatory data through Procrustes transformations and analysis-by-synthesis **Proceedings Article**

In: Proc. 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3051-3055, 2014.

Links | BibTeX | Tags: Accent conversion, Articulatory synthesis, Speech

---

Parnandi, A; Gutierrez-Osuna, R

A Comparative Study of Game Mechanics and Control Laws for an Adaptive Physiological Game **Journal Article**

In: Journal on Multimodal User Interfaces, 2014.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Stress, Wearable sensors

Harris, J; Vance, S; Fernandes, O; Parnandi, A; Gutierrez-Osuna, R

Sonic Respiration: Controlling Respiration Rate Through Auditory Biofeedback Proceedings Article

In: Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI 2014) Works-in-Progress, 2014

Links | BibTeX | Tags: Games, Health, Heart rate variability, Stress, Wearable sensors

---

Perera-Lluna, A; Manivannan, K; Xu, P; Gutierrez-Osuna, R; Benner, C; Russell, B D

Automatic Capacitor Bank Identification in Power Distribution Systems Journal Article

In: Electric Power Systems Research, vol. in press, 2014.

Links | BibTeX | Tags: Other, Pattern recognition

Yu, Y; Choe, Y; Gutierrez-Osuna, R

Context-sensitive Intra-class Clustering Journal Article

In: Pattern Recognition Letters, vol. 37, pp. 85-93, 2014.

Links | BibTeX | Tags: Pattern recognition

Gosangi, R; Gutierrez-Osuna, R

Active classification with arrays of tunable chemical sensors Journal Article

In: Chemometrics and Intelligent Laboratory Systems, vol. 132, pp. 91-103, 2014.

Links | BibTeX | Tags: Active sensing, Metal-oxide sensors

2013

Parnandi, A; Ahmed, B; Shipp, E; Gutierrez-Osuna, R

Chill-Out: Relaxation training through respiratory biofeedback in a mobile casual game Conference

Fifth International Conference on Mobile Computing, Applications and Services (MobiCASE 2013), 2013

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Heart rate variability, Stress, Wearable sensors

---

Shipman, F; Gutierrez-Osuna, R; Monteiro, C

Identifying Sign Language Videos in Video Sharing Sites Journal Article

In: ACM Transactions on Accessible Computing, vol. in press, 2013.

Links | BibTeX | Tags: Computer vision, Gestures

Parnandi, A; Karappa, V; Son, Y; Shahin, M; McKechnie, J; Ballard, K; Ahmed, B; Gutierrez-Osuna, R

Architecture of an automated therapy tool for childhood apraxia of speech Conference

The 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS), 2013.

Links | BibTeX | Tags: Childhood apraxia of speech, Games, Health, Mobile computing, Speech

---

Parnandi, A; Son, Y; Gutierrez-Osuna, R

A Control-Theoretic Approach to Adaptive Physiological Games Conference

Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), 2013.

Links | BibTeX | Tags: Electrodermal activity, Games, Health, Stress, Wearable sensors

---

Parnandi, A; Gutierrez-Osuna, R

Contactless Measurement of Heart Rate Variability from Fluctuations in Pupillary Dilation Conference

Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), 2013.

Links | BibTeX | Tags: Contactless sensing, Health, Heart rate variability, Stress

---

Aryal, S; Felps, D; Gutierrez-Osuna, R

Foreign Accent Conversion through Voice Morphing Proceedings Article

In: Interspeech, pp. 3077-3081, 2013.

Links | BibTeX | Tags: Accent conversion, Speech

Khan, H; Ahmed, B; Choi, J; Gutierrez-Osuna, R

Using an Ambulatory Stress Monitoring Device to Identify Relaxation Due to Untrained Deep Breathing Conference

35th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, 2013

Abstract | Links | BibTeX | Tags: Stress, Wearable sensors

Gosangi, R; Gutierrez-Osuna, R

Active temperature modulation of metal-oxide sensors for quantitative analysis of gas mixtures Journal Article

In: Sensors and Actuators B: Chemical, vol. 185, pp. 201-210, 2013.

Links | BibTeX | Tags: Active sensing, Chemical sensors, Metal-oxide sensors

---

Aryal, S; Gutierrez-Osuna, R

Articulatory inversion and synthesis: towards articulatory-based modification of speech Proceedings Article

In: 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7952-7956, 2013.

Links | BibTeX | Tags: Articulatory inversion, Articulatory synthesis, Speech

---

Huang, J; Gutierrez-Osuna, R

Active analysis of chemical mixtures with multi-modal sparse non-negative least squares Conference

38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2013.

Links | BibTeX | Tags: Active sensing, Chemical sensors, Infrared spectroscopy

---

Lee, J; Gutierrez-Osuna, R; Yeung, S.

SILK: Scale-space integrated Lucas-Kanade image registration for super-resolution from video Conference

38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2013.

Links | BibTeX | Tags: Computer vision

2012

Gosangi, R; Gutierrez-Osuna, R.

Active Decomposition and Sensing in Networks of Distributed Chemical Sensors Technical Report

2012.

Abstract | Links | BibTeX | Tags: Active sensing, Chemical sensors

---

Huang, J; Gutierrez-Osuna, R.

Active Analysis of Chemical Mixtures with Multi-modal Sparse Non-negative Least Sqares Technical Report

2012.

Abstract | Links | BibTeX | Tags: Active sensing, Chemical sensors, Infrared spectroscopy

---

Parnandi, A; Gutierrez-Osuna, R.

Contactless Measurement of Heart Rate Variability from Pupillary Fluctuations Technical Report

2012.

Abstract | Links | BibTeX | Tags: Contactless sensing, Health, Heart rate variability, Stress, Wearable sensors

---

Aryal, S; Gutierrez-Osuna, R.

Articulatory Inversion and Synthesis: Towards Articulatory-Based Modification of Speech Technical Report

2012.

Abstract | Links | BibTeX | Tags: Articulatory inversion, Articulatory synthesis, Speech

---

Aryal, S; Huang, J; Felps, D; Gutierrez-Osuna, R.

Boosting Automatic Speech Recognition through Articulatory Inversion Technical Report

2012.

Abstract | Links | BibTeX | Tags: Articulatory inversion

Montano, C; Gutierrez-Osuna, R; Shipman, F.

Design and Evaluation of Classifier for Identifying Sign Language Videos in Video Sharing Sites Conference

13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2012), 2012.

Abstract | Links | BibTeX | Tags: Computer vision, Gestures

Son, Y; Parnandi, A; Gutierrez-Osuna, R.

A Control-Theoretic Approach to Adaptive Physiological Games Technical Report

2012.

Abstract | Links | BibTeX | Tags: Electrodermal activity, Games, Health, Stress, Wearable sensors

---

Huang, J; Gosangi, R; Gutierrez-Osuna, R.

Active Concentration-Independent Chemical Identification with a Tunable Infrared Sensor Journal Article

In Sensors Journal, IEEE, 2012.

Abstract | Links | BibTeX | Tags: Active sensing, Chemical sensors, Infrared spectroscopy

---

Masood, K; Choi, J; Ahmad, B; Gutierrez-Osuna, R.

Consistency and Validity of Self-reporting Scores in Stress Measurement Surveys Conference

2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), 2012.

Abstract | Links | BibTeX | Tags: Health, Stress, Wearable sensors

---

Parnandi, A; Son, Y; Shahin, M; Ahmad, B; Gutierrez-Osuna, R.

Architecture of an Automated Therapy Tool for Childhood Apraxia of Speech Technical Report

2012.

Abstract | Links | BibTeX | Tags: Games, Health, Mobile computing, Speech

---

Alamudun, F; Choi, J; Gutierrez-Osuna, R; Khan, H; Ahmed, B.

Removal of Subject-Dependent and Activity-Dependent Variation in Physiological Measures of Stress Conference

Proceedings of Pervasive Computing Technologies for Healthcare, 2012. ISBN: 978-1-936968-43-5

Abstract | Links | BibTeX | Tags: Electrodermal activity, Health, Heart rate variability, Stress, Wearable sensors

---

Choi, J; Ahmed, B; Gutierrez-Osuna, R.

Development and Evaluation of an Ambulatory Stress Monitor Based on Wearable Sensors Journal Article

In IEEE Transactions on Information Technology in Biomedicine, no 99, pp. 279 - 286, 2012.

Abstract | Links | BibTeX | Tags: Electrodermal activity, Health, Heart rate variability, Stress, Wearable sensors

---

Lucchese, G; Field, M; Ho, J; Gutierrez-Osuna, R; Hammond, T.

GestureCommander: continuous touch-based gesture prediction **Conference**

Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts. ACM 2012.

Abstract | Links | BibTeX | Tags: Gestures, Mobile computing

Felps, D; Gong, C; Gutierrez-Osuna, R

Foreign accent conversion through concatenative synthesis in the articulatory domain **Journal Article**

In: IEEE Transactions on Audio, Speech and Language Processing, 2012.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

2011

Lee, J; Young, S S; Gutierrez-Osuna, R

An Iterative Image Registration Technique Using a Scale-Space Model **Technical Report**

2011.

Abstract | Links | BibTeX | Tags: Computer vision

Choi, J; Gutierrez-Osuna, R

Removal of respiratory influences from heart rate variability in stress monitoring **Journal Article**

In: IEEE Sensors Journal, vol. 11, no. 11, pp. 2649-2656, 2011.

Abstract | Links | BibTeX | Tags: Health, Heart rate variability, Stress, Wearable sensors

---

Gutierrez-Osuna, R; Gosangi, R; Hierlemann, A

Invited: Advances in Active and Adaptive Chemical Sensing **Conference**

Proceedings of the 14th International Symposium on Olfaction and Electronic Nose, 2011.

Abstract | Links | BibTeX | Tags: Active sensing

Gosangi, R; Gutierrez-Osuna, R

Quantification of Gas Mixtures with Active Recursive Estimation **Conference**

Proceedings of the 14th International Symposium on Olfaction and Electronic Nose, 2011.

Abstract | Links | BibTeX | Tags: Active sensing, Metal-oxide sensors

---

Huang, J; Gosangi, R; Gutierrez-Osuna, R

Active Sensing with Fabry-Perot Infrared Interferometers **Conference**

Proceedings of the 14th International Symposium on Olfaction and Electronic Nose, 2011.

Abstract | Links | BibTeX | Tags: Active sensing, Infrared spectroscopy

---

Boulos, M N K; Blanchard, B J; Walker, C; Montero, J; Tripathy, A; Gutierrez-Osuna, R

Web GIS in practice X: a Microsoft Kinect natural user interface for Google Earth navigation **Journal Article**

In: International Journal of Health Geographics, vol. 10, no. 1, pp. 45, 2011.

Abstract | Links | BibTeX | Tags: Gestures

Rodriguez, J; Gutierrez-Osuna, R

Reverse caricatures effects on three-dimensional facial reconstructions **Journal Article**

In: Image and Vision Computing, vol. 29, no. 5, pp. 329-334, 2011.

Abstract | Links | BibTeX | Tags: Face perception

Slipman, F; Park, S I; Gutierrez-Osuna, R; Choi, J

PerCon: Supporting the Management and Reuse of Wearable Sensor Data **Technical Report**

2011.

Abstract | Links | BibTeX | Tags: Health, Wearable sensors

Gosangi, R; Gutierrez-Osuna, R

Data-driven Modeling of Metal-oxide Sensors with Dynamic Bayesian Networks **Conference**

Proceedings of the 14th International Symposium on Olfaction and Electronic Nose, 2011.

Abstract | Links | BibTeX | Tags: Active sensing, Metal-oxide sensors

---

2010

Choi, J; Ahmed, B; Gutierrez-Osuna, R

Ambulatory Stress Monitoring with Minimally-Invasive Wearable Sensors **Technical Report**

2010.

Abstract | Links | BibTeX | Tags: Electrodermal activity, Health, Heart rate variability, Stress, Wearable sensors

---

Felps, D; Gutierrez-Osuna, R

Normalization of Articulatory Data through Procrustes Transformations and Analysis-by-synthesis **Technical Report**

2010.

Abstract | Links | BibTeX | Tags: Articulatory synthesis, Speech

---

Gutierrez-Osuna, R; Felps, D

Foreign Accent Conversion through Voice Morphing **Technical Report**

2010

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Yu, N.Y.; Yamauchi, T.; Yang, H.R.; Chen, Y.L.; Gutierrez-Osuna, R.

Feature selection for inductive generalization **Journal Article**

In: Cognitive science, vol. 34, no. 8, pp. 1574–1593, 2010.

Abstract | Links | BibTeX | Tags: Face perception

Gosangi, R.; Gutierrez-Osuna, R.

Energy-aware active chemical sensing **Conference**

Proceedings of IEEE Sensors, IEEE 2010

Abstract | Links | BibTeX | Tags: Active sensing, Chemical sensors, Metal-oxide sensors, Temperature modulation

---

Choi, J.; Gutierrez-Osuna, R.

Estimating mental stress using a wearable cardio-respiratory sensor **Conference**

Proceedings of IEEE Sensors, IEEE 2010

Abstract | Links | BibTeX | Tags: Health, Heart rate variability, Stress, Wearable sensors

---

Gutierrez-Osuna, R.; Hierlemann, A.

Adaptive Microsensor Systems **Journal Article**

In: Annual Review of Analytical Chemistry, vol. 3, pp. 255–276, 2010

Abstract | Links | BibTeX | Tags: Active sensing, Chemical sensors

---

Felps, D.; Gutierrez-Osuna, R.

Developing objective measures of foreign-accent conversion **Journal Article**

In: Audio, Speech, and Language Processing, IEEE Transactions on, vol. 18, no. 5, pp. 1030–1040, 2010

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Gosangi, R.; Gutierrez-Osuna, R.

Active temperature programming for metal-oxide chemoresistors **Journal Article**

In: Sensors Journal, IEEE, vol. 10, no. 6, pp. 1075–1083, 2010

Abstract | Links | BibTeX | Tags: Active sensing, Chemical sensors, Metal-oxide sensors, Temperature modulation

---

Felps, D.; Geng, C.; Berger, M.; Richmond, K.; Gutierrez-Osuna, R.

Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database **Conference**

Interspeech, 2010

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Gutierrez-Osuna, R.

Electronic Nose **Book Chapter**

In: SAGE Encyclopedia of Perception, Sage Publications, 2010

Abstract | Links | BibTeX | Tags: Electronic nose

2009

Felps, D.; Bortfeld, H.; Gutierrez-Osuna, R.

Foreign accent conversion in computer assisted pronunciation training **Journal Article**

In: Speech communication, vol. 51, no. 10, pp. 920–932, 2009

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Rodriguez, J.; Bortfeld, H.; Ruidomin, I.; Hernández, B.; Gutierrez-Osuna, R.

The reverse-caricature effect revisited: Familiarization with frontal facial caricatures improves veridical face recognition **Journal Article**

In: Applied cognitive psychology, vol. 23, no. 5, pp. 733–742, 2009

Abstract | Links | BibTeX | Tags: Face perception

Choi, J.; Gutierrez-Osuna, R.

Using heart rate monitors to detect mental stress **Conference**

Sixth International Workshop on Wearable and Implantable Body Sensor Networks, IEEE 2009

Abstract | Links | BibTeX | Tags: Stress, Wearable sensors

Gosangi, R.; Gutierrez-Osuna, R.

Active chemical sensing with partially observable Markov decision processes **Conference**

Proceedings of 13th International Symposium on Olfaction and Electronic Noses, 2009

Abstract | Links | BibTeX | Tags: Active sensing

Pazarloglou, A.; George, S.; Stoleru, R.; Gutierrez-Osuna, R.

Demo abstract: Signal reconstruction with subnyquist sampling using wireless sensor networks **Conference**

International Conference on Information Processing in Sensor Networks, IEEE 2009

Abstract | Links | BibTeX | Tags: Other

Pazalioglou, A; Stoleru, R; Gutierrez-Osuna, R

High-resolution speech signal reconstruction in wireless sensor networks Conference

Consumer Communications and Networking Conference, IEEE 2009.

Abstract | Links | BibTeX | Tags: Speech

Raman, B; Gutierrez-Osuna, R

Relating sensor responses of odorants to their organoleptic properties by means of a biologically-inspired model of receptor neuron convergence onto olfactory bulb Book Chapter

In: Biologically inspired signal processing for chemical sensing, pp. 93–108, Springer Berlin/Heidelberg, 2009.

Abstract | Links | BibTeX | Tags: Neuromorphic models

Yu, N Y; Yamauchi, T; Gutierrez-Osuna, R

Similarity perception of visual objects: A machine-learning approach Conference

Proceedings of International Conference on Asia Pacific Psychology, 2009.

Links | BibTeX | Tags: Face perception

2008

Yamanaka, T; Gutierrez-Osuna, R

Extracting functional clusters of glomeruli in rat olfactory bulb by non-negative matrix factorization Technical Report

2008.

Abstract | Links | BibTeX | Tags: Olfaction

Yamanaka, T; Perera-Lluna, A; Raman, B; Gutierrez-Galvez, A; Gutierrez-Osuna, R

Learning Sparse Basis Vectors in Small-Sample Datasets through Regularized Non-Negative Matrix Factorization Technical Report

2008.

Abstract | Links | BibTeX | Tags: Face perception

Choi, J; Gutierrez-Osuna, R

Estimating the principal dynamic modes of autonomic state with wearable sensors Technical Report

2008.

Abstract | Links | BibTeX | Tags: Wearable sensors

Felps, D; Bortfeld, H; Gutierrez-Osuna, R

Prosodic and segmental factors in foreign-accent conversion Technical Report

2008.

Abstract | Links | BibTeX | Tags: Accent conversion, Speech

Choi, H; Gutierrez-Osuna, R; Choi, S; Choe, Y

Kernel oriented discriminant analysis for speaker-independent phoneme spaces Conference

International Conference on Pattern Recognition, IEEE 2008.

Abstract | Links | BibTeX | Tags: Speech

Rodriguez, J; Bortfeld, H; Gutierrez-Osuna, R

Reducing the other-race effect through caricatures Conference

IEEE International Conference on Automatic Face & Gesture Recognition, IEEE 2008.

Abstract | Links | BibTeX | Tags: Face perception

Paulson, B; Rajan, P; Davalos, P; Gutierrez-Osuna, R; Hammond, T

What!?! no Rubine features?: using geometric-based features to produce normalized confidence values for sketch recognition Conference

HCC Workshop: Sketch Tools for Diagramming, 2008.

Abstract | Links | BibTeX | Tags: Gestures

Hierlemann, A; Gutierrez-Osuna, R

Higher-order chemical sensing Journal Article

In: Chemical reviews, vol. 108, no. 2, pp. 563, 2008.

Links | BibTeX | Tags: Active sensing, Chemical sensors

Yu, N Y; Yamauchi, T; Yang, H F; Chen, Y L; Gutierrez-Osuna, R

A Computational Method to Find Salient Features Conference

Proceedings of the 6th International Conference of Cognitive Science, 2008.

Abstract | Links | BibTeX | Tags: Face perception

2007

Koh, S; Caruso, D; Keene, A; Gutierrez-Osuna, R

Elimination of junk document surrogate candidates through pattern recognition Conference

Proceedings of the 2007 ACM symposium on Document engineering, ACM 2007.

Abstract | Links | BibTeX | Tags: Pattern recognition

Nogueira, F. G; Felps, D; Gutierrez-Osuna, R

Development of an infrared absorption spectroscope based on linear variable filters Journal Article

In: Sensors Journal, IEEE, vol. 7, no. 8, pp. 1183–1190, 2007.

Abstract | Links | BibTeX | Tags: Infrared spectroscopy

Raman, B; Kotsopoulou, T; Clark, L; Léal, M; Gutierrez-Osuna, R

Neuromorphic processing for optical microbead arrays: dimensionality reduction and contrast enhancement Journal Article

In: Sensors Journal, IEEE, vol. 7, no. 4, pp. 506–514, 2007.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

2006

Raman, B; Yamanaka, T; Gutierrez-Osuna, R

Contrast enhancement of gas sensor array patterns with a neurodynamics model of the olfactory bulb Journal Article

In: Sensors and Actuators B: Chemical, vol. 119, no. 2, pp. 547–555, 2006.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Gutierrez-Galvez, A; Gutierrez-Osuna, R

Contrast enhancement and background suppression of chemosensor array patterns with the KIII model Journal Article

In: International Journal of Intelligent Systems, vol. 21, no. 9, pp. 937–953, 2006.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Gutierrez-Galvez, A; Gutierrez-Osuna, R

Increasing the separability of chemosensor array patterns with Hebbian/anti-Hebbian learning Journal Article

In: Sensors and Actuators B: Chemical, vol. 116, no. 1, pp. 29–35, 2006.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Raman, B; Gutierrez-Osuna, R

Concentration normalization with a model of gain control in the olfactory bulb Journal Article

In: Sensors and Actuators B: Chemical, vol. 116, no. 1, pp. 36–42, 2006.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Perera-Lluna, A; Yamanaka, T; Gutierrez-Galvez, A; Raman, B; Gutierrez-Osuna, R

A dimensionality-reduction technique inspired by receptor convergence in the olfactory system Journal Article

In: Sensors and Actuators B: Chemical, vol. 116, no. 1, pp. 17–22, 2006.

Abstract | Links | BibTeX | Tags: Neuromorphic models

Raman, B; Sun, P A; Gutierrez-Galvez, A; Gutierrez-Osuna, R

Processing of chemical sensor arrays with a biologically inspired model of olfactory coding Journal Article

In: Neural Networks, IEEE Transactions on, vol. 17, no. 4, pp. 1015–1024, 2006.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Kakumanu, P; Esposito, A; Garcia, O N; Gutierrez-Osuna, R

A comparison of acoustic coding models for speech-driven facial animation Journal Article

In: Speech communication, vol. 48, no. 6, pp. 598–615, 2006.

Abstract | Links | BibTeX | Tags: Facial animation

2005

Raman, B; Gutierrez-Osuna, R

Concentration normalization with a model of gain control in the olfactory bulb Journal Article

In: Sensors and Actuators B: Chemical, vol. 116, no. 1, pp. 36–42, 2005.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models, Temperature modulation

---

Yu, Y; Gutierrez-Osuna, R; Choe, Y

Preserving Class Discriminatory Information by Context-Sensitve Intra-Class Clustering Algorithm Technical Report

2005.

Abstract | Links | BibTeX | Tags: Pattern recognition

Raman, B; Gutierrez-Osuna, R

Mixture segmentation and background suppression in chemosensor arrays with a model of olfactory bulb-cortex interaction Conference

Proceedings of IEEE International Joint Conference on Neural Networks, IEEE 2005.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Gutierrez-Galvez, A; Gutierrez-Osuna, R

Contrast enhancement of sensor-array patterns through hebbian/antihebbian learning Conference

Proceedings of the 11th International Symposium on Olfaction and Electronic Nose, 2005.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Fu, S; Gutierrez-Osuna, R; Esposito, A; Kakumanu, P; Garcia, O N

Audio/visual mapping with cross-modal hidden Markov models **Journal Article**

In: Multimedia, IEEE Transactions on, vol. 7, no. 2, pp. 243–252, 2005.

Abstract | Links | BibTeX | Tags: Facial animation

Gutierrez-Osuna, R; Kakumanu, P; Esposito, A; Garcia, ON; Bojorquez, A; Castillo, JL; Rudomin, I

Speech-driven facial animation with realistic dynamics **Journal Article**

In: Multimedia, IEEE Transactions on, vol. 7, no. 1, pp. 33–42, 2005.

Abstract | Links | BibTeX | Tags: Facial animation, Speech

Perera-Lluna, A; Yamanaka, T; Gutierrez-Galvez, A; Raman, B; Gutierrez-Osuna, R

A dimensionality-reduction technique inspired by receptor convergence in the olfactory system **Conference**

Proceedings of the 11th International Symposium on Olfaction and Electronic Nose, 2005.

Abstract | Links | BibTeX | Tags: Neuromorphic models

2004

Gutierrez-Galvez, A; Gutierrez-Osuna, R; Raman, B

Pattern recognition for chemosensor arrays with the KIII model **Conference**

Proceedings of the 2004 Symposium on Intentional Dynamic Systems, 2004.

Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

Gutierrez-Osuna, R; Raman, B

Cancellation of chemical backgrounds with generalized Fisher's linear discriminants **Conference**

Proceedings of IEEE Sensors, IEEE 2004.

Abstract | Links | BibTeX | Tags: Chemical sensors

Wang, M; Perera-Lluna, A; Gutierrez-Osuna, R

Principal Discriminants Analysis for small-sample-size problems: application to chemical sensing **Conference**

Proceedings of IEEE Sensors, IEEE 2004.

Abstract | Links | BibTeX | Tags: Chemical sensors

Raman, B; Gutierrez-Galvez, A; Perera-Lluna, A; Gutierrez-Osuna, R

Sensor-based machine olfaction with a neurodynamics model of the olfactory bulb **Conference**

Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE 2004.

Abstract | Links | BibTeX | Tags: Machine olfaction, Metal-oxide sensors, Neuromorphic models

---

Rodriguez-Mendez, M L; Arrieta, A; Parra, V; Bernal, A; Vegas, A; Villanueva, S; Gutierrez-Osuna, R; de Saja, J A

Fusion of three sensory modalities for the multimodal characterization of red wines **Journal Article**

In: Sensors Journal, IEEE, vol. 4, no. 3, pp. 348–354, 2004.

Abstract | Links | BibTeX | Tags: Chemical sensors, Electronic nose

---

Pasini, P; Power, N; Gutierrez-Osuna, R; Daunert, S; Roda, A

Use of a gas-sensor array for detecting volatile organic compounds (VOC) in chemically induced cells **Journal Article**

In: Analytical and bioanalytical chemistry, vol. 378, no. 1, pp. 76–83, 2004.

Abstract | Links | BibTeX | Tags: Chemical sensors, Metal-oxide sensors, Temperature modulation

---

Raman, B; Gutierrez-Osuna, R

Chemosensory processing in a spiking model of the olfactory bulb: chemotopic convergence and center surround inhibition **Conference**

NIPS, 2004.

Abstract | Links | BibTeX | Tags: Neuromorphic models

Gutierrez-Osuna, R

Olfactory Interaction **Book Chapter**

In: Encyclopedia of Human-Computer Interaction, pp. 507–511, Berkshire Publishing Group, 2004.

BibTeX | Tags: Olfaction

2003

Fiorkey, L; Gutierrez-Osuna, R; Bullock, J D

Computer-Aided Pattern Recognition in the Classification of Superficial Corneal Injuries **Conference**

Proceedings of the Annual Meeting of the American Academy of Optometry, 2003.

BibTeX | Tags: Pattern recognition

Pasini, P; Powar, N; Gutierrez-Osuna, R; Daunert, S; Roda, A

Use of an electronic nose for the identification of genetically modified organisms **Conference**

Proceedings of the 21st National Congress of the Italian Chemical Society, 2003.

BibTeX | Tags: Electronic nose

Gutierrez-Osuna, R; Gutierrez-Galvez, A

Habituation in the KIII olfactory model with chemical sensor arrays Journal Article

In: Neural Networks, IEEE Transactions on, vol. 14, no. 6, pp. 1568–1568, 2003.

Abstract | Links | BibTeX | Tags: Chemical sensors, Neuromorphic models

---

Moutavi, M. J.; Butler-Purry, K. L.; Gutierrez-Osuna, R.; Najafi, M.

Classification of load change transients and incipient abnormalities in underground cable using pattern analysis techniques Conference

Proceedings of 2003 IEEE PES Transmission and Distribution Conference and Exposition, IEEE 2003.

Abstract | Links | BibTeX | Tags: Pattern recognition

Gutierrez-Osuna, R.; Gutierrez-Galvez, A.; Powar, N.

Transient response analysis for temperature-modulated chemoresistors Journal Article

In: Sensors and Actuators B: Chemical, vol. 93, no. 1–3, pp. 57–66, 2003.

Abstract | Links | BibTeX | Tags: Chemical sensors, Metal-oxide sensors, Temperature modulation

---

Gutierrez-Galvez, A.; Gutierrez-Osuna, R.

Pattern completion through phase coding in population neurodynamics Journal Article

In: Neural networks, vol. 16, no. 5–6, pp. 649–656, 2003.

Abstract | Links | BibTeX | Tags: Neuromorphic models

Gutierrez-Galvez, A.; Gutierrez-Osuna, R.

Coherent oscillations as a neural code in a model of the olfactory system Conference

Proceedings of the International Joint Conference on Neural Networks, IEEE 2003.

Abstract | Links | BibTeX | Tags: Neuromorphic models

Bryll, R.; Gutierrez-Osuna, R.; Quek, F.

Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets Journal Article

In: Pattern Recognition, vol. 36, no. 6, pp. 1291–1302, 2003.

Abstract | Links | BibTeX | Tags: Pattern recognition

Courte, D. E.; Rizki, M. M.; Tamburino, L. A.; Gutierrez-Osuna, R.

Evolutionary optimization of Gaussian windowing functions for data preprocessing Journal Article

In: International Journal on Artificial Intelligence Tools, vol. 12, no. 1, pp. 17–38, 2003.

Abstract | Links | BibTeX | Tags: Electronic nose, Pattern recognition

---

Gutierrez-Osuna, R.; Powar, N.

Odor mixtures and chemosensory adaptation in gas sensor arrays Journal Article

In: International Journal on Artificial Intelligence Tools, vol. 12, no. 1, pp. 1–16, 2003.

Abstract | Links | BibTeX | Tags: Metal-oxide sensors, Neuromorphic models

---

Gudmundsson, K. S.; Awwal, A. A. S.; Gutierrez-Osuna, R.

Application of phase only filters on e-nose data Conference

SPIE Conference Photonic Devices and Algorithms for Computing VI, 2003.

Abstract | Links | BibTeX | Tags: Electronic nose

2002

Nagle, H. T.; Gutierrez-Osuna, R.; Kermani, B.; Schiffman, S. S.

Environmental monitoring Book Chapter

In: Handbook of Machine Olfaction, pp. 419–444, Wiley Online Library, 2002.

Abstract | Links | BibTeX | Tags: Machine olfaction

Schiffman, S. S.; Gutierrez-Osuna, R.; Nagle, H. T.

Measuring odor intensity with e-noses and other sensor types Conference

Proceedings of the 6th International Symposium on Olfaction and Electronic Nose, 2002.

Abstract | Links | BibTeX | Tags: Electronic nose

Gutierrez-Osuna, R.; Gutierrez-Galvez, A.

Habituation in the KIII olfactory model using gas sensor arrays Conference

Proceedings of the 9th International Symposium on Olfaction and Electronic Nose, 2002.

Abstract | Links | BibTeX | Tags: Metal-oxide sensors, Neuromorphic models

---

Gutierrez-Osuna, R.; Sun, P.

A biologically-plausible computational architecture for sensor-based machine olfaction Conference

Proceedings of the ninth International Symposium on Olfaction and Electronic Nose, 2002.

Abstract | Links | BibTeX | Tags: Machine olfaction, Neuromorphic models

---

Pavesi-Llona, A.; Sundic, T.; Pardo, A.; Gutierrez-Osuna, R.; Marco, S.

A portable electronic nose based on embedded PC technology and GNU/Linux: Hardware, software and applications Journal Article

In: Sensors Journal, IEEE, vol. 2, no. 3, pp. 236–248, 2002.

Abstract | Links | BibTeX | Tags: Electronic nose, Metal-oxide sensors

Gutierrez-Osuna, R.

Pattern analysis for machine olfaction: a review Journal Article

In: Sensors Journal, IEEE, vol. 2, no. 3, pp. 189–202, 2002.

Abstract | Links | BibTeX | Tags: Machine olfaction

Gutierrez-Osuna, R.

A self-organizing model of chemotopic convergence for olfactory coding Conference

Proceedings of the 2nd Joint EMBS-BMES Conference, vol. 1, IEEE 2002.

Abstract | Links | BibTeX | Tags: Neuromorphic models

Gutierrez-Osuna, R.; Nagle, H.T.; Kermani, B. Schiffman, S.S.

Signal conditioning and preprocessing Book Chapter

In: Handbook of Machine Olfaction, pp. 105–132, Wiley Online Library, 2002.

Abstract | Links | BibTeX | Tags: Machine olfaction

Rodriguez-Mendez, M.L.; Arrieta, A.; Parra, V.; Vegas, A.; Villanueva, S.; Gutierrez-Osuna, R.; de Saja, J.A.

Sensor system for the complete organoleptic characterization (smell, taste and color) of wines Conference

Proceedings of the 9th International Symposium on Olfaction and Electronic Nose, 2002.

BibTeX | Tags: Chemical sensors

Gutierrez-Osuna, R.; Gutierrez-Galvez, A.; Powar, N.

Transient Response Analysis for Temperature Modulated Chemoresistors Conference

Proceedings of the 6th International Meeting on Chemical Sensors, 2002.

Links | BibTeX | Tags: Chemical sensors, Metal-oxide sensors, Temperature modulation

---

2001

Perera-Lluna, A.; Pardo, T.; Sundic, T.; Marco, S.; Gutierrez-Osuna, R.

IpNose: Electronic nose for distributed air quality monitoring system Conference

Proceedings of the 3rd European Congress on Odours, Metrology and Electronic Noses, 2001.

Abstract | BibTeX | Tags: Electronic nose

Perera-Lluna, A.; Gutierrez-Osuna, R.; Marco, S.

ipNOSE: A portable volatile analyzer based on embedded technology for intensive computation and time dependent signal processing Conference

Proceedings of the 8th International Symposium on Olfaction and the Electronic Nose, 2001.

Abstract | Links | BibTeX | Tags: Electronic nose, Metal-oxide sensors

---

Kakumanu, P.; Gutierrez-Osuna, R.; Esposito, A.; Bryll, R.; Goshtasby, A.; Garcia, O.N.

Speech driven facial animation Conference

Proceedings of the 2001 workshop on Perceptive user interfaces, ACM 2001.

Abstract | Links | BibTeX | Tags: Facial animation, Speech

Gutierrez-Osuna, R.; Powar, N.; Sun, P.

Chemosensory adaptation in an electronic nose Conference

Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering Conference, IEEE 2001.

Abstract | Links | BibTeX | Tags: Chemical sensors, Electronic nose, Metal-oxide sensors, Neuromorphic models, Temperature modulation

---

Gutierrez-Osuna, R.; Schiffman, S.S.; Nagle, H.T.

Correlation of sensory analysis with electronic nose data for swine odor remediation assessment Conference

Proceedings of the 3rd European Congress on Odours, Metrology and Electronic Noses, 2001.

Abstract | Links | BibTeX | Tags: Electronic nose, Olfaction

Perera-Lluna, A.; Pardo, T.; Sundi, T.; Gutierrez-Osuna, R.; Marco, S.; Nicolas, J.

IpNose: Electronic nose for remote bad odour monitoring system in landfill sites Conference

Proceedings of the 8th conference EURODEUR, 2001.

Abstract | Links | BibTeX | Tags: Electronic nose

Schiffman, S.S.; Gutierrez-Osuna, R.; Nagle, H.T.

Use of an electronic nose to evaluate odors from swine operations Conference

Proceedings of the 8th International Symposium on Olfaction and Electronic Nose, 2001.

Abstract | Links | BibTeX | Tags: Electronic nose

Gutierrez-Osuna, R.; Korah, S.; Perera, A.

Multi-frequency temperature modulation for metal-oxide gas sensors Conference

Proceedings of the 8th International Symposium on Olfaction and the Electronic Nose, 2001.

Abstract | Links | BibTeX | Tags: Chemical sensors, Metal-oxide sensors, Temperature modulation

---

Garcia, O.N.; Gutierrez-Osuna, R.

An Updated Approach to Complexity from an Agent-Centered Artificial Intelligence Perspective Book Chapter

In: Encyclopedia of Library and Information Science (ELIS), vol. 68, pp. 348-366, 2001.

Abstract | Links | BibTeX | Tags: Other

2000

Gutierrez-Osuna, R; Garcia, O N

An application of Rissanen's Complexity Measure to Optimizing Abstraction Level Conference

Proceedings of the 5th Annual Conference on Human Interaction with Complex Systems, 2000

Abstract | Links | BibTeX | Tags: Other

Gutierrez-Osuna, R

Drift reduction for metal-oxide sensor arrays using canonical correlation regression and partial least squares Conference

Proceedings of the 7th International Symposium On Olfaction & Electronic Noses, 2000

Abstract | Links | BibTeX | Tags: Metal-oxide sensors

Schiffman, S S; Wyrick, D; Gutierrez-Osuna, R; Nagle, H T

Effectiveness of an electronic nose for monitoring bacterial and fungal growth Conference

Proceedings of the 7th International Symposium On Olfaction & Electronic Noses, 2000

Abstract | Links | BibTeX | Tags: Electronic nose

1999

Gutierrez-Osuna, R; Nagle, H T; Schiffman, S S

Transient response analysis of an electronic nose using multi-exponential models Journal Article

In: Sensors and Actuators B: Chemical, vol. 61, no. 1-3, pp. 170–182, 1999

Abstract | Links | BibTeX | Tags: Chemical sensors, Electronic nose, Temperature modulation

Gutierrez-Osuna, R; Nagle, H T

A method for evaluating data-preprocessing techniques for odour classification with an array of gas sensors Journal Article

In: Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 29, no. 5, pp. 626–632, 1999

Abstract | Links | BibTeX | Tags: Chemical sensors

1998

Nagle, H T; Gutierrez-Osuna, R; Schiffman, S S

The how and why of electronic noses Book Chapter

In: IEEE Spectrum, pp. 22–34, 1998

Links | BibTeX | Tags: Electronic nose

Gutierrez-Osuna, R; Janet, J A; Luo, R C

Modeling of ultrasonic range sensors for localization of autonomous mobile robots Journal Article

In: Industrial Electronics, IEEE Transactions on, vol. 45, no. 4, pp. 654–662, 1998

Abstract | Links | BibTeX | Tags: Robotics

1997

Janet, Jason A; Gutierrez-Osuna, R; Chase, Troy A; White, M; Sutton, John L

Autonomous mobile robot global self-localization using Kohonen and region-feature neural networks Journal Article

In: Journal of Robotic Systems, vol. 14, no. 4, pp. 263–282, 1997

Abstract | Links | BibTeX | Tags: Robotics

Classen, J J; Schiffman, S S; Nagle, H T; Gutierrez-Osuna, R

Electronic nose evaluation of synthetic hog farm odor Conference

Ammonia and odour emissions from animal production facilities: International Symposium, 1997

BibTeX | Tags: Electronic nose, Olfaction

1996

Gutierrez-Osuna, R; Schudel, D S; Janet, J A; Luo, R C

Lola, the mobile robot from NC State Conference

Proceedings of the thirteenth national conference on Artificial intelligence, 1996

Links | BibTeX | Tags: Robotics

Gutierrez-Osuna, R; Luo, R C

LOLA Probabilistic Navigation for Topological Maps Journal Article

In: AI Magazine, vol. 17, no. 1, pp. 55–62, 1996

Abstract | Links | BibTeX | Tags: Robotics

1995

Janet, J A; Gutierrez-Osuna, R; Kay, M G; Luo, RC

Autonomous mobile robot self-referencing with sensor windows and neural networks Conference

Proceedings of the IEEE 21st International Conference on Industrial Electronics, Control and Instrumentation. IEEE 1995.

Abstract | Links | BibTeX | Tags: Robotics

Janet, J A; Gutierrez-Osuna, R; Chase, T A; White, M; Luo, R C

Global self-localization for autonomous mobile robots using region and feature-based neural networks Conference

Proceedings of the IEEE 21st International Conference on Industrial Electronics, Control and Instrumentation. IEEE 1995

Abstract | Links | BibTeX | Tags: Robotics

Janet, J A; Gutierrez-Osuna, R; Chase, T A; White, M; Luo, R C

Global self-localization for autonomous mobile robots using self-organizing Kohonen neural networks Conference

Proceedings of the IEEE International Conference on Intelligent Robots and Systems. IEEE 1995

Abstract | Links | BibTeX | Tags: Robotics

0000

Branan, Kimberly L.; Reyes, Gilberto O. Flores; Abel, John A.; Erraguntla, Madhav; Gutierrez-Osuna, Ricardo; Cole, Gerald L.

[No title] Proceedings Article

In: Proceedings of SPIE, 0000.

BibTeX | Tags:

Bobak, Mortazavi Anurag Das, Nikolaos Deetz

Modeling Individual Differences in Food Metabolism through Alternating Least Squares Proceedings Article

In: 0000.

Links | BibTeX | Tags: Continuous glucose monitors, Diabetes

Physical: 309 Peterson Building, Texas A&M University
Mailing: PSI Lab, 3112 TAMU, College Station, TX 77843

IPR2026-00272
Krisp Technologies EX1010 Page 184

Appendix SISMAN01

Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion

Berrak Sisman, *Student Member, IEEE*, Mingyang Zhang, *Student Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—The statistical approach to voice conversion typically consists of a feature conversion module followed by a vocoder. So far, the feature conversion studies are mainly focused on the conversion of spectrum. However, speaker identity is also characterized by prosodic features, such as fundamental frequency (F0) and energy contour among others. In this paper, we study the transformation of speaker characteristics both in terms of spectrum and prosody. We propose two novel techniques that effectively use a limited amount of source-target training data and leverage a large general speech corpus to improve the voice conversion quality. First, we study the phonetic sparse representation under the group sparsity mathematical formulation. We use phonetic posteriorgrams (PPGs) together with spectral and prosody features to form tandem feature in the phonetic dictionary. The tandem feature allow us to estimate an activation matrix that is less dependent on source speakers, thus providing a better voice conversion quality. Second, we study the use of WaveNet vocoder that can be trained on general speech corpus from multiple speakers and adapted on target speaker data to improve the vocoding quality. We benefit from the large general speech databases that are used to train the PPG generator, and the WaveNet vocoder. The experiments show that the proposed conversion framework outperforms the traditional spectrum and prosody conversion techniques in both objective and subjective evaluations.**

***Index Terms*—Voice conversion, prosody conversion, phonetic posteriorgrams (PPGs), tandem feature, wavenet vocoder.**

I. Introduction

HUMAN voice carries the information of speaker identity. The goal of voice conversion (VC) is to convert one speaker's voice to sound like that of another. Voice conversion has become an enabling technology for many new applications such as personalized speech synthesis, speaker de-identification, and voice dubbing of movies, etc.

Generally speaking, speaker identity can be characterized by 1) linguistic factors that are reflected in sentence structure and lexical choice; 2) supra-segmental factors such as the prosodic characteristics of a speech signal, and 3) segmental factors that are related to short term features, such as spectrum and formants. When the language content is fixed, the supra-segment and the segmental factors are the relevant factors to describe speaker individuality. An effective voice conversion technique is expected to convert both the supra-segment and the segmental factors.

The early studies of voice conversion were focused on spectrum mapping, that include vector quantization (VQ) [1] and fuzzy vector quantization [2]. While the statistical parametric approaches such as Gaussian mixture model [3]–[6], partial least square regression [7] and dynamic kernel partial least squares regression (DKPLS) [8] greatly improved the performance, the resulting speech quality remains much to be desired, particularly when the amount of training data are highly limited.

As a solution to the limited training data problem, non-negative matrix factorization (NMF) and sparse representation techniques were proposed [9]. Following the encouraging results in noise robust automatic speech recognition [10], the exemplar-based sparse representation techniques were studied [9], [11]–[13] to address the over-smoothing problem in voice conversion. Some more recent studies show that, by taking phonetic information into consideration [14], [15], the phonetic sparse representation technique outperforms the traditional sparse representation technique, that doesn't impose a phonetic sparse constraint on the activation matrix. In this paper, we formulate the phonetic sparse representation under the mathematical framework of group sparsity, or group sparse representation.

We note that both traditional sparse representation [9], [11], and phonetic sparse representation [14], [15] are under the assumption that the target speaker can share the activation matrix derived from source speech. However, such activation matrix is highly dependent on the source speaker. As a result, the activation matrix is affected by unwanted information such as speaker characteristics of source speakers, noise and channel distortions, that we don't want to carry over to the target voice. To alleviate this problem, we further the idea of phonetic sparse representation [14], [15] by taking into account phone-level as well as frame-level phonetic information. We construct the phonetic dictionaries by using Phonetic PosteriorGrams together with spectrum and prosody features, that we call Tandem Feature (TF). Phonetic PosteriorGrams represent the posterior probability of each phonetic class for each speech frame [16], [17] that are believed to be speaker independent [18].

The statistical approach to voice conversion typically consists of a feature conversion module that is followed by a traditional parametric vocoder. The speech synthesized by traditional

Manuscript received June 6, 2018; revised October 24, 2018; accepted March 29, 2019. Date of publication April 11, 2019; date of current version April 24, 2019. This work was supported by the Ministry of Education, Singapore AcRF Tier1 NUS Start-up Grant FY2016. Thw work of B. Sisman is supported by the SINGA Scholarship under A*STAR Graduate Academy. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ani Nenkova *(Corresponding author: Berrak Sisman.)*

B. Sisman and M. Zhang are with the National University of Singapore, Singapore 117583 (e-mail: berraksisman@u.nus.edu; zhangmy@seu.edu.cn).

H. Li is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/TASLP.2019.2910637

2329-9290 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 186

parametric vocoders lacks naturalness due to the over-simplified assumptions in the signal processing process. WaveNet vocoder [19], that directly estimates waveform samples from the input feature vectors, potentially addresses the problem. There have been studies of speaker dependent and independent WaveNet vocoders [20], [21] that generate natural sounding synthetic voices. The WaveNet approach transforms the vocoder design into a data-driven learnable process. Through the learning, the network is expected to capture the dynamics of the complex mechanics of the human speech generation process. In this paper, we study the use of both speaker independent and speaker adapted WaveNet vocoders for voice generation.

There has been reported work on GMM-based voice conversion followed by WaveNet vocoder [22]. For the first time, this paper studies the interaction between the WaveNet vocoder and the sparse representation technique. We note that the sparse representation is known for producing high similarity voice, while WaveNet vocoder offers natural voice quality. We aim to benefit from the best of the two.

The main contributions of this paper are in three areas: a) we formulate group sparse representation as the theoretical foundation of phonetic sparse representation for the integrated solution to spectrum and prosody conversion; b) we study the speaker dependent and independent characteristics to motivate the design of sparse representation framework, namely phonetic activation matrix, that is based on Tandem Feature and differentiated multi-scale prosody transfer; c) we implement a voice conversion framework with phonetic sparse representation and WaveNet vocoder that provides superior performance for limited training data.

This paper is organized as follows: In Section II, we present the related work. In Section III, we study the modeling of prosody features with wavelet transform. In Section IV, we propose group sparse representation for voice conversion. In Section V, we propose the use of Tandem Feature in a joint spectrum and prosody conversion framework. In Section VI, we discuss the use of WaveNet vocoder in the sparse representation voice conversion. We report the experiments in Section VII and conclude in Section VIII.

II. Related Work

This work is built on several previous studies on sparse representation voice conversion, and phonetic and phonological prosody. Here we briefly summarize the previous work in the context of voice conversion to set the stage for our study.

A. Exemplar-based Sparse Representation

Non-negative matrix factorization (NMF) [23] is an effective data mining tool that has been widely used, especially when reconstruction of high quality signals is required, such as in speech enhancement [24], [25], speech de-noising [26], [27], noise and speech estimation [28]. It factorizes a matrix into two matrices (a dictionary and an activation matrix) with the property that all three matrices have no negative elements. The NMF-based techniques were shown effective in voice conversion with very limited training data. We just name a few as examples, non-negative spectrogram deconvolution [29], locally linear embedding (LLE) [30], and unit selection [13]. In NMF-based approaches, a target spectrogram is constructed as a linear combination of exemplars. Therefore, over-smoothing problem can arise.

To alleviate such over-smoothing problem, one effective solution is to apply sparsity constraint to the activation matrix [14], [15], that we call exemplar-based sparse representation. The recent works on sparsity constraint includes discriminative graph-embedded NMF approach [12], phonetic sparse representation for spectrum conversion [15], and conversion of timbre and prosody [31], [32].

In exemplar-based sparse representation, a pair of dictionaries, denoted as $\mathbf{A}$ and $\mathbf{B}$, consists of aligned exemplars from source and target. $[\mathbf{A}; \mathbf{B}]$ is called the coupled dictionary. At run-time, the spectrogram of a source utterance $\mathbf{X}$ can be represented as

$$\mathbf{X} \approx \mathbf{A}\hat{\mathbf{H}} \tag{1}$$

Due to the nonnegative nature of spectrogram, NMF technique is employed to estimate the source activation matrix $\hat{\mathbf{H}}$, which is constrained to be sparse. Mathematically, we estimate $\hat{\mathbf{H}}$ by minimizing an objective function,

$$\hat{\mathbf{H}} = \underset{\mathbf{H}\geq 0}{\operatorname{argmin}}\, d\left(\mathbf{X}, \mathbf{A}\mathbf{H}\right) + \lambda||\mathbf{H}|| \tag{2}$$

where λ is the sparsity penalty factor. To estimate activation matrix $\hat{\mathbf{H}}$, a generalised Kullback-Leibler (KL) divergence is used. It is assumed that source and target dictionaries $\mathbf{A}$ and $\mathbf{B}$ can share the same source activation matrix $\hat{\mathbf{H}}$ from Eq.(2). Therefore, the converted spectrogram can be written as

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\mathbf{H}}. \tag{3}$$

We note that the source activation matrix $\hat{\mathbf{H}}$ is estimated from the source speech, therefore, is biased very much towards the source signals. In this paper, we further the idea of exemplar-based sparse representation by studying how to reduce the bias, and the way to improve the sparse representation through phonetic group sparsity.

B. Spectrum Conversion with Phonetic Sparse Representation

It is known that a compact and effective dictionary has an important role in achieving a good voice conversion quality [33]. In addition, multiple dictionaries with a dictionary selection strategy [14] also help to improve the conversion performance of sparse representation. Our recent work in phonetic sparse representation [15] voice conversion is another example that employs a phonetic dictionary with phonetically labeled exemplars, or phonetic sub-dictionaries. It has been shown that multiple sub-dictionaries outperform single dictionary in exemplar-based sparse representation voice conversion [14], [15].

The studies in [14] and[15] haven't formulated the problem of multiple sub-dictionaries under a theoretical framework. They assume that source and target speakers can share the same activation matrix, without looking into how to estimate such a shareable activation matrix. Moreover, they focus on spectrum conversion without studying the prosody transfer between the

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 187

speakers. To gain insights into the previous work, we introduce group sparse representation in this paper to describe both exemplar-based sparse representation, that is referred to as traditional sparse representation, and phonetic sparse representation under an unified mathematical framework. We further study the use of Tandem Feature for phonetic dictionary to obtain speaker invariant activation matrix for spectrum and prosody conversion.

C. Phonetic and Phonological Prosody

The prosody of speech manifests itself prominently at two different levels [34], [35]: 1) the phonetically-caused aspects of prosody, referred to phonetic prosody, that arise from purely physical phonetic factors. Phonetic prosody are associated with phones and not reflected in the mental lexicon; and 2) the phonologically-maintained aspects of prosody, that is related to the mental lexicon, phrase, and sentence, referred to as phonological prosody. It is observed that prosody pattern is influenced by short term as well as long term dependencies [36] as it is hierarchical in nature [37], [38].

Prosody conveys linguistic, para-linguistic and various types of non-linguistic information, such as speaker identity, intention, attitude and mood [39], [40]. It is inherently supra-segmental [41], [42] due to the fact that the variations of prosody cannot be derived solely from the segmental sequence [37]. Speech prosody is affected by long-term dependencies at different levels such as word, phrase and utterance. At the same time, it is also affected by segmental differences [37]. For example, voiceless segments lack explicit F0 values and high vowels generally have higher F0 than low vowels. Unfortunately, there have been a few studies on how voice conversion can benefit from the findings in prosodic analysis.

The continuous wavelet transform (CWT) was used for the analysis and modeling of F0 within an hidden Markov model (HMM) framework [43],[44]. The CWT decomposition of F0 and energy contours was also used in voice conversion framework such as DKPLS [36], exemplar-based prosody conversion [31], [32], [45], and emotional voice conversion [46].

Motivated by the previous work, we would like to further the study of CWT decomposition for prosody modeling under the sparse representation voice conversion framework. We also believe that the phonetic sparse representation is a logical choice of platform for phonetic-phonological prosody conversion.

III. Prosody Modeling with CWT

Prosody is hierarchical in nature [32], [37], [38] and it can be affected by both short term as well as long term dependencies [36]. Fundamental frequency (F0) is a crucial prosodic feature in speech, hence previous studies of prosody conversion mainly are focused on transformation of F0. We will study a technique to convert both F0 and energy contours next.

One of the challenges is that prosody is highly dynamic. The limited amount of training utterances are generally not enough to characterize the prosody patterns of a speaker. The traditional technique for F0 conversion usually transforms the mean and variance of source speaker's F0 to those of the target [11]. One of the disadvantages of this technique is that it is based on a frame-level operation. We note that human manipulates F0 in a segmental manner, including phone, syllable, word, phrase or sentence level. GMM-based mapping [47], piecewise linear mapping [48] represent some promising extensions of this traditional approach [11].

Recently, it has been shown that CWT can effectively model F0 in different temporal scales that improves the speech synthesis performance [42]–[44]. Inspired by this, CWT is studied in F0 modeling within sparse representation [31], DKPLS for voice conversion [36], emotion conversion with arbitrary F0 scales [49] and adaptive scales for emotional voice conversion [50]. It was also studied that CWT decompositions of F0 and energy contours provide a way to manipulate the prosody of utterances [46].

The continuous wavelet transform of an input signal f_0 can be written as

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x) \psi \left(\frac{x - t}{\tau} \right) dx, \tag{4}$$

where ψ is the Mexican hat mother wavelet. If we fix the analysis at 10 discrete scales, f_0 can be represented as follows [31]

$$W_i(f_0)(t) = W_i(f_0)(2^{i+1}\tau_0, t)(i + 2.5)^{-5/2}, \tag{5}$$

where $i = 1, \ldots, 10$ and $\tau_0 = 5$ ms. These timing scales were originally proposed in [44] and in hierarchical prosody model [43], that were further used in voice conversion [31], [32], [45], [46]. The use of these particular scales is motivated by the attempt to relate the scales to different levels of linguistic structure. In this representation, lower scales capture short-term variations and higher scales capture the long-term variations associated with the utterance. The reconstruction formula is given as follows:

$$f_0(t) = \sum_{i=1}^{10} W_i(f_0)(t)(i + 2.5)^{-5/2}. \tag{6}$$

In this study, we adopt CWT to decompose both F0 and energy contour into 10 temporal scales, that can be used to model different prosodic levels ranging from micro-prosody to sentence levels. As CWT is sensitive to discontinuties in the prosody features, the following pre-processing steps are needed: 1) transformation of F0 and energy values from linear to logarithmic scale, 2) smoothing F0 and energy contour by using 3-point mean filter 3) linear interpolation over unvoiced regions, 4) normalizing the resulting F0 and energy contour to zero mean and unit variance.

Prosody typically reflects a combination of features from both the speaker and the linguistic content of the utterance. The examples of speaker features include the dialectal background, the lexical choice in the expression, the speaking rate, the emotional state of the speaker etc; while the examples of linguistic content, referred to as utterance features, include the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus etc. In general, we consider speaker features are speaker dependent, and utterance features are speaker independent because they are manifested in the same way by all speakers. It is generally believed that the elements of prosody can be identified as either speaker dependent and independent. In voice conversion [51], [52], we would like

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 188

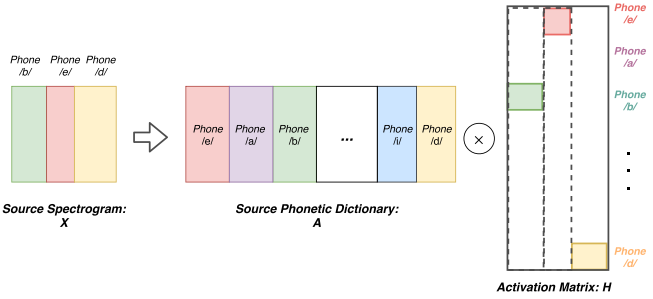


Fig. 1. The group sparse representation of a source spectrogram $\mathbf{X}$ for the word bed that has 3 phonetic groups highlighted in 3 colors. $\mathbf{X}$ is factorized with group sparsity into the product of the source phonetic dictionary $\mathbf{A}$ and the activation matrix $\mathbf{H}$. The colored rows of $\mathbf{H}$ represent nonzero sections of the activation matrix.

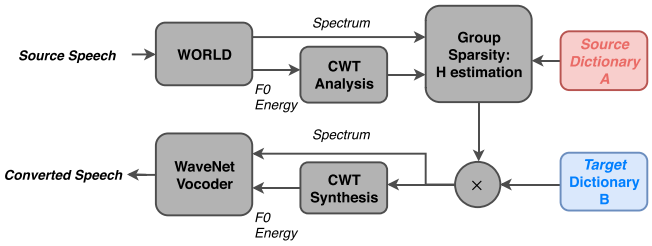


Fig. 2. The activation matrix, H, estimation and run-time conversion of the sparse representation framework under the group sparsity constraint. Exemplars in source and target dictionaries are with the phonetic labels.

to carry over the speaker independent prosody from the source to the target, but to replace the speaker dependent prosody of source speaker with that of target speaker.

We believe that a better way to convert phonetic-phonological prosody (F0, energy) is to associate the phonetic prosody with the phonetic dictionary. The 10-scale decomposition of F0 and energy contour provides a way to model the short term as well as long term dependencies of prosody. With wavelets of different time spans, some scales are more language dependent, and others are more speaker dependent. For example, studies have shown that scales 4-8 in a 10-scale decomposition represent the phonetic prosody that are more speaker dependent than other scales [32], [53], [54].

IV. Group Sparse Representation for Voice Conversion

A. Group Structure and Group Sparsity

Many real world data are inherently organized in groups with group prior information. For example, phoneme in a speech signal, speaker identity in a telephone recording, and topic in document archive form natural groupings. NMF is an effective data mining tool that has been widely used in many pattern recognition tasks. The idea of group sparsity is to incorporate group prior information into a mixed norm regularization process for non-negative matrix factorization, that constrains the activation matrix to the relevant data group in the optimization process. NMF with group sparsity has been used to improve speaker identity discrimination [55] and speaker identification [56] where speaker identity is used to induce group sparsity. In this paper, we would like to show that, given the phonetic class as the group prior information, sparse representation actually leads to phonetically aware sparse representation [15].

Let's now describe the phonetic group structure that we consider in Fig. 1. Suppose that we would like to convert a word bed with 3 phonemes from the source to the target speaker. The spectrum features are characterized by 3 prevalent phonetic groups $/b/, /e/$ and $/d/$. The motivation of our work is that the spectrum features belonging to a specific phoneme may share the same sparsity pattern, thus the activation matrix should be constrained to be sparse similarly within the the relevant phonetic group. We propose an extension of NMF that incorporates phonetic identity as the prior information to show the theoretical intuition of phonetic sparse representation.

B. Formulation of Group Sparsity for Voice Conversion

The traditional sparse representation framework for voice conversion constrains the activation matrix to be sparse without taking into account the group membership information. By defining phoneme class as the group membership, we formulate the problem of sparse representation. Next we show how the proposed group sparse representation works in voice conversion.

During training, we first use a vocoder, for example, WORLD [57], to analyze the spectrum and prosody features. Then, we decompose F0 and energy into 10 temporal scales by using Continuous Wavelet Transform. At the same time, we use an Automatic Speech Recognizer (ASR) to find the phoneme labels and boundaries for each pair of source-target parallel utterances. We construct a coupled dictionary $[\mathbf{A}; \mathbf{B}]$, of which the spectrum and prosody exemplars are phonetically aligned, referred to as phonetic dictionary. The phonetic dictionaries $\mathbf{A}$ and $\mathbf{B}$ are divided into sections that are also called $\mathbf{A}_i$ and $\mathbf{B}_i$ sub-dictionaries, where $i = 1, \ldots, N$ with N being the number of phone entries in the phonetic dictionary.

The run-time conversion of group sparse representation is given in Fig. 2. We note that ASR is only used during training when we construct the phonetic dictionary. The group sparsity criterion for sparse representation [58] can be written as follows:

$$\hat{\mathbf{H}} = \underset{\mathbf{H} \geq 0}{\operatorname{argmin}}\, d\left(\mathbf{X}, \sum_{i=1}^{N} \mathbf{A}_i \mathbf{H}_i\right) + \lambda \sum_{i=1}^{N} \sqrt{p_i} ||\mathbf{H}_i|| \qquad (7)$$

where A_i is the subdictionary for phone i and p_i is the length of $\mathbf{H}_i$. This criterion exploits the non-differentiability of $||\mathbf{H}_i||$ at $\mathbf{H}_i = 0$; setting some groups of coefficients to exactly 0. In this way, the group sparsity leads to a sparse set of groups. If the group sparsity decides to activate a group in the model, all coefficients in the group carry non-zero values.

To understand the difference between the group sparse representation and the traditional sparse representation, let's compare Eq. (2) and Eq. (7). We note that Eq. (2) decomposes an source spectrogram into $\mathbf{A} \times \hat{\mathbf{H}}$ where $\hat{\mathbf{H}}$ is sparse over the speech exemplars in the dictionary $\mathbf{A}$. Eq.(7) decomposes an source

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 189

TABLE I
THE COMPARISON OF SPARSE REPRESENTATION APPROACHES IN TERMS OF THE USAGE OF ASR AND PHONETIC DICTIONARY. WE CONSIDER GROUP SPARSE REPRESENTATION IS A CONSTRAINED VERSION OF TRADITIONAL SPARSE REPRESENTATION, AND PHONETIC SPARSE REPRESENTATION ASSIGNS THE PHONETIC CLASS AS THE GROUP IN GROUP SPARSE REPRESENTATION

Framework	ASR	Phonetic Dict.	H Estimation
Group Sparse Representation	Training only	Yes	Eq. (7)
Phonetic Sparse Representation [15]	Training & Testing	Yes	Eq. (8)
Sparse Representation [11]	N. A.	No	Eq. (2)

spectrogram into $\sum \mathbf{A}_i \times \hat{\mathbf{H}}_i$ where $\hat{\mathbf{H}}_i$ is sparse over the phonetic group of speech exemplars. In the latter case, only a few phonetic groups in the phonetic dictionary $\mathbf{A}_i$ are activated in the factorization. During training, the group sparse representation uses an ASR to derive a phonetic dictionary where each exemplar is phonetically labeled, while the traditional sparse representation puts all exemplars into a single dictionary indiscriminatively.

With the group sparsity regularization, only the phonetic sub-dictionary that is relevant to the input features is likely to be activated at run-time. As the activated sub-dictionary only contains the phonetic exemplars that are similar to the input features, we can expect the group sparse representation offers a better activation matrix, thus better converted voice than the traditional sparse representation [11].

We consider that the proposed formulation of group sparse representation provides the theoretical intuition to the phonetic sparse representation that was reported in [15]. To summarize, we compare the family of sparse representation frameworks in Table I. Group sparse representation and phonetic sparse presentation both use ASR to derive phonetic dictionary during training. At run-time, phonetic sparse representation estimates activation matrix using phonetic labels of source features from ASR [15], while group sparse representation doesn't use phonetic labels from ASR. In phonetic sparse representation, the objection function for estimating the activation matrix $\hat{\mathbf{H}}_k$ for phone $i = k$ is formulated [15] by selecting the corresponding phonetic sub-dictionary $\mathbf{A}_k$,

$$\hat{\mathbf{H}}_k = \underset{\mathbf{H}_k \geq 0}{\operatorname{argmin}}\, d\left(\mathbf{X}_k, \mathbf{A}_k \mathbf{H}_k\right) + \lambda ||\mathbf{H}_k|| \tag{8}$$

which is similar to Eq. (2) except that $\hat{\mathbf{H}}_k$ is now phone dependent [15] and the overall $\hat{\mathbf{H}}$ can be written as $[\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2, \ldots, \hat{\mathbf{H}}_k, \ldots, \hat{\mathbf{H}}_M]$, where M is the number of phones in a given utterance. The phonetic labels come from either an ASR or the transcript of the source speech at run-time. The estimation of activation matrix for group sparse representation, or Eq. (7), is different from Eq. (8) in that it doesn't explicitly require the phonetic labels of source speech. It optimizes Eq. (7) numerically to find $\hat{\mathbf{H}}_k$. In the ideal case, we end up with $\hat{\mathbf{H}}_k$ that corresponds to the phone i section of in the source speech features through Eq. (7). In this way, we can expect that Eq. (7) and Eq. (8) lead to almost the same matrix decomposition. As the traditional sparse representation does not take into account the phonetic information, it doesn't involve ASR either during training or testing.

V. JOINT SPECTRUM AND PROSODY CONVERSION WITH TANDEM FEATURES

In this section, we propose an integrated voice conversion framework that jointly transforms both spectrum and prosody under the formulation of group sparse representation. As shown in Table I, phonetic sparse representation is the most constrained version of sparse representation among the three variants, and offers the best performance. Next we study the use of speaker invariant features under the phonetic sparse representation framework. Without loss of generality, the same feature scheme will be applicable to the implementation of group sparse representation and traditional sparse representation frameworks as well.

A. Tandem Feature

The sparse representation marks a success in voice conversion when only a very limited amount of training data is available [11], [15]. In the sparse representation framework, the activation matrix is shared between the source and target speakers. Ideally, the activation matrix should capture the speaker invariant phonetic information [59]. We now argue that such sharing might not be grounded well either in theory or in practice.

Suppose that we have a coupled dictionary and a pair of parallel utterances. For ease of discussion, we use the traditional sparse representation framework here as an example to discuss the sharing of activation matrix. The same argument is valid for both group sparse representation and phonetic sparse representation.

Using Eq. (2), we can derive an activation matrix for the source utterance, that we call source activation matrix $\hat{\mathbf{H}}$; If we replace $\mathbf{X}$ with $\mathbf{Y}$, and $\mathbf{A}$ with $\mathbf{B}$ in Eq.(1), we can derive an activation matrix for the target utterance, that we call target activation matrix $\hat{\mathbf{H}}^t$, as follows:

$$\hat{\mathbf{H}}^t = \underset{\mathbf{H}^t \geq 0}{\operatorname{argmin}}\, d\left(\mathbf{Y}, \mathbf{B}\mathbf{H}^t\right) + \lambda ||\mathbf{H}^t|| \tag{9}$$

In [9], [11] and [15], it is assumed that the source activation matrix $\hat{\mathbf{H}}$ and target activation matrix $\hat{\mathbf{H}}^t$ are similar enough, therefore the source activation matrix is used for the target speaker. Such sharing is based on the assumption that the source activation matrix $\hat{\mathbf{H}}$ mostly captures the phonetic content. However, the source activation matrix in reality carries speaker characteristics from the input source utterance, that we don't want to carry over to the target utterance. The same problem exists in to Eq. (7) and Eq. (8).

To alleviate this problem, we propose to use Tandem Feature that consists of Phonetic PosteriorGrams (PPGs), spectrum and prosody features to represent the exemplars, thus, the dictionaries. Phonetic PosteriorGrams represents the posterior

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 190

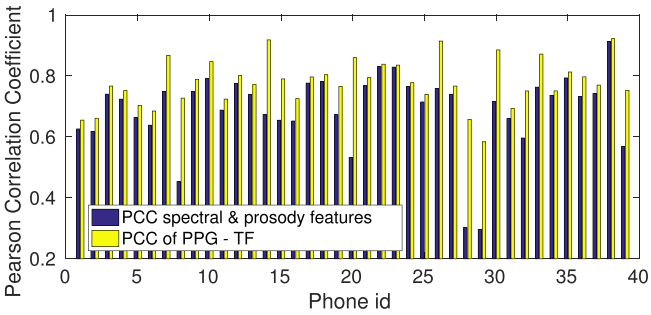


Fig. 3. Pearson Correlation Coefficient of spectrum &amp; prosody features vs. Tandem Feature. As Tandem Feature shows consistently high PCCs, thus is more speaker invariant than the spectrum and prosody features.

probability of each phonetic class for each speech frame [16]–[18], [60]. We propose to convert all speaker's voice into PPGs, and construct dictionaries that consist of source spectrum and prosody features as well as PPGs. As the dictionaries now consist of PPGs, that are considered speaker-independent phonetic features, we believe that the resulting activation matrix depends less on the source speaker, thus represents a better transformation function for voice conversion.

In Fig. 3, we compare the two types of features by examining their inter-speaker Pearson's Correlation Coefficients (PCCs): a) spectrum and prosody features; b) Tandem Feature: PPGs, spectral features, and prosodic features. We report the phone-dependent PCCs across all the speaker pairs on Voice Conversion Challenge (VCC) 2016 database. As a pair of speakers share the same linguistic content, the feature of a higher PCC suggests that it is more speaker invariant.

B. Joint Spectrum and Prosody Conversion

Recent studies show that phonetic sparse representation [14], [15], [32] achieves better voice conversion quality by using phonetic transcription of the speech both during training and at run-time. We now extend the idea of Tandem Feature to the phonetic sparse representation framework to perform joint spectrum and prosody conversion. We call the proposed framework as PSR-TF. When the idea of Tandem Feature is extended to group sparse representation, we call the framework as GSR-TF. For simplicity, we only discuss PSR-TF in detail.

We believe that a better way to convert the spectrum and prosody (F0, energy) features is to use phonetic dictionary [32]. Furthermore, phonetic dictionary with Tandem Feature allows for a better estimation of activation matrix that is less dependent on source speaker. We consider that the phonetic sparse representation framework will benefit from better activation matrix with Tandem Feature.

Fig. 4 shows the training and run-time phases of the proposed joint spectrum and prosody conversion framework. During training, we use WORLD to extract spectrum and fundamental frequency. To obtain the energy contour of the speech signal, we first find the energy of each speech frame. With the energy of every speech frame, we obtain the energy contour of the speech signal. We then use CWT to obtain 10-scale decomposition of F0 and energy contour. In addition, we use a DNN-HMM based Automatic Speech Recognizer (ASR) to find the phone labels, boundaries and PPGs for each utterance. Instead of using a single coupled dictionary, we now have multiple coupled dictionaries $[\mathbf{A}_i; \mathbf{B}_i]$, one for each phone i, where $i = 1, \ldots, N$, $\mathbf{A}_i$ is the source phonetic dictionary, and $\mathbf{B}_i$ the target phonetic dictionary. The source phonetic dictionary here consists of both PPGs, source spectrum features as well as CWT-based prosody features, called $sourceTF - dictionary$ while the target phonetic dictionary $\mathbf{B}_i$ only includes spectrum features and prosody features, called $targetdictionary$.

At run-time conversion, we use CWT to obtain the 10-scale representations of F0 and energy contour, and the spectrum features and its corresponding PPGs denoted as $\mathbf{P}_i$, for each phone of the source speaker with the same ASR system that was used in the training phase. For phone $i = k$, a source speech is represented by a matrix, denoted as $\mathbf{X}_k$, that consists of spectrogram, and 10-scale representations of F0 and energy contour, can be written as

$$\mathbf{X}_k \approx \mathbf{A}_k \hat{\mathbf{H}}_k \tag{10}$$

The objective function for estimating the activation matrix $\mathbf{H}_k$ can be formulated as follows.

$$\hat{\mathbf{H}}_k = \underset{\mathbf{H}_k \geq 0}{\operatorname{argmin}}\, d\left([\mathbf{X}_k; \mathbf{P}_k], \mathbf{A}_k \mathbf{H}_k\right) + \lambda ||\mathbf{H}_k|| \tag{11}$$

Finally, the activation matrix is applied to the target phonetic dictionary to perform conversion. The converted spectrum and prosody features for phone k can be generated as

$$\hat{\mathbf{Y}}_k = \mathbf{B}_k \hat{\mathbf{H}}_k. \tag{12}$$

We note that the 10-scale decomposition of F0 and energy contour represents the short term as well as long term dependencies of prosody. With wavelets of different time spans, some scales are more language dependent, and others are more speaker dependent. We transform scales 3-8 and carry over scales 1, 2, 9 and 10 from source to target speakers because we consider they are more speaker independent [32].

In phonetic sparse representation, we use the phonetic labels of the run-time input speech from either an ASR or the given scripts. Next we discuss three ways to take advantage of such phonetic labels, the phone-dependent residual compensation, context information, and back-off scheme.

C. Phone-dependent Residual Compensation

It is well-known that residual compensation (RC) reduces spectral distortion arising from reconstruction of speech features [11]. If we have the phonetic label of the speech at run-time, either from ASR or given the scripts, we can introduce phone dependent residual estimation and compensation to phonetic sparse representation for the first time, which is different from [11] and [14] where residual compensation is only introduced to spectrum conversion. Here we use dynamic partial least square (DPLS) regression [8] for spectrum and prosody features, that we call *PSR-TF-RC*. During training, we model the residual mapping from source to target speaker for both spectrum and

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 191

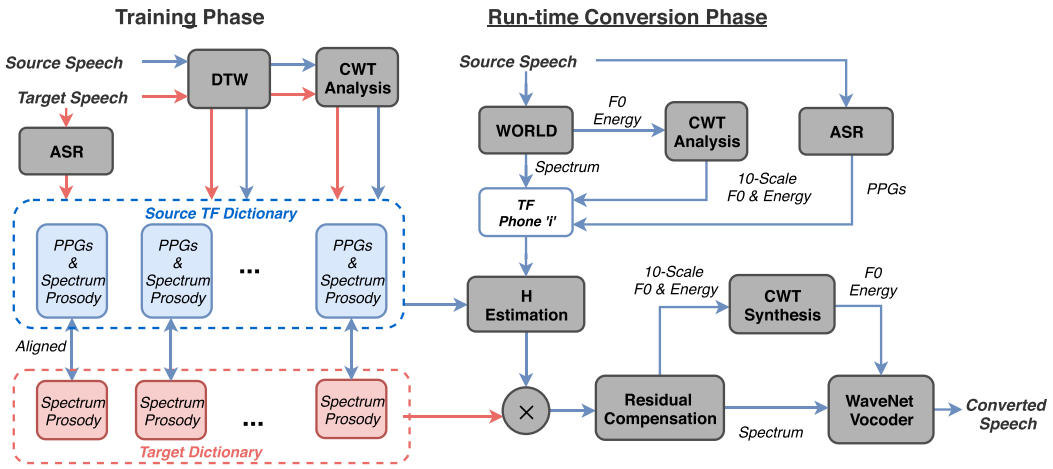


Fig. 4. The training and run-time conversion phases of the proposed joint spectrum and prosody conversion framework with Tandem Feature.

prosody features. At run-time, we use the estimated residual mapping function to compensate the conversion error.

For phone $i = k$, the modeling error between the observed source features $\mathbf{X}_k$ and the reconstructed source features $\hat{\mathbf{X}}_k$ is called source residual, and the modeling error between the observed target features $\mathbf{Y}_k$ and the converted target features $\hat{\mathbf{Y}}_k$ is called target residual. The process of finding source and target residuals can be formulated as:

$$\mathbf{R}_{S,k} = \log(\mathbf{X}_k) - \log(\hat{\mathbf{X}}_k) \tag{13}$$

$$\mathbf{R}_{T,k} = \log(\mathbf{Y}_k) - \log(\hat{\mathbf{Y}}_k) \tag{14}$$

where $\mathbf{R}_{S,k}$ and $\mathbf{R}_{T,k}$ are source and target residuals, respectively. As given below, we can train a DPLS to estimate a residual mapping function denoted as F_k, from each phone of source speaker to target speaker by using the paired source-target residuals.

$$\mathbf{R}_{T,k} = F_k\left(\mathbf{R}_{S,k}\right) \tag{15}$$

D. Contextual Information

So far, each frame is converted independently, in other words, contextual information is not taken into account. This may lead to sharp changes across frames. By considering contextual information, one can expect a better conversion performance [11], [15], [32]. To achieve a more reliable activation matrix estimation, we implement exemplars that span multiple consecutive frames in TF phonetic dictionary.

In unit selection approach to speech synthesis [13], [61], we favor speech units that share similar phonetic context as the intended context by using bi-phones or tri-phone context. In a similar way, if we know the phonetic labels of the input speech at run-time, we can use biphones together with monophones in TF phonetic dictionary to achieve a smoother phone transition. As far as prosody conversion is concerned, the studies on prosody and phonology [34], [35] reveal that the phonetic prosody is highly influenced by the transition between phonemes. We consider that bi-phone phonetic dictionary with Tandem Feature captures the prosodic patterns arising from specific phonetic transitions.

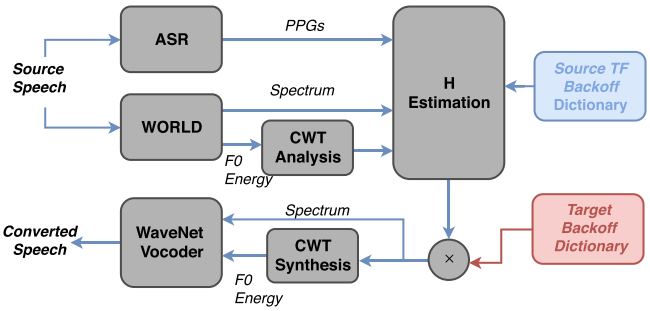


Fig. 5. The run-time workflow of the back-off scheme, that is Sparse Representation with Tandem Feature (SR-TF) for the joint spectrum and prosody conversion.

E. Back-off Scheme

As we have limited training data, it is not guaranteed that there are always enough exemplars for every phone in the phonetic dictionary that is needed at run-time. In such cases, we propose a back-off scheme that uses exemplars from all phones.

During training phase, we construct a back-off dictionary that includes all exemplars from source and target speakers. We incorporate PPGs, spectrum features and CWT-based prosodic features to form Tandem Feature in the source dictionary, while we only include spectrum and prosody features in the target dictionary. In practice, such a back-off dictionary is the union of all phonetic sub-dictionaries. The run-time conversion phase of the proposed backoff scheme is given in Fig. 5.

The proposed back-off scheme is an extension to the traditional sparse representation (SR) framework [31] by incorporating PPG features. Therefore, we call it *SR-TF* for spectrum and prosody conversion. We note that SR can perform prosody conversion under very limited training data. As the proposed back-off scheme incorporates PPGs as frame-level phonetic information, we expect that it outperforms the traditional SR

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 192

framework. Last but not least, with the proposed back-off dictionary, SR-TF can also be used alone as a system for voice conversion.

VI. WaveNet as Vocoder

Many well-known voice conversion frameworks [1]–[4], [62]–[65] including sparse representation [9], [14], [15], typically use a statistical parametric vocoder. Traditional parametric vocoder is generally designed to simulate the complex mechanics of the human speech generation process under certain simple assumptions, for example, the interaction between F0 and formant structure is ignored, the phase information is discarded [66], the assumption of stationary process in the short-time window, a time-invariant linear filter. Therefore, the speech that is generated from the traditional vocoders, lacks naturalness in general. This problem is even more serious in voice conversion as the feature conversion changes both F0 and the formant structure of speech among others. We believe that a good vocoder can help reconstruct the speech by harmonizing various changes.

A. WaveNet

WaveNet [19] is a deep neural network that can generate high quality raw audio waveforms. The joint probability of a waveform $x = x_1, x_2, \ldots, x_N$ can be factorized as a product of conditional probabilities.

$$p(x) = \prod_{n=1}^{N} p(x_n | x_1, x_2, \ldots, x_{n-1}) \tag{16}$$

WaveNet approximates the conditional probability by canceling the effect of past samples. The structure of WaveNet consists of many residual blocks. Each of these residual blocks consist of 2×1 dilated causal convolutions, a gated activation function and 1×1 convolutions. Using additional auxiliary features h, WaveNet can also model conditional distribution $p(x|h)$. Eq. (16) can then be written as follows:

$$p(x|h) = \prod_{t=1}^{T} p(x_t | x_1, x_2, \ldots, x_{t-1}, h) \tag{17}$$

We note that we can control the characteristics of generated samples by conditioning the network with auxiliary features. In the original WaveNet [19], linguistic features and/or speaker codes are conditioned to generate speech samples according to the given text input while keeping specific speaker characteristics.

B. WaveNet Vocoder

We note that WaveNet vocoder [21] achieves remarkable sound quality improvement over the traditional vocoders. Not only can it learn the relationship between input features and output waveforms, but also it learns the interaction among the input features. Moreover, it has been sucessfully used for speech synthesis [67] and in GMM-based voice conversion [22] to improve the naturalness of the synthetic voice. Recently, a speaker independent WaveNet vocoder [21] is studied by utilizing the acoustic features such as F0, aperiodicity, and spectrum as the additional inputs of WaveNet. In doing so, WaveNet learns a sample-by-sample correspondence between the time-domain waveform and the corresponding acoustic features. To leverage from human speech of a larger speaker population, we apply the recently proposed speaker independent WaveNet vocoder [21] to generate the speech waveforms. We also propose to adapt the speaker independent WaveNet vocoder towards the target speaker to make full use of the training samples from the target speaker.

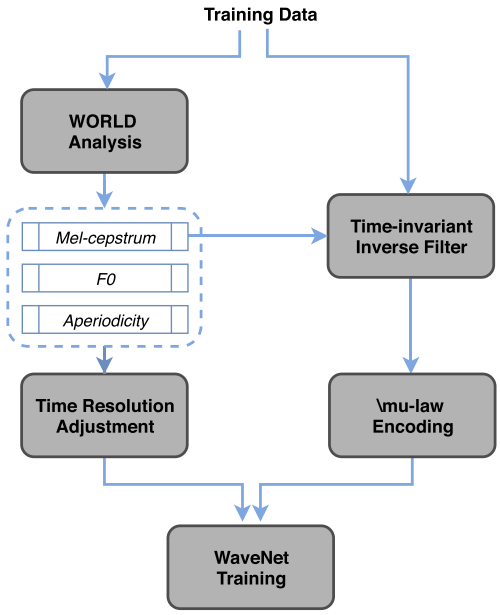


Fig. 6. The training phase of speaker independent WaveNet vocoder [21].

C. Adaptation of WaveNet Vocoder

In this paper, we study the case where we only have a limited number of speech samples from the target speaker, which is a challenging task in voice conversion. While WaveNet vocoder is a deep neural network that typically requires a large amount of training data, we believe that it can benefit from the limited speech sample through adaptation process. We first train the speaker independent WaveNet vocoder with CMU Arctic and VCC 2016 datasets, in a similar way that is described in [21] and given in Fig. 6. During training, we exclude the speech data from any of the source or target speakers to make sure we have a speaker-independent WaveNet Vocoder to start with. Then, we propose to adapt the WaveNet vocoder for a target speaker by using the same target utterances that are used in the training of sparse representation. The adaptation brings the speaker independent WaveNet vocoder close to the target speaker.

At run-time conversion, we use the proposed adapted WaveNet vocoder to generate the target speech waveforms. It is not straightforward to report the performance of the WaveNet vocoder using objective metrics such as Mel-cepstral distortion [68]. We report the listening test of the WaveNet vocoder under various conditions in Section VII.

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 193

TABLE II
A COMPARISON OF SPECTRAL DISTORTION (MCD) [68] BETWEEN THE BASELINE EXEMPLAR-BASED SPARSE REPRESENTATION [9] AND A VARIANT OF GROUP SPARSE REPRESENTATION. WE COMPARE THE USE OF MONOPHONE (M) AND THE COMBINATION OF MONOPHONE AND BIPHONE (M+B) IN THE PHONETIC DICTIONARY THAT ARE DERIVED FROM 25 AND 35 PARALLEL UTTERANCES FOR EACH SPEAKER PAIR

Spectrum Conversion	Features	# Frames	Phonetic Dict.	Script: Training	Script: Testing	MCD-35	MCD-25
PSR-TF	Tandem Feature	1	M	yes	yes	5.09	5.21
	Tandem Feature	3	M+B	yes	yes	4.98	5.15
	Tandem Feature	3	M+B	yes	no	5.06	5.19
	Tandem Feature	3	M+B	no	no	5.12	5.25
PSR [15]	spectrum+prosody	1	M	yes	yes	5.18	5.31
	spectrum+prosody	3	M+B	yes	yes	5.09	5.23
	spectrum+prosody	3	M+B	yes	no	5.16	5.29
	spectrum+prosody	3	M+B	no	no	5.22	5.34
GSR-TF	Tandem Feature	3	M	yes	no	5.31	5.43
SR-TF	Tandem Feature	3	-	-	-	5.58	5.72
SR [11]	spectrum+prosody	3	-	-	-	5.78	5.92

VII. EXPERIMENTS

We report the voice conversion experiments on VCC 2016 database [69]. The VCC 2016 dataset is recorded from 5 female and 5 male US English speakers. The spectrum compression factor is set to 0.4 in all experiments [11]. In addition, we use a DNN-HMM based ASR [70] to obtain phone labels, phone boundaries and PPGs. The ASR is reported with 18.0% WER on WSJ Eval92 database. While we conduct joint spectrum and prosody (F0 and energy) conversion, we report objective evaluation on spectrum and prosody conversion separately. We conduct experiments with 25 and 35 pairs of parallel training utterances.

Mel-cepstral distortion (MCD) [68] between the converted and reference Mel-cepstra was employed as an objective evaluation measure for spectrum conversion.

We use Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) as the evaluation metrics for prosody conversion. The PCC measures the linear dependence of two signals, that can be defined as:

$$p(S,T) = \frac{cov(S,T)}{\sigma_S \sigma_T} \tag{18}$$

where σ_S and σ_T are the standard deviations of signals S and T, respectively. We report the PCCs for F0 and energy contour between the converted and the reference target speech.

A. Evaluation of WaveNet Vocoder

We use MCD to assess the spectrum conversion, and PCC to assess the prosody conversion. Both are calculated before the vocoding. As a result, it is not straightforward to compare the vocoding effect between WaveNet vocoder and traditional vocoders using objective measures. Therefore, we only evaluate the performance of WaveNet through listening test in the subjective evaluation.

B. Reference Methods and Setups

To validate our proposals, we implement the state-of-the-art methods for joint spectrum and prosody conversion. In addition, we also implement a family of sparse representation schemes. The setups are summarized as follows:

- Reference baseline: linear F0 conversion.
- SR: Exemplar-based sparse representation [31].
- PSR: Phonetic sparse representation [15], [32].
- GSR: Group sparse representation, as described in Section IV *(proposed)*.
- PSR-TF: Phonetic sparse representation with Tandem Feature, as described in Section V *(proposed)*.
- PSR-TF with Speaker Independent WaveNet Vocoder *(proposed)*.
- PSR-TF with Speaker Adapted WaveNet Vocoder *(proposed)*.
- PSR-TF with Residual Compensation, denoted as *PSR-TF-RC*. *(proposed)*.

We conduct both objective and subjective evaluations to assess their performance in a comparative study.

The GMM-based approaches [3], [6], such as ML-GMM, represent a group of successful voice conversion frameworks. The dynamic kernel partial least squares (DKPLS) regression method has been shown to be effective for implementing a non-linear conversion function [8]. The studies [11] have shown that the sparse representation approach in general outperforms ML-GMM [3] and DKPLS [8]. The recent work on phonetic sparse representation [14], [15] shows that it further improves the sparse representation framework by incorporating phonetic information. As the phonetic sparse representation (PSR) provides a competitive performance among the state-of-the-art techniques for the limited training data problem, we use PSR as the baseline in the experiments.

C. Objective Evaluation

1) Spectrum Conversion Performance: Table II reports the MCD values for a number of settings in a comparative study. Firstly, we observed that all PSR-TF settings consistently outperform those of other sparse representation frameworks both with 25 and 35 training utterances. Secondly, when scripts are available for training and/or test utterances, we perform forced alignment to obtain better phone segmentation, therefore, we get lower MCD values. Thirdly, we observed that multiple-frame exemplars is apparently helpful to avoid sharp changes across frames. Furthermore, we observed that using monophones together with biphones enhances the conversion performance.

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 194

TABLE III
THE EFFECT OF RESIDUAL COMPENSATION FOR SPECTRUM CONVERSION. 35 PARALLEL UTTERANCES ARE USED DURING TRAINING. MONOPHONES TOGETHER WITH BIPHONES ARE USED IN PHONETIC DICTIONARY

Spectrum Conversion	Script: Training	Script: Testing	MCD
PSR-TF	yes	yes	4.98
	yes	no	5.06
	no	no	5.12
PSR-TF-RC	yes	yes	4.85
	yes	no	4.90
	no	no	4.99

TABLE IV
THE EFFECT OF RESIDUAL COMPENSATION IN PROSODY CONVERSION. 35 PARALLEL UTTERANCES ARE USED DURING TRAINING. MONOPHONES TOGETHER WITH BIPHONES ARE USED IN PHONETIC DICTIONARY

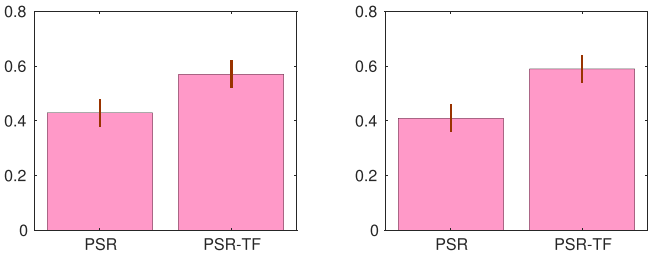
F0 Conversion	Script: Training	Script: Testing	PCC
PSR-TF	yes	yes	0.908
	yes	no	0.899
	no	no	0.891
PSR-TF-RC	yes	yes	0.914
	yes	no	0.906
	no	no	0.901

Lastly, by comparing PSR with PSR-TF, we show the effect of Tandem Feature in achieving a better conversion.

2) Prosody Conversion Performance: Table V reports the PCC and RMSE values of F0 and energy contour conversion for a number of settings in a comparative study. To start with, we observe that the proposed prosody conversion, based on phonetic sparse representation with Tandem Feature (PSR-TF), outperforms the PSR [32] that doesn't use Tandem Feature. The proposed PSR-TF uses the frame-level phonetic information. As a result, the activation matrix of PSR-TF depends less on the source speaker than that of PSR approach, hence yields a better prosody conversion. We also note that group sparse representation (GSR) outperforms the linear conversion [71] and the exemplar-based sparse representation (SR) [31] that doesn't use phonetic dictionary.

3) The Effect of Tandem Feature: Table II and V confirm the effectiveness of the novel idea to incorporate PPGs as phonetic features. The phonetic features allow for an improved activation matrix in Eq. (2), Eq. (7), and Eq. (11), that depends less on the source speaker. If we consider PSR [15], [32] performs well because of the use of phonetic information through a phonetic dictionary, the proposed PSR-TF benefits from both segmental level (phonetic dictionary) and frame level (PPGs) phonetic information. We observe that the use of phonetic information in both segmental level and frame level is rewarding for spectrum as well as prosody conversion.

4) The Effect of Residual Compensation: We report a phone dependent residual estimation and compensation by using dynamic partial least square (DPLS) regression for spectrum and prosody features under the PSR framework. Table III reports the results in a comperative study for spectrum conversion. We observed that all PSR-TF-RC settings outperform the PSR-TF settings by yielding lower MCD values. Moreover, we have conducted some experiments to see the effect of residual compensation in prosody conversion, as reported in Table IV. As expected, with residual compensation, we achieve higher correlation, hence a better prosody conversion.



(a) Voice quality (b) Speaker similarity

Fig. 7. The preference percentage tests with 95 % confidence interval between PSR and PSR-TF for joint spectrum and prosody conversion. The p-values [72] are (a) $7.0382e^{-21}$ and (b) $2.0564e^{-21}$ respectively.

D. Subjective Evaluation

We further conduct listening experiments to assess the performance of Tandem Feature for joint spectrum and prosody conversion, and the effect of WaveNet vocoder in terms of voice quality and speaker similarity. In all listening experiments, we use the scripts of the training utterances to perform forced alignment. It should be noted that at run-time conversion, we do not perform forced alignment. 15 subjects participated in all the listening tests, each of them listens to 30 converted utterances. The p-values are calculated in a similar way that is reported in [72].

1) Tandem Feature: We first conduct listening experiments, as reported in Fig. 7.a and 7.b, to examine the effect of Tandem Feature in terms of voice quality and speaker similarity for both spectrum and prosody conversion. We note that the original PSR framework [15] does not use the WaveNet vocoder to generate speech waveform. To assess the effect of Tandem Feature, without changing the vocoding process, we use WORLD in both PSR and PSR-TF in these experiments. Each listener is asked to decide the better sample in terms of voice quality and speaker similarity. We observe that PSR-TF outperforms the baseline PSR consistently in both voice quality and speaker similarity, that validates the proposal of the use of Tandem Feature in sparse representation voice conversion.

2) Phone-dependent Residual Compensation: We also conduct listening experiments, as reported in Fig. 8.a and 8.b, to examine the effect of phone-dependent residual compensation. We observe that PSR-TF with residual compensation outperforms PSR-TF consistently in both voice quality and speaker similarity.

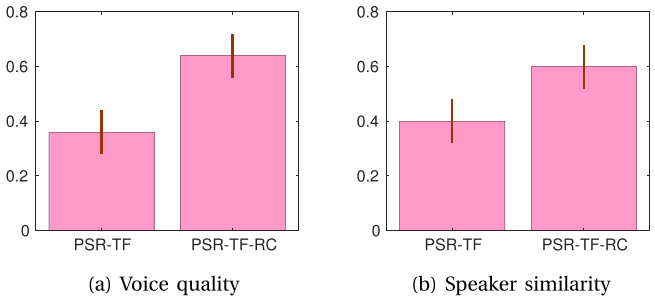
3) Speaker Independent and Adapted WaveNet Vocoder: We futher conduct a listening experiment, that is reported in Fig. 9, to study the listener preference of WaveNet vocoders. We train the speaker independent WaveNet vocoder with CMU Arctic and VCC 2016 datasets, in a similar way that is described in [21]. It is important to note that the speech data from any of the source or target speakers in the experiment are excluded during training. To perform WaveNet adaptation, we use the same 35 target utterances that are used in the training of sparse represen-

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 195

TABLE V
A COMPARISON OF PCC VALUES OF CONVERTED F0 AND ENERGY, DENOTED AS E, ACROSS SEVERAL SPARSE REPRESENTATION IMPLEMENTATIONS. WE ALSO REPORT LINEAR F0 CONVERSION RESULT AS THE F0 CONVERSION BASELINE. PSR-TF AND GSR-TF ARE PROPOSED IN THIS PAPER, WHILE LINEAR CONVERSION, PSR [15] AND SR [11] WERE STUDIED PREVIOUSLY

Prosody Conversion	Features	Number of Frames	Phonetic Dict.	Script: Training	Script: Testing	PCC: F0	PCC: E
PSR-TF	Tandem Feature	1	M	yes	yes	0.861	0.831
	Tandem Feature	3	M+B	yes	yes	0.901	0.850
	Tandem Feature	3	M+B	yes	no	0.892	0.843
	Tandem Feature	3	M+B	no	no	0.881	0.837
PSR [15]	spectrum+prosody	1	M	yes	yes	0.847	0.819
	spectrum+prosody	3	M+B	yes	yes	0.893	0.836
	spectrum+prosody	3	M+B	yes	no	0.879	0.831
	spectrum+prosody	3	M+B	no	no	0.865	0.825
GSR-TF	Tandem Feature	3	M	yes	no	0.839	0.805
	Tandem Feature	3	M	no	no	0.830	0.793
SR-TF	spectrum+prosody	3	-	-	-	0.827	0.779
SR [11]	spectrum+prosody	3	-	-	-	0.810	0.765
Linear Conversion	-	3	-	-	-	0.780	N. A.



(a) Voice quality (b) Speaker similarity

Fig. 8. The preference percentage tests with 95 % confidence interval between PSR-TF and PSR-TF-RC for joint spectrum and prosody conversion. The p-values [72] are (a) $4.2350e^{-23}$ and (b) $2.0564e^{-21}$ respectively.

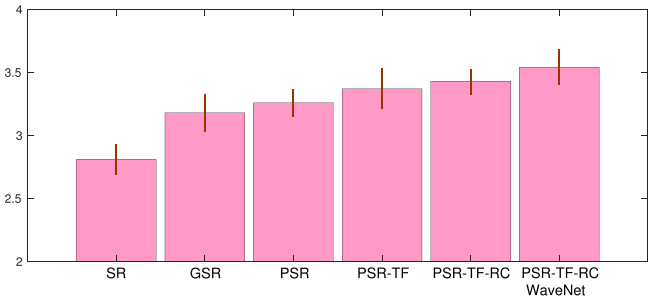


Fig. 10. A comparison of MOS for joint spectrum and prosody conversion over the sparse representation schemes discussed in this paper, that include SR [11], PSR[15], GSR, PSR-TF, PSR-TF-RC and PSR-TF-RC with adapted WaveNet vocoder (PSR-TF-RC-WaveNet).

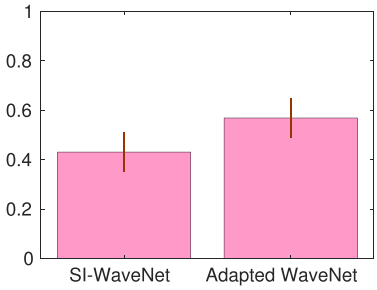


Fig. 9. The preference test between the speaker independent WaveNet vocoder [21] and the adapted WaveNet vocoder in terms of voice quality. The p-value [72] is $2.3393e^{-20}$.

tation. At run-time conversion, we use the speaker independent WaveNet [21] and the adapted WaveNet vocoder to generate the target speech waveforms. We observe that Adapted WaveNet outperforms the speaker independent WaveNet in terms of voice quality, that validates our proposed idea. Motivated by the encouraging results of Tandem Feature and the adapted WaveNet for joint spectrum and prosody conversion, we now move on to the last listening experiment, that is reported in Fig. 10, to assess the performance of the proposed PSR-TF-RC with adapted WaveNet vocoder. We perform adaptation by using the same 35 utterances from the target speaker, that have been used for TF-dictionary construction. We evaluate the sound quality of the converted voices by using the mean opinion score (MOS). The listeners rate the quality of the converted voice using a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. Fig. 10 shows that the voice conversion framework PSR-TF-RC with adapted WaveNet vocoder outperforms others that include SR, GSR, PSR, PSR-TF and PSR-TF-RC.

VIII. CONCLUSION

We have proposed a novel voice conversion framework that includes spectrum, F0 and energy contour transformation. We provide a theoretical formulation for phonetic sparse representation by proposing group sparse representation approach for voice conversion. Moreover, we study the speaker dependent and independent characteristics (phonetic activation matrix and differentiated multi-scale prosody transfer) to motivate the design of sparse representation framework. We show two effective ways to improve the conversion performance under limited training data. Firstly, we propose the Tandem Feature sparse representation strategy for spectrum and prosody features. We implement the TF-dictionary in phonetic sparse representation frameworks as a solution to improve the conversion performance under limited training data. Secondly, we propose the use of speaker adapted WaveNet vocoder together with a conversion framework to achieve higher voice quality. Experiment results show that the proposed framework makes good use of

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 196

the limited training data and outperforms the baselines in both objective and subjective evaluations.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1988, vol. 2, pp. 655–658.

[2] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *Proc IEEE Int. Symp. Circuits Syst.*, 1991, pp. 594–597.

[3] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[4] H. Zen, Y. Nankaku, and K. Tokuda, "Probabilistic feature mapping based on trajectory HMMs," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 1068–1071.

[5] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. IV-1249–IV-1252.

[6] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 1667–1671.

[7] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 912–921, Jul. 2010.

[8] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Workshop Spoken Lang. Technol.*, 2012, pp. 313–317.

[10] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, and K. Hirose, "Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1764–1767.

[11] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1506–1521, Oct. 2014.

[12] R. Aihara, K. Masaka, T. Takiguchi, and Y. Ariki, "Parallel dictionary learning for multimodal voice conversion using matrix factorization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 27–40. [Online]. Available: http://link.springer.com/10.1007/978-3-319-40171-3_3

[13] Z. Jin, A. Finkelstein, S. Di Verdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar- based unit selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 2–6.

[14] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 7894–7898.

[15] B. Sisman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *Proc. IEEE Autom. Speech Recog. Understanding Workshop*, 2017, pp. 677–684.

[16] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, 2009, pp. 421–426.

[17] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1905–1908.

[18] L. Sun, H. Wang, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 322–326.

[19] A. Van Den Oord *et al.*, "Wavenet: A generative model for raw audio," 2016, arXiv preprint arXiv:1609.03499.

[20] A. Tamamori, T. Hayashi, and K. Kobayashi, "Speaker-dependent wavenet vocoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1118–1122.

[21] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 712–718.

[22] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1138–1142.

[23] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[24] S.-S. Wang *et al.*, "Wavelet speech enhancement based on nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 23, no. 8, pp. 1101–1105, Aug. 2016.

[25] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[26] K. A. Akarsh, "Speech enhancement using non negative matrix factorization and enhanced NMF," in *Proc. Int. Conf. Circuit, Power Comput. Technol.*, 2015, pp. 1–7.

[27] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4029–4032.

[28] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback - Leibler divergence," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[29] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Examplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[30] Y. C. Wu, H. T. Hwang, C. C. Hsu, Y. Tsao, and H. M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 1652–1656.

[31] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *Proc. IEEE Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 5175–5179.

[32] B. Sisman, H. Li, and K. C. Tan, "Transformation of Prosody in voice conversion," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 1537–1546.

[33] H. Chin-Cheng, H. Hsin-Te, W. Yi-Chiao, T. Yu, and W. Hsin-Min, "Dictionary update for NMF-based voice conversion using an encoder-decoder network," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process.*, 2016, vol. 22, no. 3, pp. 293–297.

[34] D. L. Bolinger, "Intonation as a universal," *Int. COngr. Linguists, Cambridge*, 1962.

[35] J. J. Ohala, A. Dunn, and R. Sprouse, "Prosody and Phonology," *ISCA Speech Prosody*, 2004.

[36] G. Sanchez, H. Silen, J. Nurminen, and M. Gabbouj, "Hierarchical modeling of F0 contours for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2318–2321.

[37] "Speech prosody: A methodological review," *J. Speech Sci.*, vol. 1, no. 1, pp. 85–115, 2015.

[38] J. Latorre, "Multilevel parametric-base F0 model for speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, May 2014, pp. 2274–2277.

[39] M. S. Ribeiro and R. A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4909–4913.

[40] A. Wennerstrom, *The Music of Everyday Speech Prosody and Discourse Analysis*. Oxford, U.K.: Oxford Univ. Press, 2001, pp. 153–158.

[41] D. R. Ladd, *Intonational Phonology*. Cambridge, U.K.: Cambridge Univ. Press, 2008, pp. 153–158.

[42] B. K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[43] M. Vainio, A. Suni, and D. Aalto, "Continuous wavelet transform for analysis of speech prosody," in *Proc. Tools Resources Anal. Speech Prosody*, 2013, pp. 78–81.

[44] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013, pp. 285–290.

[45] H. Ming, D. Huang, M. Dong, H. Li, L. Xei, and S. Zhang, "Fundamental frequency modeling using wavelets for emotional voice conversion," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 804–809.

[46] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2016, pp. 2453–2457.

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 197

[47] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1732–1735.
[48] B. Gillett, S. King, and U. Kingdom, "Transforming F0 Contours," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2003, pp. 101–104.
[49] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP J. Audio, Speech, Music Process.*, vol. 2017, no. 1, 2017, Art. no. 116.
[50] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3399–3403.
[51] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," 2018, arXiv preprint arXiv:1808.05092.
[52] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," 2018, arXiv preprint arXiv:1806.02169.
[53] B. Sisman, G. Lee, and H. Li, "Phonetically aware exemplar-based prosody transformation," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018.
[54] B. Sisman and H. Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 52–56.
[55] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 2135–2138.
[56] R. Serizel, S. Essid, and G. Richard, "Group non-negative matrix factorisation with speaker and session variability compensation for speaker identification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 5470–5474.
[57] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.
[58] J. Kim and H. Park, "Group sparsity in nonnegative matrix factorization," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 851–862.
[59] R. Aihara, T. Takiguchi, and A. Yasuo, "Activity-mapping non-negative matrix factorization for exemplar-based voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 4899–4903.
[60] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.
[61] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[62] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2017, arXiv preprint arXiv:1409.0473.
[63] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances Neural Inf. Process. Syst.*, 2017, pp. 1878–1889.
[64] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," 2017, arXiv.
[65] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," 2017, arXiv.
[66] S. Furui, "Digital speech processing, synthesis, and recognition(revised and expanded)," in *Digital Speech Processing, Synthesis, and Recognition*. Boca Raton, FL, USA: CRC Press, 2000.
[67] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on Mel spectrogram predictions," in *IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2018, pp. 4779–4783.
[68] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process.*, 1993, pp. 125–128.
[69] T. Toda *et al.*, "The voice conversion challenge 2016," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 1632–1636.
[70] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011.
[71] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "A style capturing approach to F0 transformation in voice conversion,"in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 6915–6919.
[72] S. Lifa, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2016, pp. 1–6.



Berrak Sisman received two Bachelor's degrees in electrical and electronic engineering, and computer engineering from Isik University, Istanbul, Turkey, in 2015. She received the Master's of Science degree in electronics engineering from Graduate School of Science and Engineering, Isik University, in 2016. She is currently working toward the Ph.D. degree with the National University of Singapore, Singapore. Her current research interests include speech synthesis and voice conversion. She is the General Coordinator of the Student Advisory Committee of International Speech Communication. She received Singapore International Graduate Award for her Ph.D studies. She was awarded APSIPA PhD Forum Best Presentation Award for her presentation titled "Limited Data Voice Conversion from Sparse Representation to GANs and WaveNet" in 2018, Hawaii, USA.



Mingyang Zhang received the B.Eng. degree in electrical information engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014. He is currently working toward the Ph.D. degree with Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, China. He has been with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current research interests include speech synthesis and voice conversion.



Haizhou Li (M'91–SM'01–F'14) received the B.Sc., M.Sc., and Ph.D. degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. He is also a Conjoint Professor with the University of New South Wales, Australia. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Before joining NUS, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), a Research Manager with the Apple-ISS Research Centre (1996–1998), a Research Director in Lernout and Hauspie Asia Pacific (1999–2001), Vice President in InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology with the Institute for Infocomm Research, Singapore (2003–2016). He is currently the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015–2018), a Member of the Editorial Board of Computer Speech and Language (2012–2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012 and INTERSPEECH 2014. He is a Fellow of the ISCA. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019.

Authorized licensed use limited to: Penn State University. Downloaded on October 27,2025 at 18:04:16 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 198

Appendix SISMAN02

MARC View

LEADER 02283nas a22004817i 4500			
001 49178302			
003 SIRSI			
005 20251014090047.0			
006 m o d			
007 cr n			
008 130709c20149999njumr p s | a0eng			
010			a| 2013203345
019			a| DO NOT EDIT this 360 MARC RECORD
022			a| 2329-9304 y| 2329-9290
030			a| ITASFA
035			a| (WaSeSS)ssj0001079974
037			b| IEEE, 445 Hoes Ln., Piscataway, NJ 08854
040			a| DLC b| eng c| DLC d| OCLCA d| DKU d| OCLCF d| OCLCO d| VT2 d| OCLCO d| OCLCA d| W2U d| UUM d| DLC d| OCLCQ d| OCLCL d| EZC d| WaSeSS
042			a| pcc a| nsdp
082	0	0	a| 621.382 2| 23 1| https://id.oclc.org/worldcat/ddc/E36HC9jBKfpGqy8cdwf3r7FPbf
210	0		a| IEEE/ACM trans. audio, speech lang. process. b| (Online)
222		0	a| IEEE/ACM transactions on audio, speech, and language processing b| (Online)
245	0	0	a| IEEE/ACM transactions on audio, speech, and language processing h| [electronic resource] : b| a publication of the Signal Processing Society.
246	1		a| Transactions on audio, speech, and language processing
246	1		a| Audio, speech, and language processing
246	1		a| TASLP
260			a| [Piscataway, NJ] : b| [Institute of Electrical and Electronics Engineers], c| [2014]-
310			a| Monthly
362	1		a| Began in 2014
538			a| Mode of access: World Wide Web.
588	0		a| Print version record.
588	1		a| Volume 28 (2020) (IEEExplore viewed Feb. 13, 2020).
650		0	a| Speech processing systems v| Periodicals.
650		0	a| Electro-acoustics v| Periodicals.
655		0	a| Electronic journals.
710	2		a| IEEE Signal Processing Society.
710	2		a| Institute of Electrical and Electronics Engineers.
710	2		a| Association for Computing Machinery.
776	0	8	i| Print version: t| IEEE/ACM transactions on audio, speech, and language processing x| 2329-9290 w| (DLC) 2013203346 w| (OCoLC)852251724
780	0	0	t| IEEE transactions on audio, speech and language processing (Online) x| 1558-7924 w| (DLC) 2005215814 w| (OCoLC)62197773
856	4	0	u| http://SK8ES4MC2L.search.serialssolutions.com/?V=1.0&L=SK8ES4MC2L&S=JCs&C=TC0001079974&T=marc
910			a| CONSER online
949			a| Electronic resource w| ASIS c| 1 i| 49178302-1001 l| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 10/18/2025 o| .STAFF. 360 MARC

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

(814) 865-6368

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

IEEE/ACM transactions on audio, speech, and language processing [electronic resource] : a publication of the Signal Processing Society

Additional Titles	IEEE/ACM trans. audio, speech lang. process. (Online), Transactions on audio, speech, and language processing, Audio, speech, and language processing, and TASLP
Published	[Piscataway, NJ] : [Institute of Electrical and Electronics Engineers], [2014]-
Additional Creators	IEEE Signal Processing Society, Institute of Electrical and Electronics Engineers, and Association for Computing Machinery

Access Online

serialssolutions.com

Continues

IEEE transactions on audio, speech and language processing (Online)

Dates of Publication and/or Sequential Designation

Began in 2014

Subject(s)

Speech processing systems—Periodicals

Electro-acoustics—Periodicals

Genre(s)

Electronic journals

ISSN

2329-9304

Publication Frequency

Monthly

Technical Details

Mode of access: World Wide Web.

ASK A LIBRARIAN

View MARC record | catkey: 49178302

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368

Libraries Intranet (Staff Only)

Accessibility Help

Website Feedback

Policies and Guidelines

Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook

Twitter

Instagram

Online Journals, Books and more

All | Journals Only | Books Only | Videos Only

Title begins with | Search

Show more search options ▼

Showing results 1 through 1 of 1

for the search: "TC0001079974"

Refine Results: All | Journals Only | Books Only | Videos Only

Limit by: Peer Reviewed | Open Access

IEEE/ACM transactions on audio, speech, and language processing

ISSN: 2329-9290

Peer Reviewed

01/01/2014 to 01/31/2024 in ACM Digital Library

2014 to Present in IEEE Electronic Library (IEL)

Search inside this journal | Search

Back to Top

Copyright ©2025 The Pennsylvania State University.
All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Last Updated on 11-14-2025 © Ex Libris, a ProQuest Company, 2000-2025

Appendix LOOTS01

MARC View

LEADER 02401nas a2200529 a 4500			
001 49559717			
003 SIRSI			
005 20251114090051.0			
006 m o d			
007 cr n			
008 980610c19829999ne br pss 0 b0eng c			
010			a| 2003233217
016	7		a| 1460279-9 2| DE-600
016	7		a| 1460279 9. 2| DE-600
016	7		a| 012435072 2| Uk
016	7		a| 1460279 9 2| DE-600
019			a| DO NOT EDIT this 360 MARC RECORD
022			a| 1872-7182 l| 0167-6393 y| 0167-6393
030			a| SCOMDH
035			a| (WaSeSS)ssj0004882
040			a| OH1 b| eng c| OH1 d| OCL d| OCLCQ d| GUA d| OCLCQ d| VRC d| U9S d| GUA d| OCLCQ d| AZU d| LGG d| OCLCQ d| OCLCO d| OCLCQ d| OCLCO d| OCLCF d| OCLCO d| OCLCQ d| BUF d| U3W d| WYU d| VT2 d| UHL d| LIP d| OCLCQ d| UKSSU d| UK7LJ d| OCLCQ d| OCLCO d| OCL d| OCLCQ d| UAB d| LDP d| UKUAL d| UKMGB d| DXU d| SFB d| OCLCL d| WaSeSS
042			a| pcc
050	1	4	a| BF455
130	0		a| Speech communication (Online)
222		0	a| Speech communication b| (Online)
245	1	0	a| Speech communication h| [electronic resource].
246	1		i| Also known as: a| Speech com
260			a| [Amersterdam] : b| Elsevier Science
310			a| Bimonthly, b| 2002-
321			a| Frequency varies, b| <1995>-2001
362	1		a| Began with vol. 1, issue 1 (May 1982).
530			a| Also issued in print.
538			a| Mode of access: World Wide Web.
550			a| A Publication of the European Association for Signal Processing (EURASIP) and the International Speech Communication Association (ISCA).
588	0		a| Vol. 16, issue 1 (Jan. 1995); title from contents screen (ScienceDirect, viewed October 17, 2003).
588	1		a| Vol. 41, issues 2/3 (Oct. 2003), viewed October 17, 2003.
650		0	a| Oral communication v| Periodicals.
650		0	a| Communication.
650		2	a| Communication
655		0	a| Electronic journals.
655		2	a| Periodical
710	2		a| European Association for Signal Processing. 1| https://id.oclc.org/worldcat/entity/E39QH7JmrGHjqb8hFgKxpqcKtg.
710	2		a| International Speech Communication Association. 1| https://id.oclc.org/worldcat/entity/E39QH7Jmp8HCf8gyJ94gjq7b9t.
776	1		t| Speech communication x| 0167-6393 w| (DLC)sc 83001666 w| (OCoLC)8843982
856	4	0	u| http://SK8ES4MC2L.search.serialssolutions.com/?V=1.0&L=SK8ES4MC2L&S=JCs&C=SPEECOM&T=marc
910			a| CONSER online
949			a| Electronic resource w| ASIS c| 1 i| 49559717-1001 l| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 11/29/2025 o| .STAFF. 360 MARC

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368

Libraries Intranet (Staff Only)

Accessibility Help

Website Feedback

Policies and Guidelines

Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook

Twitter

Instagram

Speech communication [electronic resource].

Uniform Title	Speech communication (Online)
Additional Titles	Also known as: Speech com
Published	[Amersterdam] : Elsevier Science
Additional Creators	European Association for Signal Processing and International Speech Communication Association

Access Online

serialssolutions.com

Dates of Publication and/or Sequential Designation
Began with vol. 1, issue 1 (May 1982).

Subject(s)
Oral communication—Periodicals
Communication

Genre(s)
Electronic journals

ISSN
1872-7182

Publication Frequency
Bimonthly, 2002-
Frequency varies, <1995>-2001

Other Forms
Also issued in print.

Technical Details
Mode of access: World Wide Web.

Issuing Body
A Publication of the European Association for Signal Processing (EURASIP) and the International Speech Communication Association (ISCA).

View MARC record | catkey: 49559717

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES
Libraries Home
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory
(814) 865-6368

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES
Facebook
Twitter
Instagram

Online Journals, Books and more

All | Journals Only | Books Only | Videos Only

Title begins with | Search

Show more search options ▼

Showing results 1 through 1 of 1
for the search: "SPEECOM"

Refine Results: All | Journals Only | Books Only | Videos Only

Limit by: Peer Reviewed | Open Access



Speech communication
ISSN: 0167-6393
Peer Reviewed
1982 to 1994 in Backfile Package - Computer Science (Legacy) [YCS]
05/01/1982 to Present in ScienceDirect Journals

Search inside this journal | Search

Back to Top

Copyright ©2025 The Pennsylvania State University.
All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Last Updated on 12-19-2025 © Ex Libris, a ProQuest Company, 2000-2025

Appendix LOOTS02

ScienceDirect

Speech Communication

Volume 53, Issue 1, January 2011, Pages 75-84

Automatic conversion between pronunciations of different English accents ☆

Linsen Loots ✉, Thomas Niesler ✉

Show more

Outline | Share Cite

https://doi.org/10.1016/j.specom.2010.07.006 ↗

Get rights and content ↗

Full text access

Abstract

We describe the application of decision trees to the automatic conversion of pronunciations between American, British and South African English accents. The resulting phoneme-to-phoneme (P2P) conversion technique derives the pronunciation of a word in a new target accent by taking advantage of its existing available pronunciation in a different source accent. We find that it is substantially more accurate to derive pronunciations in this way than directly from the orthography and available target accent pronunciations using more conventional grapheme-to-phoneme (G2P) conversion. Furthermore, by including both the graphemes and the phonemes of the source accent, grapheme-and-phoneme-to-phoneme (GP2P) conversion delivers additional increases in accuracy in relation to P2P. These findings are particularly important for less-resourced varieties of English, for which extensive manually-prepared pronunciation dictionaries are not available. By means of the P2P and GP2P approaches, the pronunciations of new words can be obtained with better accuracy than is possible using G2P methods.

Previous Next

Keywords

English accents; Pronunciation modelling; G2P; P2P; GP2P; Decision trees; South African English

1. Introduction

The pronunciation dictionary is a key component of an Automatic Speech Recognition (ASR) as well as a text to speech (TTS) system. However, the development of this dictionary involves a great deal of time and effort by linguistic experts, a process that may be prohibitively expensive for under-resourced languages and accents. Nevertheless phonetic transcriptions in the appropriate accent can improve both ASR accuracy (Humphries et al., 2001) and TTS quality (Damper et al., 1999).

IPR2026-00272
Krisp Technologies EX1010 Page 208

We consider the scenario where the pronunciations of words not in the dictionary are sought, and where these pronunciations are already available in a different accent of the language in question. It will be shown that, in this situation, data-driven techniques similar to those used in grapheme-to-phoneme (G2P) conversion can be used to obtain the desired pronunciation from the existing pronunciation, instead of from the graphemes. This amounts to a conversion between the two accents.

In our experimental evaluation, we focus on General American (GenAm), Received Pronunciation (RP) and Standard South African English (SSAE). The first two are commonly-used and widely-studied reference accents for American and British English, respectively (Wells, 1982). In South Africa, a number of different English accents are prevalent. The accent generally used by first-language speakers is commonly referred to as Standard South African English and has been described as fairly similar to RP (Bowerman, 2004, Bekker, 2009). An important distinguishing factor of SSAE is the characteristic split of Wells' *KIT* vowel. This vowel appears in words like *kit*, *thin* and *sit*, and in SSAE has two clear allophonic variations: one close and one more central (Wells, 1982). While these words contain a single vowel sound in most English accents, they are pronounced using two different vowel sounds in SSAE.

From our perspective the objective of accent conversion is to determine whether British or American pronunciations can be used to complement our still fairly small SSAE pronunciation dictionary, and how this can be done best. However, we believe that the techniques we consider are generally applicable to any accent pair.

The following section will describe the data upon which our experimental evaluation is based. Sections 3 Grapheme-to-phoneme conversion (G2P), 4 Phoneme-to-phoneme conversion (P2P), 5 Grapheme-and-phoneme-to-phoneme conversion (GP2P) then describe three alternative techniques that can be used to obtain pronunciations of new words. Finally, Section 7 discusses the results and Section 8 concludes.

2. Dictionaries

Four dictionaries were used to represent the three accents which we consider, and are described shortly in the following.

CMUDICT and PRONLEX were used to represent GenAm pronunciations. CMUDICT was developed at Carnegie Mellon University (CMU, 2009). This dictionary contains words drawn from a variety of sources, including an initial subset of over 20,000 words developed by hand and extensively verified at CMU. Other sources of pronunciations for CMUDICT include the Shoup and the Dragon dictionaries. PRONLEX is the COMLEX English Pronouncing Lexicon, and was hand-transcribed at the Linguistic Data Consortium (LDC, 2009).

The BEEP dictionary was used to represent RP and is drawn primarily from the Medical Research Council (MRC) Psycholinguistic Database and the Computer Usable Oxford Advanced Learners Dictionary (BEEP, 2009). In addition, it contains pronunciations provided by sources at Durham and Oxford Universities.

SSAE pronunciations were obtained from SAEDICT, a pronunciation dictionary under development at Stellenbosch University. All pronunciations in SAEDICT were transcribed by the same linguistic specialist to ensure consistency. Transcriptions were chosen to reflect commonly accepted SSAE pronunciations.

Table 1 describes each dictionary in terms of the number of words it contains, the total number of pronunciations and the average number of pronunciations per word (P/W). The number of pronunciations and P/W is also given for a set of 23,031 words common to all four dictionaries, as will be described in Section 2.2.

Table 1. Number of entries in the four pronunciation dictionaries.

Dictionary	Entire dictionary			Common set		
	Words	Prons	P/W	Words	Prons	P/W
SAEDICT	30,390	36,954	1.22	23,031	28,087	1.22

IPR2026-00272
Krisp Technologies EX1010 Page 209

Dictionary	Entire dictionary			Common set		
	Words	Prons	P/W	Words	Prons	P/W
BEEP	237,592	256,998	1.08	23,031	25,576	1.11
CMUDICT	123,292	133,694	1.08	23,031	26,771	1.16
PRONLEX	90,919	99,978	1.11	23,031	25,471	1.11

2.1. Phoneset

ARPABET (Rabiner and Juang, 1993) was chosen as the common phoneset in which to analyse the four dictionaries. Both BEEP and CMUDICT already use ARPABET, while PRONLEX uses an ARPABET short-hand that can easily be converted to standard ARPABET. SAEDICT uses a phoneset, based on IPA, developed to describe the languages of Southern Africa (Niesler et al., 2005). This was converted to ARPABET by means of a mapping based on the closest IPA symbol.

A small number of phonemes was found to be present in only a single dictionary, and these were mapped to alternative phonemes as indicated in Table 2. This normalisation ensures that the phonesets used by the four dictionaries match exactly.

Table 2. Phoneme mappings used to achieve a common phoneset.

Dictionary	Mapping
BEEP	*/oh/⇒/aa/*
	/ia/⇒/ih ax/
	/ea/⇒/eh ax/
	/ua/⇒/uh ax/
PRONLEX	*/wh/⇒/w/*
	/en/⇒/ax n/

2.2. Wordlist

In order to compare pronunciations, the set of words common to all four dictionaries was determined. Before extracting this set, all words were automatically converted to standard UK spelling using tools developed at the University of Sheffield (Wan et al., 2008). The set of common words contained 23,031 entries.

There are a number of words in each dictionary with multiple pronunciations, either homographs (words with the same spelling but different pronunciations), or words with alternate pronunciations. The number of pronunciations vary, but all four dictionaries contain an average of between 1.11 and 1.22 pronunciations per word, as reflected in Table 1. Of the 23,031 words common to all four dictionaries, 9344 have multiple pronunciations in at least one dictionary.

2.3. Pronunciation alignments

The pronunciations of corresponding words in the three accents were aligned using a dynamic programming algorithm that will be described in Section 3.1. This was done so that accents could be compared directly, as will be described in the following section. However, the alignments will also be used to obtain the one-to-one correspondences between phonemes in different accents. These correspondences are needed to train the decision trees that will be responsible for

IPR2026-00272
Krisp Technologies EX1010 Page 210

inter-accent conversions (described in Sections 4 Phoneme-to-phoneme conversion (P2P), 5 Grapheme-and-phoneme-to-phoneme conversion (GP2P)).

As an example, the alignment between SSAE and RP pronunciations of the word “reactions” is shown in Table 3. Nulls have been introduced where a phoneme has no counterpart in the other string (i.e. insertions and deletions). In this case the /*ax*/ in the SSAE pronunciation is deleted in RP.

Table 3. Two aligned pronunciations of the word “reactions”.

SAEDICT	***r***	***ih***	***ae***	***k***	***sh***	***ax***	***n***	***s***
BEEP	*r*	*ih*	*ae*	*k*	*sh*	–	*n*	*z*

Words with multiple pronunciations pose a problem both when comparing two dictionaries, and when subsequently using the aligned pronunciations to train decision trees for accent conversion. Our solution has been to use only the single pair of pronunciations per word (one from each dictionary) that gives the best alignment. An alternative approach is to consider all the possible pairings of pronunciations between the two dictionaries. However, preliminary experiments indicated that this did not influence the relative accuracies obtained for the various accents. Hence, the simpler approach of using the best-aligned pair was chosen for all further experiments.

2.4. Direct pronunciation comparison

In order to obtain a baseline for inter-accent similarity against which accent conversion can later be compared, a direct phonetic comparison of the different accents was performed. A summary of the results obtained for the 23,031 common words described in Section 2.2 is given in Table 4. The phoneme accuracy was calculated according to Eq. (1), where N_c, N_i and N_t are the numbers of correct, inserted and total phonemes, respectively. This accuracy is given for all phonemes, and separately for vowels and for consonants. The last column indicates the percentage of words that have identical pronunciations.

$$Acc = \frac{N_c - N_i}{N_t}. \tag{1}$$

Table 4. Accuracies (%) of direct pronunciation alignments between dictionaries.

Dictionary pair	**Phoneme**	**Vowel**	**Cons**	**Word**
BEEP ⇔ SAEDICT	92.1	84.5	96.9	59.6
CMUDICT ⇔ SAEDICT	81.6	59.8	95.2	29.4
PRONLEX ⇔ SAEDICT	87.3	72.8	96.2	43.0
CMUDICT ⇔ BEEP	84.7	67.7	94.6	37.8
PRONLEX ⇔ BEEP	89.4	79.0	95.6	50.6
PRONLEX ⇔ CMUDICT	89.0	73.2	98.6	50.7

It is clear from Table 4 that SAEDICT and BEEP match most closely, while the poorest correspondence is found between CMUDICT and SAEDICT. On average, the American pronunciations are closer to RP than to SSAE. Furthermore, vowels differ more strongly than consonants between accents. While at least 94.6% of consonants matched between any pair of dictionaries, this figure varied between 59.8% and 84.5% for vowels. A more detailed analysis of the phonetic differences between the three accents can be found in (Loots and Niesler, 2009).

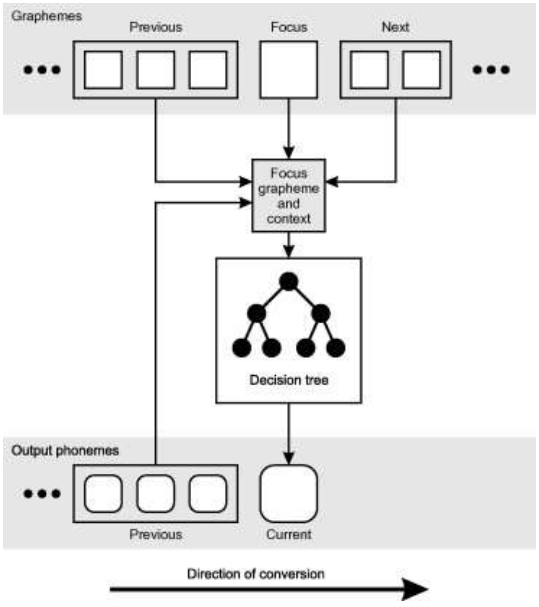
3. Grapheme-to-phoneme conversion (G2P)

IPR2026-00272
Krisp Technologies EX1010 Page 211

G2P conversion is the process of determining a word's pronunciation from its spelling. Several data-driven G2P methods have been suggested in the literature, including decision trees (Black et al., 1998, Suontausta and Häkkinen, 2000, Kienappel and Kneser, 2001, Webster and Braunschweiler, 2008), HMMs (Taylor, 2005), pronunciation by analogy (Marchand and Damper, 2000) and memory-based learning (Daelemans et al., 1999). Decision trees have been shown to yield competitive accuracy in G2P conversion (Han and Chen, 2004) and to be a highly competitive approach to the generation of accent-specific pronunciation variants for ASR (Humphries et al., 2001). We have chosen decision trees as the method with which we will investigate the success of accent conversion.

Although this paper does not focus on G2P systems, the methods used for accent conversion in Sections 4 Phoneme-to-phoneme conversion (P2P), 5 Grapheme-and-phoneme-to-phoneme conversion (GP2P) are based on analogous techniques. Hence, G2P conversion by means of decision trees will be reviewed first, and used as a further baseline against which the later systems can be measured.

For G2P conversion, the graphemes and their context form the input of the decision tree classifier, and the phonemes the output class. Pronunciations are generated by sequentially passing graphemes and their context through the tree, assigning each to a phoneme class, and then concatenating these output phonemes. This process is illustrated in Fig. 1.



Download: Download full-size image

Fig. 1. Diagrammatic representation of grapheme-to-phoneme (G2P) conversion using decision trees.

3.1. Aligning graphemes and phonemes

In order to sequentially generate an output phoneme for each grapheme, the decision tree requires a one-to-one correspondence between the graphemes and the phonemes. An example of such an alignment is shown in Table 5.

Table 5. An example of grapheme-to-phoneme (G2P) alignment for the word "extreme".

Grapheme	**e**	**x**	**t**	**r**	**e**	**m**	**e**
Phoneme	*eh*	*k s*	*t*	*r*	*iy*	*m*	–

Discrepancies in length and alignment between the graphemes and phonemes are dealt with by inserting nulls into the phoneme string (for letters not corresponding to a phoneme), and by combining pairs of phonemes into double phonemes

IPR2026-00272
Krisp Technologies EX1010 Page 212

for the relatively few cases where one grapheme corresponds to two phonemes (Damper et al., 2004, Suontausta and Häkkinen, 2000).

The primary method of aligning graphemic and phonetic strings uses the Viterbi algorithm (Daelemans and van den Bosch, 1994, Andersen et al., 1996, Pagel et al., 1998). The process begins by estimating $P(G,P)$, the unigram probability that a given grapheme is associated with a given phoneme. This is estimated using those words that have graphemic and phonetic strings of equal length (Andersen et al., 1996). Optimal alignments based on these estimated probabilities are then generated by means of dynamic programming for all words in the corpus. The newly-aligned words are then used to recalculate the unigram probabilities, and the process is repeated until the distributions converge (Pagel et al., 1998, Webster and Braunschweiler, 2008).

Double phonemes can be determined automatically by iteratively repeating this process. First, suitable double phonemes are identified from unsuccessful alignments between grapheme and phoneme strings. Alignment is subsequently re-attempted, and the process repeated until all words align with their pronunciations.

3.2. Decision trees for G2P

We have employed deterministic binary decision trees for G2P conversion, for which each node is associated with a true/false question regarding the input grapheme and its context. The tree is traversed from the root by recursively using the answer of each node's question to determine which child node to choose. Each leaf node is associated with an output phoneme, which constitutes the classification result (Breiman et al., 1984). This process produces a single pronunciation for the presented orthography.

Decision trees are grown recursively. For each new node the available training data are split according to all possible questions. The question which results in the greatest entropy gain is then chosen for that node (Suontausta and Häkkinen, 2000). For a node t with entropy $i(t)$, children t_L and t_R with respective entropies $i(t_L)$ and $i(t_R)$, and the proportion of node t's data associated with each child given by p_L and p_R such that both are non-negative and sum to unity, the entropy gain is (Breiman et al., 1984):

$$\Delta i = i(t) - p_L i(t_L) - p_R i(t_R). \tag{2}$$

When choosing the question with the largest Δi, $i(t)$ can be omitted from the calculation with no loss of generality as it remains constant for all questions at a given node. Furthermore p_k can be approximated by the relative frequency N_{t_k}/N_t for a child node t_k, while $1/N_t$ is constant for all questions.

$$\Delta i = \sum_{\forall\, children\ \ k}\ \sum_{\forall\, phonemes\ \ p} N_{t_k,p} \log\left(\frac{N_{t_k,p}}{N_{t_k}}\right). \tag{3}$$

It is therefore sufficient to maximise Eq. (3) in order to determine the optimal question for a new decision tree node (Kienappel and Kneser, 2001).

3.3. Decision tree training and testing

A G2P decision tree includes many parameters, the most important of which are the number of graphemes and previously generated phonemes included in the context, the direction of the conversion, the use of clusters, and the use of pruning (Webster and Braunschweiler, 2008).

To obtain values for these parameters in our G2P experiments, the phoneme accuracy of the decision tree was optimised on a held-out development set within a 10-fold cross-validation framework. This was achieved by dividing the 23,031 common words described in Section 2.2 into 10 non-overlapping and approximately equally-sized partitions. Reserving each partition in turn for later use as a test set, the remaining 9 partitions (corresponding to 90% of the data) were again subdivided into 10 equal partitions. The first of these sub-partitions served as a development set, and the remaining 9 as a training set for parameter optimisation. Decision trees were trained on this training set for a range of parameter values, and those that lead to the highest phoneme accuracy measured on the associated development set were identified as

IPR2026-00272
Krisp Technologies EX1010 Page 213

optimal. This process was repeated for each of the 10 partitions of the 23,031-word data set. In general, different optimal parameter values were obtained in each case.

The data partitioning must be carefully considered, as training, development and testing datasets should be properly representative of the dictionary (and thus the language) as a whole. Furthermore, it is important that there is not too much overlap between the data partitions, as might occur between words with the same root but different affixes. This should be avoided because it can cause the test or development data to be unrealistically similar to the training data, leading to results that are not properly indicative of the system's performance on completely unknown words. Error rates have been reported to drop by a relative margin of up to 60% when training and test data contain words with overlapping roots (Demberg et al., 2007).

We have chosen to first sort the 23,031 words alphabetically, and then assign groups of 10 consecutive words at a time to the 10 cross-validation partitions. The same procedure was followed for the nested subdivision into development and training partitions for parameter optimisation. Manual inspection showed this strategy, described in (Kienappel and Kneser, 2001), to consign most related words to the same dataset. In comparison, the use of random selection gave more optimistic results, with an absolute increase in word accuracy of approximately 10% observed for G2P conversion.

The optimal number of graphemes and previously generated phonemes included in the G2P context, the best direction of the conversion, whether and which clusters should be included, and whether and how pruning should be applied, was determined by 10-fold cross-validation for each of the four dictionaries, leading to a total of 40 optimised parameter sets.

In agreement with the findings of others, the highest accuracy was obtained when including on average three graphemes to the left and four to the right of the focus grapheme, as well as the four phonemes most recently produced as part of the G2P process, while converting from right-to-left (Pagel et al., 1998, Torkkola, 1993). Note that, during training, the phonemes in question are taken from the dictionary pronunciation, while during G2P conversion the phonemes hypothesised by the decision tree are used. Since the G2P process is not perfect, this will lead to some mismatch between training and testing conditions.

The inclusion of questions relating to clusters of phonemes or graphemes, rather than merely individual phonemes or graphemes, was found to improve G2P performance in 27 of the 40 optimisations, measured in each case on the respective development set. The clusters were automatically determined using the greedy optimisation algorithm described in (Kienappel and Kneser, 2001). For each context position, this algorithm hierarchically groups graphemes or phonemes into clusters by means of repeated entropy-based binary splits. Optimal performance was achieved when using on average 26 graphemic and 39 phonemic clusters, although the precise number depends on the context position as well the development set in question. Examples of frequent grapheme clusters include *(c, k, q)*, *(a, o, u)* and *(s, z)*, while *(/aa/, /ae/, /eh/, /ao/, /y/)* and *(/th/, /dh/)* were common phoneme clusters.

Finally, decision tree pruning removes child nodes and their associated subtrees when these do not contribute at least a threshold improvement to the classification accuracy on a held-out set (Breiman et al., 1984). In 9 of the 40 optimisations, using 5% of the training data for pruning led to improved performance measured on the respective development set, but in most cases pruning was not used.

All experiments were performed using software developed at the University of Stellenbosch.

3.4. Experimental results for G2P

Using the optimal parameters determined in the previous section, G2P decision trees were trained for each dictionary individually. This allowed us to determine how accurately pronunciations could be produced based on training data drawn from the same accent. Together with the direct pronunciation comparison of Section 2.4, these results serve as a baseline for the accent conversion techniques that are presented in Sections 4 Phoneme-to-phoneme conversion (P2P), 5 Grapheme-and-phoneme-to-phoneme conversion (GP2P).

IPR2026-00272
Krisp Technologies EX1010 Page 214

The same 10 disjoint partitions of the 23,031-word data set described in Section 3.3 were used for training and testing the decision trees. Each of the 10 partitions was employed as a test set, while decision trees were trained using the remaining 9 partitions and the corresponding optimal parameters.

The generated pronunciations for each of the 10 test partitions were aligned with the pronunciations in the dictionary by dynamic programming. From these alignments the number of substitutions, insertions and deletions were determined. The phoneme accuracy was subsequently calculated using Eq. (1), and is presented in Table 6 for each of the four dictionaries. Word accuracy indicates the percentage of words for which the generated and correct pronunciations are identical. Where words have multiple pronunciations in the target accent, any of these are considered as correct. In each case the reported percentages are averages for the 10 cross-validation splits, while 95% confidence intervals were calculated using the bootstrap method described in (Bisani and Ney, 2004).

Table 6. Grapheme-to-phoneme (G2P) accuracies for the four dictionaries, with 95% confidence intervals and decision tree sizes.

Dictionary	Phoneme (%)	Word (%)	Size (nodes)
SAEDICT	90.67±0.54	59.64	18,064
BEEP	91.76±0.53	65.24	14,164
CMUDICT	91.26±0.54	63.29	15,328
PRONLEX	92.22±0.51	65.82	14,044

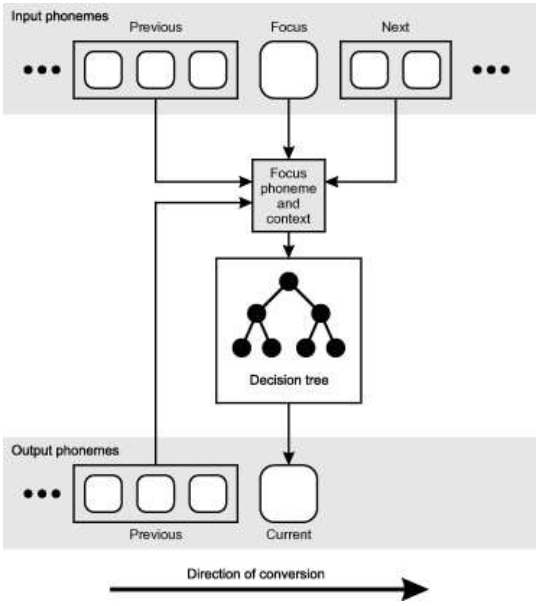
The results show little variation among the three accents. It is clear, however, that SAEDICT has the lowest G2P accuracy, while PRONLEX has the highest. This suggests that SAEDICT has the least regular relationship between its graphemes and phonemes. This is also reflected by the substantially larger size of SAEDICT decision trees. The results are comparable to those found by other researchers, where accuracies of approximately 91% are common (Black et al., 1998, Suontausta and Häkkinen, 2000).

4. Phoneme-to-phoneme conversion (P2P)

Grapheme-to-phoneme conversion can be used to estimate the pronunciations of new words, given a set of training pronunciations in the same accent or language. It is sometimes the case, however, that pronunciations for the desired words are already available, but in a different accent of the same language. For example, the pronunciations of many words currently not in SAEDICT are available in CMUDICT, PRONLEX or BEEP. We now address the question of whether such existing pronunciations in a different (source) accent can be used to estimate the pronunciation in the desired (target) accent. Since this process involves the conversion of one phoneme string (the source pronunciation) to another (the target pronunciation), it will be referred to as phoneme-to-phoneme (P2P) conversion.

P2P conversion has been used successfully to correct the output of a G2P system when applied to the derivation of pronunciations for proper names (van den Heuvel et al., 2007). We will not perform an explicit G2P conversion, but employ the same decision tree based algorithm used by this approach. In the case of P2P conversion, the phonemes of the source accent take the role of the “graphemes” and those of the target accent of the “phonemes”. This process is shown diagrammatically in Fig. 2.

IPR2026-00272
Krisp Technologies EX1010 Page 215



Download: Download full-size image

Fig. 2. Diagrammatic representation of phoneme-to-phoneme (P2P) conversion using decision trees.

Parameter optimisation for the P2P decision trees was carried out using a held-out development set within a 10-fold cross-validation framework, in an analogous fashion to the process described in Section 3.3. The optimal number of phonemes included in the P2P context, the best direction of the conversion, whether and which clusters should be included, and whether and how pruning should be applied, was determined for each of the 12 possible dictionary pairs, leading to a total of 120 optimised parameter sets.

It was found that the inclusion of on average two phonemes to the left and two to the right of the focus phoneme was optimal. Including previously generated output phonemes was beneficial in 65 of the 120 optimisations, while right-to-left conversion was best in 63 cases. The use of phoneme clusters improved performance in all 120 cases. In contrast to the G2P findings, decision tree pruning also led to improved performance, with the best results obtained when between 40% and 50% of the dictionary was used for training and the remainder for pruning. A notable difference between the G2P and P2P approaches is a higher incidence of insertions when aligning two pronunciations. While there are relatively few cases of a single grapheme corresponding to multiple phonemes, there are many more cases of a single source phoneme corresponding to multiple target phonemes. This led to the use by the decision tree of a larger number of double phonemes (treated as a single symbol). A total of 151 double phones was needed across all dictionary pairs, compared with just 63 for G2P. Double phonemes were determined automatically by iteratively aligning the training pronunciations of all words in a dictionary pair, finding pronunciations that could not be aligned, and adding the double phonemes required for the alignment to succeed. This process was also employed for G2P conversion, as set out in Section 3.1.

The P2P accent conversion accuracies are given in Table 7. The phoneme and word accuracies are again calculated as described in Section 3.4. The same 10-fold split into training and testing data was used in order to make Table 6, Table 7 directly comparable.

Table 7. Phoneme-to-phoneme (P2P) conversion accuracies between dictionary pairs, with 95% confidence intervals and decision tree sizes.

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
BEEP⇒SAEDICT	96.19±0.32	79.53	2708

IPR2026-00272
Krisp Technologies EX1010 Page 216

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
CMUDICT⇒SAEDICT	94.76±0.36	72.79	3471
PRONLEX⇒SAEDICT	95.62±0.34	76.69	2951
SAEDICT⇒BEEP	96.30±0.32	80.77	2794
CMUDICT⇒BEEP	95.43±0.36	76.89	3143
PRONLEX⇒BEEP	96.25±0.33	80.82	2776
SAEDICT⇒CMUDICT	94.54±0.39	72.46	3511
BEEP⇒CMUDICT	95.28±0.37	75.91	3308
PRONLEX⇒CMUDICT	97.64±0.26	87.20	1995
SAEDICT⇒PRONLEX	95.16±0.36	74.44	3221
BEEP⇒PRONLEX	95.66±0.35	77.31	3005
CMUDICT⇒PRONLEX	96.81±0.29	82.37	2219

It is clear that the pronunciations that can most accurately be derived from each other are CMUDICT and PRONLEX. This suggests that the apparent differences when directly comparing their pronunciations (Table 4) are at least in part due to systematic differences in the transcription conventions followed, rather than due to fundamental differences in the pronunciations.

Overall, the accuracy of the P2P process is considerably higher than that of G2P conversion. It is also evident that the decision trees used for P2P conversion are substantially smaller than those used for G2P conversion. A detailed discussion of these results is deferred to Section 7.

5. Grapheme-and-phoneme-to-phoneme conversion (GP2P)

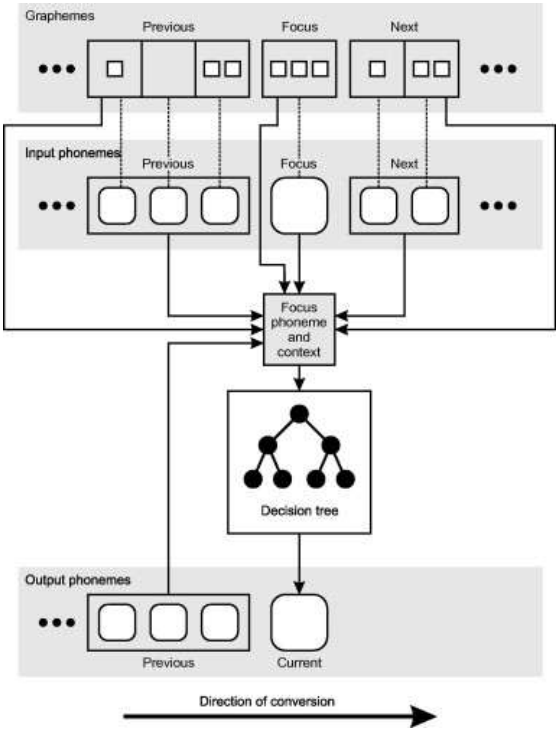
For some words, there may be information present in the orthography that is not represented phonetically. For example, words with different spellings can have identical pronunciations in one accent but not in another. In our case this occurs for “spokesmen” and “spokesman”, which are both pronounced /*s p ow k s m ax n*/ in BEEP, while “spokesmen” is pronounced /*s p ow k s m eh n*/ in SAEDICT.

The hypothesis that the orthography can provide information not already captured by the pronunciation is tested in the following by using both the graphemes and the phonemes in the source accent to derive the pronunciation in the target accent. This modification will be referred to as grapheme-and-phoneme-to-phoneme conversion (GP2P).

In order to employ the source graphemes as additional context for the decision tree, they are first aligned with the source phonemes, as described in Section 3.1. In this case the alignment was used to associate a sequence of zero or more graphemes with every source phoneme (rather than associating zero or more phonemes with each grapheme).

The decision tree question set was extended to include questions of the type “Does the grapheme sequence aligned with the phoneme contain the letter X?”. These questions could be asked of the focus phoneme or of any phoneme in the input context of the source accent. Fig. 3 illustrates the GP2P conversion process.

IPR2026-00272
Krisp Technologies EX1010 Page 217



Download: Download full-size image

Fig. 3. Diagrammatic representation of grapheme-and-phoneme-to-phoneme (GP2P) conversion using decision trees.

Optimal decision tree parameters were found using the same process described in Section 4. It was found that the parameters that produced the best performance for P2P systems did so also for GP2P systems. The only departure from this finding was that between 55% and 65% of the training data is best used for pruning, an absolute increase of approximately 5% relative to the P2P systems. Further refinements that were considered were allowing questions to be based on grapheme clusters obtained in an analogous fashion to the phoneme clusters, including the length of the grapheme sequence aligned with a source phoneme, and allowing questions to include fixed grapheme sequences (n-tuples). However, none of these were found to yield any further performance improvements during cross-validation. The results of the GP2P experiments are presented in Table 8, and will be discussed in relation to the performance of G2P and P2P conversion in Section 7.

Table 8. Grapheme-and-phoneme-to-phoneme (GP2P) conversion accuracies between dictionary pairs, with 95% confidence intervals and decision tree sizes.

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
BEEP⇒SAEDICT	96.55±0.31	81.58	2362
CMUDICT⇒SAEDICT	95.60±0.34	76.79	2831
PRONLEX⇒SAEDICT	96.05±0.33	78.99	2591
SAEDICT⇒BEEP	97.46±0.27	86.78	1880
CMUDICT⇒BEEP	96.58±0.32	82.28	2596
PRONLEX⇒BEEP	96.97±0.30	84.41	2078

IPR2026-00272
Krisp Technologies EX1010 Page 218

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
SAEDICT⇒CMUDICT	96.61±0.32	82.38	2313
BEEP⇒CMUDICT	96.85±0.31	83.96	2362
PRONLEX⇒CMUDICT	97.87±0.25	88.44	1613
SAEDICT⇒PRONLEX	97.26±0.27	84.60	2075
BEEP⇒PRONLEX	97.30±0.28	85.44	1943
CMUDICT⇒PRONLEX	97.75±0.25	87.22	1551

6. Multiple source dictionaries

A final question addressed by our work is whether the use of more than one source dictionary could further improve the performance of the GP2P approaches described in the previous section. As a specific example, we considered the two GP2P systems most accurately able to derive SSAE pronunciations. From Table 8, these are the BEEP to SAEDICT and PRONLEX to SAEDICT configurations. The GP2P algorithm was extended by allowing the decision trees to query the graphemes and phonemes in both BEEP and PRONLEX. To achieve this, BEEP was considered the primary and PRONLEX the additional source, and PRONLEX pronunciations were aligned with BEEP pronunciations as described in Section 4.

The decision tree question set was extended to include dictionary-specific questions, such as “Is the preceding phoneme in PRONLEX a plosive?”. This set of questions is a superset of those used for GP2P conversion with a single source dictionary, as was described in Section 5. Optimal decision tree parameters, again obtained by 10-fold cross-validation, were found to coincide with those that were optimal for the single source dictionary GP2P systems. The results, presented in Table 9, indicate that a small improvement in performance over the baseline BEEP to SAEDICT GP2P system was achieved, but that this improvement is not statistically significant.

Table 9. Grapheme-and-phoneme-to-phoneme (GP2P) conversion accuracies for SSAE when using one (BEEP) and two (BEEP and PRONLEX) source dictionaries, with 95% confidence intervals and decision tree sizes.

GP2P system	Phoneme (%)	Word (%)	Size (nodes)
BEEP⇒SAEDICT	96.55±0.31	81.58	2362
BEEP+PRONLEX⇒SAEDICT	96.62±0.31	81.84	2272

It therefore appears that further significant gains using multiple source dictionaries are not achievable for SSAE. However, it is possible that the result would be different for other English accents, or for other languages.

7. Discussion

When comparing Table 6, Table 7, it is clear that the pronunciation of a word in a new accent is in all cases substantially more accurately determined from its pronunciation in another accent by means of P2P conversion, than by means of G2P conversion within the same accent. Furthermore, Table 8 shows that the use of source graphemes in addition to source phonemes in GP2P conversion is rewarded with a further improvement in accuracy. This additional improvement is beyond the 95% confidence interval for all dictionary pairs except CMUDICT to PRONLEX, which is also the pair for which the highest P2P conversion accuracy was obtained.

IPR2026-00272
Krisp Technologies EX1010 Page 219

Converting between different accents also gives considerably more accurate results than merely using another accent's dictionary, as can be seen from Table 4. Substituting one English dictionary for another with no modifications gives at best a 60% and at worst a 30% word accuracy. For GP2P, however, even the poorest matching pair of dictionaries achieves a word accuracy above 76%.

Table 10 summarises the results presented in Table 6, Table 7, Table 8, and indicates the average performance across all four dictionaries. The results in the table show that, on average, P2P conversion leads to a 4.3% absolute improvement in phoneme accuracy over G2P conversion, which corresponds to a 51% relative decrease in error rate. Furthermore, the percentage of words with pronunciation errors drops by 14.6%, which is a 40% relative reduction in error. The move from P2P to GP2P conversion results in an additional average absolute phoneme accuracy improvement of 1.1%, which corresponds to a 26% relative reduction in the error rate. The corresponding percentage of incorrectly pronounced words drops by 5.5%, which is a relative improvement of 25%. Overall, GP2P conversion achieves relative reductions of 64% and 55% in the phoneme and word error rates, respectively, when compared with G2P conversion.

Table 10. Average accuracies achieved by the G2P, P2P and GP2P conversion approaches.

Dictionary	Phoneme (%)			Word (%)		
	G2P	P2P	GP2P	G2P	P2P	GP2P
⇒SAEDICT	90.67	95.52	96.07	59.64	76.34	79.12
⇒BEEP	91.76	96.00	97.00	65.24	79.50	84.49
⇒CMUDICT	91.26	95.82	97.11	63.29	78.52	84.93
⇒PRONLEX	92.22	95.88	97.44	65.82	78.04	85.75
Average	91.48	95.80	96.90	63.50	78.10	83.57

The introduction of source graphemes in GP2P conversion also led to a considerable reduction in the size of the decision trees. While P2P conversion employed 2925 nodes on average, the corresponding figure for GP2P falls to 2183. However, both P2P and GP2P trees are much smaller than G2P trees which contained 15,400 nodes on average.

Hence, when a word's pronunciation is not available, as is fairly common in less prominent accents, it is better to first search for it in a different pronunciation dictionary. If it is found, it should be converted to the target accent by means of GP2P or P2P techniques. G2P methods should be reserved as a last resort.

8. Summary and conclusions

This study has investigated methods that can be used to derive the pronunciation of a word in a new accent from its known pronunciation in another accent of the same language.

The performance of these techniques has been evaluated by applying them to three accents of English: RP, GenAm and South African English. In order to achieve this, four pronunciation dictionaries were employed: BEEP for RP, CMUDICT and PRONLEX for GenAm, and SAEDICT for South African English.

As a baseline for comparison, G2P techniques were first applied to the dictionaries individually, with comparable results in all four cases. The G2P algorithm was then modified to allow conversion of pronunciations from one accent to another, first using only the phonemes of the source accent (P2P) and then using both the phonemes and the graphemes (GP2P).

If we consider the specific case of South African English, for which pronunciation resources are limited, our experiments show that the pronunciations of 82% of new words obtained using the developed GP2P conversion technique are correct.

IPR2026-00272
Krisp Technologies EX1010 Page 220

In comparison, this figure is 60% when using conventional G2P conversion.

Our research can be extended by performing listening tests, as well as by performing Automatic Speech Recognition, to determine on a perceptual and an acoustic level how well the generated pronunciations match the desired accent. This would provide a further indication of the validity of our findings. Furthermore, it is possible that the techniques which we have presented would bear fruit if applied to other accents of the same language, or even to related languages, such as Dutch and Afrikaans.

Acknowledgements

This material is based on work supported financially by the South African National Research Foundation (NRF) under Grants FA2007022300015 and TTK2007041000010, and was executed using the High-Performance Computing (HPC) facility at Stellenbosch University.

Recommended articles

References

Andersen et al., 1996 Andersen, O., Kuhn, R., Lazaridès, A., Dalsgaard, P., Haas, J., Nöth, E., 1996. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In: Proc. ICSLP, Philadelphia, USA.

Google Scholar ↗

BEEP, 2009 BEEP, 2009. The British Example Pronunciation (BEEP) Dictionary. <http://svr-www.eng.cam.ac.uk/comp.speech/section1/lexical/beep.html> ↗ (accessed March 2009).

Google Scholar ↗

Bekker, 2009 Bekker, I., 2009. The vowels of South African English. Ph.D. Thesis, North-West University, Potchefstroom, South Africa.

Google Scholar ↗

Bisani and Ney, 2004 Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proc. ICASSP, Montreal, Canada.

Google Scholar ↗

Black et al., 1998 Black, A., Lenzo, K., Pagel, V., 1998. Issues in building general letter to sound rules. In: Proc. ESCA Speech Synthesis Workshop, Jenolan Caves, Australia.

Google Scholar ↗

Bowerman, 2004 S. Bowerman

White South African English: phonology

E. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, C. Upton (Eds.), A Handbook of Varieties of English, Vol. 1, Mouton de Gruyter, Berlin (2004), pp. 931-942

Google Scholar ↗

Breiman et al., 1984 L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone

Classification and Regression Trees

Wadsworth and Brooks, Pacific Grove (1984)

Google Scholar ↗

CMU, 2009 CMU, 2009. Carnegie Mellon University Pronouncing Dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> ↗ (accessed March 2009).

Google Scholar ↗

IPR2026-00272
Krisp Technologies EX1010 Page 221

Daelemans and van den Bosch, 1994 W. Daelemans, A. van den Bosch
Language-independent data-oriented grapheme-to-phoneme conversion
J. van Santen, R. Sproat, J. Olive, J. Hirschberg (Eds.), Progress in Speech Synthesis, Springer, New York (1994), pp. 77-89
Google Scholar ↗

Daelemans et al., 1999 W. Daelemans, A. van den Bosch, J. Zavrel
Forgetting exceptions is harmful in language learning
Machine Learn., 34 (1999), pp. 11-41
View in Scopus ↗ Google Scholar ↗

Damper et al., 1999 R. Damper, Y. Marchand, M. Adamson, K. Gustafson
Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches
Comput. Speech Lang., 13 (2) (1999), pp. 155-176
View PDF View article View in Scopus ↗ Google Scholar ↗

Damper et al., 2004 Damper, R., Marchand, Y., Marsters, J., Bazin, A., 2004. Aligning letters and phonemes for speech synthesis. In: Proc. 5th ISCA Speech Synthesis Workshop, Pittsburg, USA.
Google Scholar ↗

Demberg et al., 2007 Demberg, V., Schmid, H., Möhler, G., 2007. Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion. In: Proc. ACL, Prague, Czech Republic.
Google Scholar ↗

Han and Chen, 2004 Han, K.-S., Chen, G.-L., 2004. Letter-to-sound for small-footprint multilingual TTS engine. In: Proc. ICSLP, Jeju, Korea.
Google Scholar ↗

Humphries et al., 2001 Humphries, J., Woodland, P., Pearce, D., 2001. Using accent-specific pronunciation modelling for robust speech recognition. In: Proc. ICASSP, Salt Lake City.
Google Scholar ↗

Kienappel and Kneser, 2001 Kienappel, A.K., Kneser, R., 2001. Designing very compact decision trees for grapheme-to-phoneme transcription. In: Proc. Eurospeech, Aalborg, Denmark.
Google Scholar ↗

LDC, 2009 LDC, 2009. COMLEX English pronouncing lexicon. <http://www.ldc.upenn.edu> ↗ (accessed March 2009).
Google Scholar ↗

Loots and Niesler, 2009 Loots, L., Niesler, T., 2009. Data-driven phonetic comparison and conversion between South African, British and American English pronunciations. In: Proc. Interspeech, Brighton, UK.
Google Scholar ↗

Marchand and Damper, 2000 Y. Marchand, R. Damper
A multistrategy approach to improving pronunciation by analogy
Comput. Linguist., 26 (2) (2000), pp. 195-219
Google Scholar ↗

Niesler et al., 2005 T. Niesler, P. Louw, J. Roux
Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases
Southern African Linguist. Appl. Lang. Studies, 23 (4) (2005), pp. 459-474
Crossref ↗ View in Scopus ↗ Google Scholar ↗

IPR2026-00272
Krisp Technologies EX1010 Page 222

Pagel et al., 1998 Pagel, V., Lenzo, K., Black, A., 1998. Letter to sound rules for accented lexicon compression. In: Proc. ICSLP, Sydney, Ausralia.

Google Scholar ↗

Rabiner and Juang, 1993 L. Rabiner, B. Juang

Fundamentals of Speech Recognition

Prentice Hall, New Jersey, USA (1993)

Google Scholar ↗

Suontausta and Häkkinen, 2000 Suontausta, J., Häkkinen, J., 2000. Decision tree based text-to-phoneme mapping for speech recognition. In: Proc. ICSLP, Beijing, China.

Google Scholar ↗

Taylor, 2005 Taylor, P., 2005. Hidden Markov models for grapheme to phoneme conversion. In: Proc. Interspeech, Lisbon, Portugal.

Google Scholar ↗

Torkkola, 1993 Torkkola, K., 1993. An efficient way to learn grapheme-to-phoneme rules automatically. In: Proc. ICASSP, Minneapolis, USA.

Google Scholar ↗

van den Heuvel et al., 2007 van den Heuvel, H., Martens, J.-P., Konings, N., 2007. G2P conversion of names. what can we do (better)? In: Proc. Interspeech, Antwerp, Belgium.

Google Scholar ↗

Wan et al., 2008 Wan, V., Dines, J., El Hannani, A., Hain, T., 2008. Bob: a lexicon and pronunciation dictionary generator. In: Proc. Spoken Language Technology Workshop, Goa, India.

Google Scholar ↗

Webster and Braunschweiler, 2008 Webster, G., Braunschweiler, N., 2008. An evaluation of non-standard features for grapheme-to-phoneme conversion. In: Proc. Interspeech, Brisbane, Australia.

Google Scholar ↗

Wells, 1982 J. Wells

Accents of English

Cambridge University Press, Cambridge (1982)

Google Scholar ↗

Cited by (21)

Controllable Accented Text-to-Speech Synthesis With Fine and Coarse-Grained Intensity Rendering ↗

2024, IEEE ACM Transactions on Audio Speech and Language Processing

Non-Autoregressive End-to-End Neural Modeling for Automatic Pronunciation Error Detection ↗

2023, Applied Sciences Switzerland

Arabic Automatic Speech Recognition: A Systematic Literature Review ↗

2022, Applied Sciences Switzerland

Automatic speech recognition system for Tunisian dialect ↗

IPR2026-00272
Krisp Technologies EX1010 Page 223

2018, Language Resources and Evaluation

Arabic Automatic Speech Recognition Enhancement ↗

2017, Proceedings 2017 Palestinian International Conference on Information and Communication Technology Picict 2017

Radio-browsing for developmental monitoring in Uganda ↗

2017, ICASSP IEEE International Conference on Acoustics Speech and Signal Processing Proceedings



View all citing articles on Scopus ↗

☆ Extended version of a paper presented at Interspeech 2009, Brighton, UK (Loots and Niesler, 2009).

View Abstract

Copyright © 2010 Elsevier B.V. All rights reserved.

ELSEVIER

All content on this site: Copyright © 2025 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the relevant licensing terms apply.

RELX™

IPR2026-00272
Krisp Technologies EX1010 Page 224

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 75–84
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Automatic conversion between pronunciations of different English accents ☆

Linsen Loots, Thomas Niesler *

Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

Received 11 January 2010; received in revised form 6 July 2010; accepted 25 July 2010

Abstract

We describe the application of decision trees to the automatic conversion of pronunciations between American, British and South African English accents. The resulting phoneme-to-phoneme (P2P) conversion technique derives the pronunciation of a word in a new target accent by taking advantage of its existing available pronunciation in a different source accent. We find that it is substantially more accurate to derive pronunciations in this way than directly from the orthography and available target accent pronunciations using more conventional grapheme-to-phoneme (G2P) conversion. Furthermore, by including both the graphemes and the phonemes of the source accent, grapheme-and-phoneme-to-phoneme (GP2P) conversion delivers additional increases in accuracy in relation to P2P. These findings are particularly important for less-resourced varieties of English, for which extensive manually-prepared pronunciation dictionaries are not available. By means of the P2P and GP2P approaches, the pronunciations of new words can be obtained with better accuracy than is possible using G2P methods.
© 2010 Elsevier B.V. All rights reserved.

Keywords: English accents; Pronunciation modelling; G2P; P2P; GP2P; Decision trees; South African English

1. Introduction

The pronunciation dictionary is a key component of an Automatic Speech Recognition (ASR) as well as a text to speech (TTS) system. However, the development of this dictionary involves a great deal of time and effort by linguistic experts, a process that may be prohibitively expensive for under-resourced languages and accents. Nevertheless phonetic transcriptions in the appropriate accent can improve both ASR accuracy (Humphries et al., 2001) and TTS quality (Damper et al., 1999).

We consider the scenario where the pronunciations of words not in the dictionary are sought, and where these pronunciations are already available in a different accent of the language in question. It will be shown that, in this situation, data-driven techniques similar to those used in grapheme-to-phoneme (G2P) conversion can be used to obtain the desired pronunciation from the existing pronunciation, instead of from the graphemes. This amounts to a conversion between the two accents.

In our experimental evaluation, we focus on General American (GenAm), Received Pronunciation (RP) and Standard South African English (SSAE). The first two are commonly-used and widely-studied reference accents for American and British English, respectively (Wells, 1982). In South Africa, a number of different English accents are prevalent. The accent generally used by first-language speakers is commonly referred to as Standard South African English and has been described as fairly similar to RP (Bowerman, 2004; Bekker, 2009). An important distinguishing factor of SSAE is the characteristic split of Wells' *KIT* vowel. This vowel appears in words like *kit*, *thin* and *sit*, and in SSAE has two clear allophonic variations: one close and one more central (Wells, 1982). While these words

☆ Extended version of a paper presented at Interspeech 2009, Brighton, UK (Loots and Niesler, 2009).

* Corresponding author. Tel.: +27 21 808 4118.
E-mail addresses: linsen@sun.ac.za (L. Loots), trn@sun.ac.za (T. Niesler).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.07.006

contain a single vowel sound in most English accents, they are pronounced using two different vowel sounds in SSAE.

From our perspective the objective of accent conversion is to determine whether British or American pronunciations can be used to complement our still fairly small SSAE pronunciation dictionary, and how this can be done best. However, we believe that the techniques we consider are generally applicable to any accent pair.

The following section will describe the data upon which our experimental evaluation is based. Sections 3–5 then describe three alternative techniques that can be used to obtain pronunciations of new words. Finally, Section 7 discusses the results and Section 8 concludes.

2. Dictionaries

Four dictionaries were used to represent the three accents which we consider, and are described shortly in the following.

CMUDICT and PRONLEX were used to represent GenAm pronunciations. CMUDICT was developed at Carnegie Mellon University (CMU, 2009). This dictionary contains words drawn from a variety of sources, including an initial subset of over 20,000 words developed by hand and extensively verified at CMU. Other sources of pronunciations for CMUDICT include the Shoup and the Dragon dictionaries. PRONLEX is the COMLEX English Pronouncing Lexicon, and was hand-transcribed at the Linguistic Data Consortium (LDC, 2009).

The BEEP dictionary was used to represent RP and is drawn primarily from the Medical Research Council (MRC) Psycholinguistic Database and the Computer Usable Oxford Advanced Learners Dictionary (BEEP, 2009). In addition, it contains pronunciations provided by sources at Durham and Oxford Universities.

SSAE pronunciations were obtained from SAEDICT, a pronunciation dictionary under development at Stellenbosch University. All pronunciations in SAEDICT were transcribed by the same linguistic specialist to ensure consistency. Transcriptions were chosen to reflect commonly accepted SSAE pronunciations.

Table 1 describes each dictionary in terms of the number of words it contains, the total number of pronunciations and the average number of pronunciations per word (P/W). The number of pronunciations and P/W is also given for a set of 23,031 words common to all four dictionaries, as will be described in Section 2.2.

2.1. Phoneset

ARPABET (Rabiner and Juang, 1993) was chosen as the common phoneset in which to analyse the four dictionaries. Both BEEP and CMUDICT already use ARPABET, while PRONLEX uses an ARPABET short-hand that can easily be converted to standard ARPABET. SAEDICT uses a phoneset, based on IPA, developed to describe the languages of Southern Africa (Niesler et al., 2005). This was converted to ARPABET by means of a mapping based on the closest IPA symbol.

A small number of phonemes was found to be present in only a single dictionary, and these were mapped to alternative phonemes as indicated in Table 2. This normalisation ensures that the phonesets used by the four dictionaries match exactly.

2.2. Wordlist

In order to compare pronunciations, the set of words common to all four dictionaries was determined. Before extracting this set, all words were automatically converted to standard UK spelling using tools developed at the University of Sheffield (Wan et al., 2008). The set of common words contained 23,031 entries.

There are a number of words in each dictionary with multiple pronunciations, either homographs (words with the same spelling but different pronunciations), or words with alternate pronunciations. The number of pronunciations vary, but all four dictionaries contain an average of between 1.11 and 1.22 pronunciations per word, as reflected in Table 1. Of the 23,031 words common to all four dictionaries, 9344 have multiple pronunciations in at least one dictionary.

Table 2
Phoneme mappings used to achieve a common phoneset.

Dictionary	Mapping
BEEP	*/oh/ ⇒ /aa/* */ia/ ⇒ /ih ax/* */ea/ ⇒ /eh ax/* */ua/ ⇒ /uh ax/*
PRONLEX	*/wh/ ⇒ /w/* */en/ ⇒ /ax n/*

Table 1
Number of entries in the four pronunciation dictionaries.

Dictionary	Entire dictionary			Common set		
	Words	Prons	P/W	Words	Prons	P/W
SAEDICT	30,390	36,954	1.22	23,031	28,087	1.22
BEEP	237,592	256,998	1.08	23,031	25,576	1.11
CMUDICT	123,292	133,694	1.08	23,031	26,771	1.16
PRONLEX	90,919	99,978	1.11	23,031	25,471	1.11

IPR2026-00272
Krisp Technologies EX1010 Page 226

2.3. Pronunciation alignments

The pronunciations of corresponding words in the three accents were aligned using a dynamic programming algorithm that will be described in Section 3.1. This was done so that accents could be compared directly, as will be described in the following section. However, the alignments will also be used to obtain the one-to-one correspondences between phonemes in different accents. These correspondences are needed to train the decision trees that will be responsible for inter-accent conversions (described in Sections 4 and 5).

As an example, the alignment between SSAE and RP pronunciations of the word "reactions" is shown in Table 3. Nulls have been introduced where a phoneme has no counterpart in the other string (i.e. insertions and deletions). In this case the /*ax*/ in the SSAE pronunciation is deleted in RP.

Words with multiple pronunciations pose a problem both when comparing two dictionaries, and when subsequently using the aligned pronunciations to train decision trees for accent conversion. Our solution has been to use only the single pair of pronunciations per word (one from each dictionary) that gives the best alignment. An alternative approach is to consider all the possible pairings of pronunciations between the two dictionaries. However, preliminary experiments indicated that this did not influence the relative accuracies obtained for the various accents. Hence, the simpler approach of using the best-aligned pair was chosen for all further experiments.

2.4. Direct pronunciation comparison

In order to obtain a baseline for inter-accent similarity against which accent conversion can later be compared, a direct phonetic comparison of the different accents was performed. A summary of the results obtained for the 23,031 common words described in Section 2.2 is given in Table 4. The phoneme accuracy was calculated according to Eq. (1), where N_c, N_i and N_t are the numbers of correct, inserted and total phonemes, respectively. This accuracy is given for all phonemes, and separately for vowels and for consonants. The last column indicates the percentage of words that have identical pronunciations.

$$Acc = \frac{N_c - N_i}{N_t}. \quad (1)$$

It is clear from Table 4 that SAEDICT and BEEP match most closely, while the poorest correspondence is found between CMUDICT and SAEDICT. On average, the American pronunciations are closer to RP than to SSAE.

Table 3
Two aligned pronunciations of the word "reactions".

SAEDICT	*r*	*ih*	*ae*	*k*	*sh*	*ax*	*n*	*s*
BEEP	*r*	*ih*	*ae*	*k*	*sh*	–	*n*	*z*

Table 4
Accuracies (%) of direct pronunciation alignments between dictionaries.

Dictionary pair	Phoneme	Vowel	Cons	Word
BEEP ⟺ SAEDICT	92.1	84.5	96.9	59.6
CMUDICT ⟺ SAEDICT	81.6	59.8	95.2	29.4
PRONLEX ⟺ SAEDICT	87.3	72.8	96.2	43.0
CMUDICT ⟺ BEEP	84.7	67.7	94.6	37.8
PRONLEX ⟺ BEEP	89.4	79.0	95.6	50.6
PRONLEX ⟺ CMUDICT	89.0	73.2	98.6	50.7

Furthermore, vowels differ more strongly than consonants between accents. While at least 94.6% of consonants matched between any pair of dictionaries, this figure varied between 59.8% and 84.5% for vowels. A more detailed analysis of the phonetic differences between the three accents can be found in (Loots and Niesler, 2009).

3. Grapheme-to-phoneme conversion (G2P)

G2P conversion is the process of determining a word's pronunciation from its spelling. Several data-driven G2P methods have been suggested in the literature, including decision trees (Black et al., 1998; Suontausta and Häkkinen, 2000; Kienappel and Kneser, 2001; Webster and Braunschweiler, 2008), HMMs (Taylor, 2005), pronunciation by analogy (Marchand and Damper, 2000) and memory-based learning (Daelemans et al., 1999). Decision trees have been shown to yield competitive accuracy in G2P conversion (Han and Chen, 2004) and to be a highly competitive approach to the generation of accent-specific pronunciation variants for ASR (Humphries et al., 2001). We have chosen decision trees as the method with which we will investigate the success of accent conversion.

Although this paper does not focus on G2P systems, the methods used for accent conversion in Sections 4 and 5 are based on analogous techniques. Hence, G2P conversion by means of decision trees will be reviewed first, and used as a further baseline against which the later systems can be measured.

For G2P conversion, the graphemes and their context form the input of the decision tree classifier, and the phonemes the output class. Pronunciations are generated by sequentially passing graphemes and their context through the tree, assigning each to a phoneme class, and then concatenating these output phonemes. This process is illustrated in Fig. 1.

3.1. Aligning graphemes and phonemes

In order to sequentially generate an output phoneme for each grapheme, the decision tree requires a one-to-one correspondence between the graphemes and the phonemes. An example of such an alignment is shown in Table 5.

Discrepancies in length and alignment between the graphemes and phonemes are dealt with by inserting nulls into the phoneme string (for letters not corresponding to a phoneme), and by combining pairs of phonemes into

IPR2026-00272
Krisp Technologies EX1010 Page 227

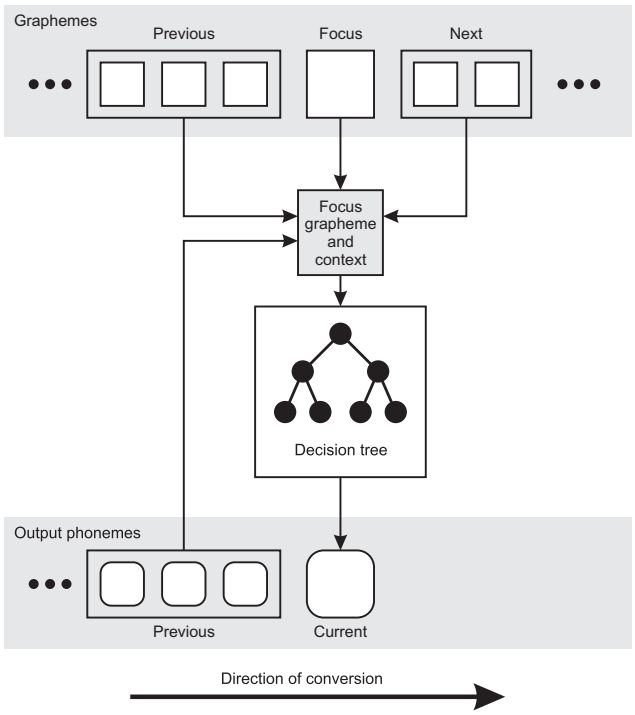


Fig. 1. Diagrammatic representation of grapheme-to-phoneme (G2P) conversion using decision trees.

double phonemes for the relatively few cases where one grapheme corresponds to two phonemes (Damper et al., 2004; Suontausta and Häkkinen, 2000).

The primary method of aligning graphemic and phonetic strings uses the Viterbi algorithm (Daelemans and van den Bosch, 1994; Andersen et al., 1996; Pagel et al., 1998). The process begins by estimating $P(G,P)$, the unigram probability that a given grapheme is associated with a given phoneme. This is estimated using those words that have graphemic and phonetic strings of equal length (Andersen et al., 1996). Optimal alignments based on these estimated probabilities are then generated by means of dynamic programming for all words in the corpus. The newly-aligned words are then used to recalculate the unigram probabilities, and the process is repeated until the distributions converge (Pagel et al., 1998; Webster and Braunschweiler, 2008).

Double phonemes can be determined automatically by iteratively repeating this process. First, suitable double phonemes are identified from unsuccessful alignments between grapheme and phoneme strings. Alignment is subsequently re-attempted, and the process repeated until all words align with their pronunciations.

Table 5
An example of grapheme-to-phoneme (G2P) alignment for the word "extreme".

Grapheme	e	x	t	r	e	m	e
Phoneme	*eh*	*k s*	*t*	*r*	*iy*	*m*	–

3.2. Decision trees for G2P

We have employed deterministic binary decision trees for G2P conversion, for which each node is associated with a true/false question regarding the input grapheme and its context. The tree is traversed from the root by recursively using the answer of each node's question to determine which child node to choose. Each leaf node is associated with an output phoneme, which constitutes the classification result (Breiman et al., 1984). This process produces a single pronunciation for the presented orthography.

Decision trees are grown recursively. For each new node the available training data are split according to all possible questions. The question which results in the greatest entropy gain is then chosen for that node (Suontausta and Häkkinen, 2000). For a node t with entropy $i(t)$, children t_L and t_R with respective entropies $i(t_L)$ and $i(t_R)$, and the proportion of node t's data associated with each child given by p_L and p_R such that both are non-negative and sum to unity, the entropy gain is (Breiman et al., 1984):

$$\Delta i = i(t) - p_L i(t_L) - p_R i(t_R). \tag{2}$$

When choosing the question with the largest Δi, $i(t)$ can be omitted from the calculation with no loss of generality as it remains constant for all questions at a given node. Furthermore p_k can be approximated by the relative frequency N_{t_k}/N_t for a child node t_k, while $1/N_t$ is constant for all questions.

$$\Delta i = \sum_{\forall children\ k} \sum_{\forall phonemes\ p} N_{t_k,p} \log\left(\frac{N_{t_k,p}}{N_{t_k}}\right). \tag{3}$$

It is therefore sufficient to maximise Eq. (3) in order to determine the optimal question for a new decision tree node (Kienappel and Kneser, 2001).

3.3. Decision tree training and testing

A G2P decision tree includes many parameters, the most important of which are the number of graphemes and previously generated phonemes included in the context, the direction of the conversion, the use of clusters, and the use of pruning (Webster and Braunschweiler, 2008).

To obtain values for these parameters in our G2P experiments, the phoneme accuracy of the decision tree was optimised on a held-out development set within a 10-fold cross-validation framework. This was achieved by dividing the 23,031 common words described in Section 2.2 into 10 non-overlapping and approximately equally-sized partitions. Reserving each partition in turn for later use as a test set, the remaining 9 partitions (corresponding to 90% of the data) were again subdivided into 10 equal partitions. The first of these sub-partitions served as a development set, and the remaining 9 as a training set for parameter optimisation. Decision trees were trained on this training set for a range of parameter values, and those that lead to the highest phoneme accuracy measured on the

IPR2026-00272
Krisp Technologies EX1010 Page 228

associated development set were identified as optimal. This process was repeated for each of the 10 partitions of the 23,031-word data set. In general, different optimal parameter values were obtained in each case.

The data partitioning must be carefully considered, as training, development and testing datasets should be properly representative of the dictionary (and thus the language) as a whole. Furthermore, it is important that there is not too much overlap between the data partitions, as might occur between words with the same root but different affixes. This should be avoided because it can cause the test or development data to be unrealistically similar to the training data, leading to results that are not properly indicative of the system's performance on completely unknown words. Error rates have been reported to drop by a relative margin of up to 60% when training and test data contain words with overlapping roots (Demberg et al., 2007).

We have chosen to first sort the 23,031 words alphabetically, and then assign groups of 10 consecutive words at a time to the 10 cross-validation partitions. The same procedure was followed for the nested subdivision into development and training partitions for parameter optimisation. Manual inspection showed this strategy, described in (Kienappel and Kneser, 2001), to consign most related words to the same dataset. In comparison, the use of random selection gave more optimistic results, with an absolute increase in word accuracy of approximately 10% observed for G2P conversion.

The optimal number of graphemes and previously generated phonemes included in the G2P context, the best direction of the conversion, whether and which clusters should be included, and whether and how pruning should be applied, was determined by 10-fold cross-validation for each of the four dictionaries, leading to a total of 40 optimised parameter sets.

In agreement with the findings of others, the highest accuracy was obtained when including on average three graphemes to the left and four to the right of the focus grapheme, as well as the four phonemes most recently produced as part of the G2P process, while converting from right-to-left (Pagel et al., 1998; Torkkola, 1993). Note that, during training, the phonemes in question are taken from the dictionary pronunciation, while during G2P conversion the phonemes hypothesised by the decision tree are used. Since the G2P process is not perfect, this will lead to some mismatch between training and testing conditions.

The inclusion of questions relating to clusters of phonemes or graphemes, rather than merely individual phonemes or graphemes, was found to improve G2P performance in 27 of the 40 optimisations, measured in each case on the respective development set. The clusters were automatically determined using the greedy optimisation algorithm described in (Kienappel and Kneser, 2001). For each context position, this algorithm hierarchically groups graphemes or phonemes into clusters by means of repeated entropy-based binary splits. Optimal performance was achieved when using on average 26 graphemic and 39 phonemic clusters, although the precise number depends on the context position as well the development set in question. Examples of frequent grapheme clusters include *(c, k, q)*, *(a, o, u)* and *(s, z)*, while *(/aa/, /ae/, /eh/, /ao/, /y/)* and *(/th/, /dh/)* were common phoneme clusters.

Finally, decision tree pruning removes child nodes and their associated subtrees when these do not contribute at least a threshold improvement to the classification accuracy on a held-out set (Breiman et al., 1984). In 9 of the 40 optimisations, using 5% of the training data for pruning led to improved performance measured on the respective development set, but in most cases pruning was not used.

All experiments were performed using software developed at the University of Stellenbosch.

3.4. Experimental results for G2P

Using the optimal parameters determined in the previous section, G2P decision trees were trained for each dictionary individually. This allowed us to determine how accurately pronunciations could be produced based on training data drawn from the same accent. Together with the direct pronunciation comparison of Section 2.4, these results serve as a baseline for the accent conversion techniques that are presented in Sections 4 and 5.

The same 10 disjoint partitions of the 23,031-word data set described in Section 3.3 were used for training and testing the decision trees. Each of the 10 partitions was employed as a test set, while decision trees were trained using the remaining 9 partitions and the corresponding optimal parameters.

The generated pronunciations for each of the 10 test partitions were aligned with the pronunciations in the dictionary by dynamic programming. From these alignments the number of substitutions, insertions and deletions were determined. The phoneme accuracy was subsequently calculated using Eq. (1), and is presented in Table 6 for each of the four dictionaries. Word accuracy indicates the percentage of words for which the generated and correct pronunciations are identical. Where words have multiple pronunciations in the target accent, any of these are considered as correct. In each case the reported percentages are averages for the 10 cross-validation splits, while 95% confidence intervals were calculated using the bootstrap method described in (Bisani and Ney, 2004).

Table 6
Grapheme-to-phoneme (G2P) accuracies for the four dictionaries, with 95% confidence intervals and decision tree sizes.

Dictionary	Phoneme (%)	Word (%)	Size (nodes)
SAEDICT	90.67 ± 0.54	59.64	18,064
BEEP	91.76 ± 0.53	65.24	14,164
CMUDICT	91.26 ± 0.54	63.29	15,328
PRONLEX	92.22 ± 0.51	65.82	14,044

IPR2026-00272
Krisp Technologies EX1010 Page 229

The results show little variation among the three accents. It is clear, however, that SAEDICT has the lowest G2P accuracy, while PRONLEX has the highest. This suggests that SAEDICT has the least regular relationship between its graphemes and phonemes. This is also reflected by the substantially larger size of SAEDICT decision trees. The results are comparable to those found by other researchers, where accuracies of approximately 91% are common (Black et al., 1998; Suontausta and Häkkinen, 2000).

4. Phoneme-to-phoneme conversion (P2P)

Grapheme-to-phoneme conversion can be used to estimate the pronunciations of new words, given a set of training pronunciations in the same accent or language. It is sometimes the case, however, that pronunciations for the desired words are already available, but in a different accent of the same language. For example, the pronunciations of many words currently not in SAEDICT are available in CMUDICT, PRONLEX or BEEP. We now address the question of whether such existing pronunciations in a different (source) accent can be used to estimate the pronunciation in the desired (target) accent. Since this process involves the conversion of one phoneme string (the source pronunciation) to another (the target pronunciation), it will be referred to as phoneme-to-phoneme (P2P) conversion.

P2P conversion has been used successfully to correct the output of a G2P system when applied to the derivation of pronunciations for proper names (van den Heuvel et al., 2007). We will not perform an explicit G2P conversion, but employ the same decision tree based algorithm used by this approach. In the case of P2P conversion, the phonemes of the source accent take the role of the "graphemes" and those of the target accent of the "phonemes". This process is shown diagrammatically in Fig. 2.

Parameter optimisation for the P2P decision trees was carried out using a held-out development set within a 10-fold cross-validation framework, in an analogous fashion to the process described in Section 3.3. The optimal number of phonemes included in the P2P context, the best direction of the conversion, whether and which clusters should be included, and whether and how pruning should be applied, was determined for each of the 12 possible dictionary pairs, leading to a total of 120 optimised parameter sets.

It was found that the inclusion of on average two phonemes to the left and two to the right of the focus phoneme was optimal. Including previously generated output phonemes was beneficial in 65 of the 120 optimisations, while right-to-left conversion was best in 63 cases. The use of phoneme clusters improved performance in all 120 cases. In contrast to the G2P findings, decision tree pruning also led to improved performance, with the best results obtained when between 40% and 50% of the dictionary was used for training and the remainder for pruning. A

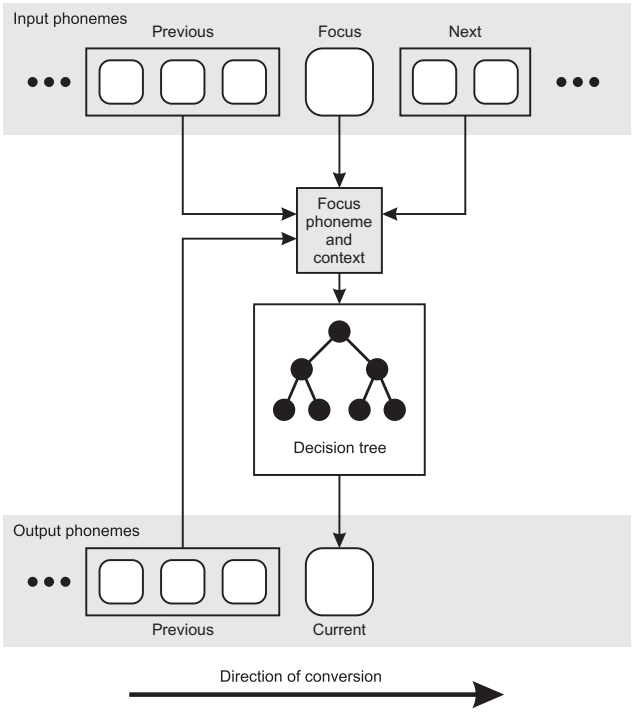


Fig. 2. Diagrammatic representation of phoneme-to-phoneme (P2P) conversion using decision trees.

notable difference between the G2P and P2P approaches is a higher incidence of insertions when aligning two pronunciations. While there are relatively few cases of a single grapheme corresponding to multiple phonemes, there are many more cases of a single source phoneme corresponding to multiple target phonemes. This led to the use by the decision tree of a larger number of double phonemes (treated as a single symbol). A total of 151 double phones was needed across all dictionary pairs, compared with just 63 for G2P. Double phonemes were determined automatically by iteratively aligning the training pronunciations of all words in a dictionary pair, finding pronunciations that could not be aligned, and adding the double phonemes required for the alignment to succeed. This process was also employed for G2P conversion, as set out in Section 3.1.

The P2P accent conversion accuracies are given in Table 7. The phoneme and word accuracies are again calculated as described in Section 3.4. The same 10-fold split into training and testing data was used in order to make Tables 6 and 7 directly comparable.

It is clear that the pronunciations that can most accurately be derived from each other are CMUDICT and PRONLEX. This suggests that the apparent differences when directly comparing their pronunciations (Table 4) are at least in part due to systematic differences in the transcription conventions followed, rather than due to fundamental differences in the pronunciations.

Overall, the accuracy of the P2P process is considerably higher than that of G2P conversion. It is also evident that

IPR2026-00272
Krisp Technologies EX1010 Page 230

Table 7
Phoneme-to-phoneme (P2P) conversion accuracies between dictionary pairs, with 95% confidence intervals and decision tree sizes.

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
BEEP ⇒ SAEDICT	96.19 ± 0.32	79.53	2708
CMUDICT ⇒ SAEDICT	94.76 ± 0.36	72.79	3471
PRONLEX ⇒ SAEDICT	95.62 ± 0.34	76.69	2951
SAEDICT ⇒ BEEP	96.30 ± 0.32	80.77	2794
CMUDICT ⇒ BEEP	95.43 ± 0.36	76.89	3143
PRONLEX ⇒ BEEP	96.25 ± 0.33	80.82	2776
SAEDICT ⇒ CMUDICT	94.54 ± 0.39	72.46	3511
BEEP ⇒ CMUDICT	95.28 ± 0.37	75.91	3308
PRONLEX ⇒ CMUDICT	97.64 ± 0.26	87.20	1995
SAEDICT ⇒ PRONLEX	95.16 ± 0.36	74.44	3221
BEEP ⇒ PRONLEX	95.66 ± 0.35	77.31	3005
CMUDICT ⇒ PRONLEX	96.81 ± 0.29	82.37	2219

the decision trees used for P2P conversion are substantially smaller than those used for G2P conversion. A detailed discussion of these results is deferred to Section 7.

5. Grapheme-and-phoneme-to-phoneme conversion (GP2P)

For some words, there may be information present in the orthography that is not represented phonetically. For example, words with different spellings can have identical pronunciations in one accent but not in another. In our case this occurs for "spokesmen" and "spokesman", which are both pronounced */s p ow k s m ax n/* in BEEP, while "spokesmen" is pronounced */s p ow k s m eh n/* in SAEDICT.

The hypothesis that the orthography can provide information not already captured by the pronunciation is tested in the following by using both the graphemes and the phonemes in the source accent to derive the pronunciation in the target accent. This modification will be referred to as grapheme-and-phoneme-to-phoneme conversion (GP2P).

In order to employ the source graphemes as additional context for the decision tree, they are first aligned with the source phonemes, as described in Section 3.1. In this case the alignment was used to associate a sequence of zero or more graphemes with every source phoneme (rather than associating zero or more phonemes with each grapheme).

The decision tree question set was extended to include questions of the type "Does the grapheme sequence aligned with the phoneme contain the letter X?". These questions could be asked of the focus phoneme or of any phoneme in the input context of the source accent. Fig. 3 illustrates the GP2P conversion process.

Optimal decision tree parameters were found using the same process described in Section 4. It was found that the parameters that produced the best performance for P2P systems did so also for GP2P systems. The only departure from this finding was that between 55% and 65% of the

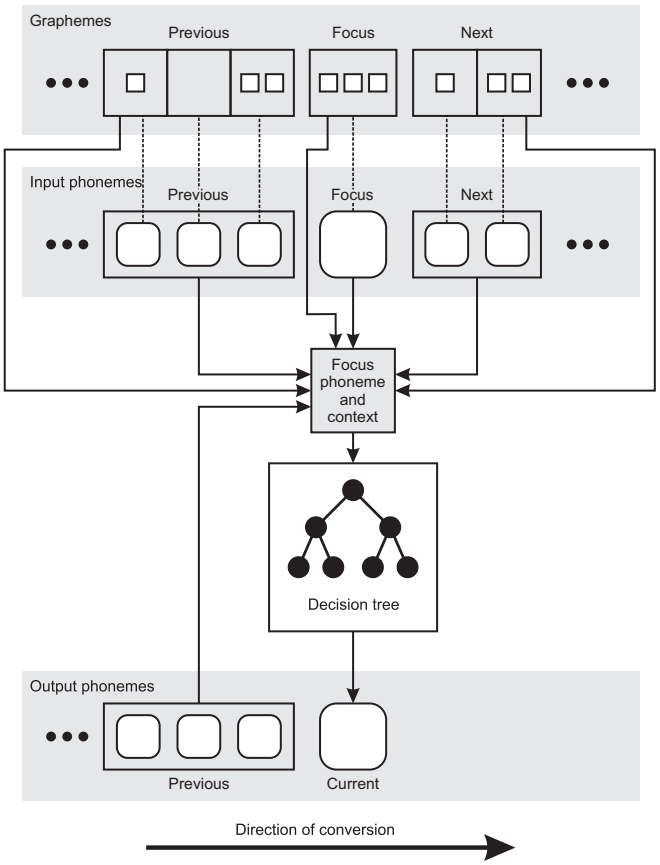


Fig. 3. Diagrammatic representation of grapheme-and-phoneme-to-phoneme (GP2P) conversion using decision trees.

training data is best used for pruning, an absolute increase of approximately 5% relative to the P2P systems. Further refinements that were considered were allowing questions to be based on grapheme clusters obtained in an analogous fashion to the phoneme clusters, including the length of the grapheme sequence aligned with a source phoneme, and allowing questions to include fixed grapheme sequences (*n*-tuples). However, none of these were found to yield any further performance improvements during cross-validation. The results of the GP2P experiments are presented in Table 8, and will be discussed in relation to the performance of G2P and P2P conversion in Section 7.

6. Multiple source dictionaries

A final question addressed by our work is whether the use of more than one source dictionary could further improve the performance of the GP2P approaches described in the previous section. As a specific example, we considered the two GP2P systems most accurately able to derive SSAE pronunciations. From Table 8, these are the BEEP to SAEDICT and PRONLEX to SAEDICT configurations. The GP2P algorithm was extended by allowing the decision trees to query the graphemes and phonemes in both BEEP and PRONLEX. To achieve this, BEEP was considered the primary and PRONLEX the

IPR2026-00272
Krisp Technologies EX1010 Page 231

Table 8
Grapheme-and-phoneme-to-phoneme (GP2P) conversion accuracies between dictionary pairs, with 95% confidence intervals and decision tree sizes.

Dictionary pair	Phoneme (%)	Word (%)	Size (nodes)
BEEP ⇒ SAEDICT	96.55 ± 0.31	81.58	2362
CMUDICT ⇒ SAEDICT	95.60 ± 0.34	76.79	2831
PRONLEX ⇒ SAEDICT	96.05 ± 0.33	78.99	2591
SAEDICT ⇒ BEEP	97.46 ± 0.27	86.78	1880
CMUDICT ⇒ BEEP	96.58 ± 0.32	82.28	2596
PRONLEX ⇒ BEEP	96.97 ± 0.30	84.41	2078
SAEDICT ⇒ CMUDICT	96.61 ± 0.32	82.38	2313
BEEP ⇒ CMUDICT	96.85 ± 0.31	83.96	2362
PRONLEX ⇒ CMUDICT	97.87 ± 0.25	88.44	1613
SAEDICT ⇒ PRONLEX	97.26 ± 0.27	84.60	2075
BEEP ⇒ PRONLEX	97.30 ± 0.28	85.44	1943
CMUDICT ⇒ PRONLEX	97.75 ± 0.25	87.22	1551

additional source, and PRONLEX pronunciations were aligned with BEEP pronunciations as described in Section 4.

The decision tree question set was extended to include dictionary-specific questions, such as "Is the preceding phoneme in PRONLEX a plosive?". This set of questions is a superset of those used for GP2P conversion with a single source dictionary, as was described in Section 5. Optimal decision tree parameters, again obtained by 10-fold cross-validation, were found to coincide with those that were optimal for the single source dictionary GP2P systems. The results, presented in Table 9, indicate that a small improvement in performance over the baseline BEEP to SAEDICT GP2P system was achieved, but that this improvement is not statistically significant.

It therefore appears that further significant gains using multiple source dictionaries are not achievable for SSAE. However, it is possible that the result would be different for other English accents, or for other languages.

7. Discussion

When comparing Tables 6 and 7, it is clear that the pronunciation of a word in a new accent is in all cases substantially more accurately determined from its pronunciation in another accent by means of P2P conversion, than by means of G2P conversion within the same accent. Furthermore, Table 8 shows that the use of source graphemes in addition to source phonemes in GP2P conversion is rewarded with a further improvement in accuracy. This additional improvement is beyond the 95% confidence interval for all dictionary pairs except CMUDICT to PRONLEX, which is also the pair for which the highest P2P conversion accuracy was obtained.

Converting between different accents also gives considerably more accurate results than merely using another accent's dictionary, as can be seen from Table 4. Substituting one English dictionary for another with no modifications gives at best a 60% and at worst a 30% word accuracy. For GP2P, however, even the poorest matching pair of dictionaries achieves a word accuracy above 76%.

Table 10 summarises the results presented in Tables 6–8, and indicates the average performance across all four dictionaries. The results in the table show that, on average, P2P conversion leads to a 4.3% absolute improvement in phoneme accuracy over G2P conversion, which corresponds to a 51% relative decrease in error rate. Furthermore, the percentage of words with pronunciation errors drops by 14.6%, which is a 40% relative reduction in error. The move from P2P to GP2P conversion results in an additional average absolute phoneme accuracy improvement of 1.1%, which corresponds to a 26% relative reduction in the error rate. The corresponding percentage of incorrectly pronounced words drops by 5.5%, which is a relative improvement of 25%. Overall, GP2P conversion achieves relative reductions of 64% and 55% in the phoneme and word error rates, respectively, when compared with G2P conversion.

Table 9
Grapheme-and-phoneme-to-phoneme (GP2P) conversion accuracies for SSAE when using one (BEEP) and two (BEEP and PRONLEX) source dictionaries, with 95% confidence intervals and decision tree sizes.

GP2P system	Phoneme (%)	Word (%)	Size (nodes)
BEEP ⇒ SAEDICT	96.55 ± 0.31	81.58	2362
BEEP + PRONLEX ⇒ SAEDICT	96.62 ± 0.31	81.84	2272

Table 10
Average accuracies achieved by the G2P, P2P and GP2P conversion approaches.

Dictionary	Phoneme (%)			Word (%)		
	G2P	P2P	GP2P	G2P	P2P	GP2P
⇒ SAEDICT	90.67	95.52	96.07	59.64	76.34	79.12
⇒ BEEP	91.76	96.00	97.00	65.24	79.50	84.49
⇒ CMUDICT	91.26	95.82	97.11	63.29	78.52	84.93
⇒ PRONLEX	92.22	95.88	97.44	65.82	78.04	85.75
Average	91.48	95.80	96.90	63.50	78.10	83.57

IPR2026-00272
Krisp Technologies EX1010 Page 232

The introduction of source graphemes in GP2P conversion also led to a considerable reduction in the size of the decision trees. While P2P conversion employed 2925 nodes on average, the corresponding figure for GP2P falls to 2183. However, both P2P and GP2P trees are much smaller than G2P trees which contained 15,400 nodes on average.

Hence, when a word's pronunciation is not available, as is fairly common in less prominent accents, it is better to first search for it in a different pronunciation dictionary. If it is found, it should be converted to the target accent by means of GP2P or P2P techniques. G2P methods should be reserved as a last resort.

8. Summary and conclusions

This study has investigated methods that can be used to derive the pronunciation of a word in a new accent from its known pronunciation in another accent of the same language.

The performance of these techniques has been evaluated by applying them to three accents of English: RP, GenAm and South African English. In order to achieve this, four pronunciation dictionaries were employed: BEEP for RP, CMUDICT and PRONLEX for GenAm, and SAEDICT for South African English.

As a baseline for comparison, G2P techniques were first applied to the dictionaries individually, with comparable results in all four cases. The G2P algorithm was then modified to allow conversion of pronunciations from one accent to another, first using only the phonemes of the source accent (P2P) and then using both the phonemes and the graphemes (GP2P).

If we consider the specific case of South African English, for which pronunciation resources are limited, our experiments show that the pronunciations of 82% of new words obtained using the developed GP2P conversion technique are correct. In comparison, this figure is 60% when using conventional G2P conversion.

Our research can be extended by performing listening tests, as well as by performing Automatic Speech Recognition, to determine on a perceptual and an acoustic level how well the generated pronunciations match the desired accent. This would provide a further indication of the validity of our findings. Furthermore, it is possible that the techniques which we have presented would bear fruit if applied to other accents of the same language, or even to related languages, such as Dutch and Afrikaans.

Acknowledgements

This material is based on work supported financially by the South African National Research Foundation (NRF) under Grants FA2007022300015 and TTK2007041000010, and was executed using the High-Performance Computing (HPC) facility at Stellenbosch University.

References

Andersen, O., Kuhn, R., Lazaridès, A., Dalsgaard, P., Haas, J., Nöth, E., 1996. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In: Proc. ICSLP, Philadelphia, USA.

BEEP, 2009. The British Example Pronunciation (BEEP) Dictionary. <http://svr-www.eng.cam.ac.uk/comp.speech/section1/lexical/beep.html> (accessed March 2009).

Bekker, I., 2009. The vowels of South African English. Ph.D. Thesis, North-West University, Potchefstroom, South Africa.

Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proc. ICASSP, Montreal, Canada.

Black, A., Lenzo, K., Pagel, V., 1998. Issues in building general letter to sound rules. In: Proc. ESCA Speech Synthesis Workshop, Jenolan Caves, Australia.

Bowerman, S., 2004. White South African English: phonology. In: Schneider, E., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), . In: A Handbook of Varieties of English, Vol. 1. Mouton de Gruyter, Berlin, pp. 931–942.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove.

CMU, 2009. Carnegie Mellon University Pronouncing Dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (accessed March 2009).

Daelemans, W., van den Bosch, A., 1994. Language-independent data-oriented grapheme-to-phoneme conversion. In: van Santen, J., Sproat, R., Olive, J., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer, New York, pp. 77–89.

Daelemans, W., van den Bosch, A., Zavrel, J., 1999. Forgetting exceptions is harmful in language learning. Machine Learn. 34, 11–41.

Damper, R., Marchand, Y., Adamson, M., Gustafson, K., 1999. Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. Comput. Speech Lang. 13 (2), 155–176.

Damper, R., Marchand, Y., Marsters, J., Bazin, A., 2004. Aligning letters and phonemes for speech synthesis. In: Proc. 5th ISCA Speech Synthesis Workshop, Pittsburg, USA.

Demberg, V., Schmid, H., Möhler, G., 2007. Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion. In: Proc. ACL, Prague, Czech Republic.

Han, K.-S., Chen, G.-L., 2004. Letter-to-sound for small-footprint multilingual TTS engine. In: Proc. ICSLP, Jeju, Korea.

Humphries, J., Woodland, P., Pearce, D., 2001. Using accent-specific pronunciation modelling for robust speech recognition. In: Proc. ICASSP, Salt Lake City.

Kienappel, A.K., Kneser, R., 2001. Designing very compact decision trees for grapheme-to-phoneme transcription. In: Proc. Eurospeech, Aalborg, Denmark.

LDC, 2009. COMLEX English pronouncing lexicon. <http://www.ldc.upenn.edu> (accessed March 2009).

Loots, L., Niesler, T., 2009. Data-driven phonetic comparison and conversion between South African, British and American English pronunciations. In: Proc. Interspeech, Brighton, UK.

Marchand, Y., Damper, R., 2000. A multistrategy approach to improving pronunciation by analogy. Comput. Linguist. 26 (2), 195–219.

Niesler, T., Louw, P., Roux, J., 2005. Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. Southern African Linguist. Appl. Lang. Studies 23 (4), 459–474.

Pagel, V., Lenzo, K., Black, A., 1998. Letter to sound rules for accented lexicon compression. In: Proc. ICSLP, Sydney, Ausralia.

Rabiner, L., Juang, B., 1993. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, USA.

Suontausta, J., Häkkinen, J., 2000. Decision tree based text-to-phoneme mapping for speech recognition. In: Proc. ICSLP, Beijing, China.

Taylor, P., 2005. Hidden Markov models for grapheme to phoneme conversion. In: Proc. Interspeech, Lisbon, Portugal.

IPR2026-00272
Krisp Technologies EX1010 Page 233

Torkkola, K., 1993. An efficient way to learn grapheme-to-phoneme rules automatically. In: Proc. ICASSP, Minneapolis, USA.

van den Heuvel, H., Martens, J.-P., Konings, N., 2007. G2P conversion of names. what can we do (better)? In: Proc. Interspeech, Antwerp, Belgium.

Wan, V., Dines, J., El Hannani, A., Hain, T., 2008. Bob: a lexicon and pronunciation dictionary generator. In: Proc. Spoken Language Technology Workshop, Goa, India.

Webster, G., Braunschweiler, N., 2008. An evaluation of non-standard features for grapheme-to-phoneme conversion. In: Proc. Interspeech, Brisbane, Australia.

Wells, J., 1982. Accents of English. Cambridge University Press, Cambridge.

IPR2026-00272
Krisp Technologies EX1010 Page 234

Appendix ZHAO01

MARC View

LEADER 02283nas a22004817i 4500
001 49457801
003 SIRSI
005 20251114090051.0
006 m o d
007 cr n
008 130709c20149999njumr p s | a0eng
010 a| 2013203345
019 a| DO NOT EDIT this 360 MARC RECORD
022 a| 2329-9304 y| 2329-9290
030 a| ITASFA
035 a| (WaSeSS)ssj0001079974
037 b| IEEE, 445 Hoes Ln., Piscataway, NJ 08854
040 a| DLC b| eng c| DLC d| OCLCA d| DKU d| OCLCF d| OCLCO d| VT2 d| OCLCO d| OCLCA d| W2U d| UUM d| DLC d| OCLCQ d| OCLCL d| EZC d| WaSeSS
042 a| pcc a| nsdp
082 0 0 a| 621.382 2| 23 1| https://id.oclc.org/worldcat/ddc/E36HC9jBKfpGqy8cdwf3r7FPbf
210 0 a| IEEE/ACM trans. audio, speech lang. process. b| (Online)
222 0 a| IEEE/ACM transactions on audio, speech, and language processing b| (Online)
245 0 0 a| IEEE/ACM transactions on audio, speech, and language processing h| [electronic resource] : b| a publication of the Signal Processing Society.
246 1 a| Transactions on audio, speech, and language processing
246 1 a| Audio, speech, and language processing
246 1 a| TASLP
260 a| [Piscataway, NJ] : b| [Institute of Electrical and Electronics Engineers], c| [2014]-
310 a| Monthly
362 1 a| Began in 2014
538 a| Mode of access: World Wide Web.
588 0 a| Print version record.
588 1 a| Volume 28 (2020) (IEEExplore viewed Feb. 13, 2020).
650 0 a| Speech processing systems v| Periodicals.
650 0 a| Electro-acoustics v| Periodicals.
655 0 a| Electronic journals.
710 2 a| IEEE Signal Processing Society.
710 2 a| Institute of Electrical and Electronics Engineers.
710 2 a| Association for Computing Machinery.
776 0 8 i| Print version: t| IEEE/ACM transactions on audio, speech, and language processing x| 2329-9290 w| (DLC) 2013203346 w| (OCoLC)852251724
780 0 0 t| IEEE transactions on audio, speech and language processing (Online) x| 1558-7924 w| (DLC) 2005215814 w| (OCoLC)62197773
856 4 0 u| http://SK8ES4MC2L.search.serialssolutions.com/?V=1.0&L=SK8ES4MC2L&S=JCs&C=TC0001079974&T=marc
910 a| CONSER online
949 a| Electronic resource w| ASIS c| 1 i| 49457801-1001 l| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 11/29/2025 o| .STAFF. 360 MARC

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES
Libraries Home
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory
(814) 865-6368

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES
Facebook
Twitter
Instagram

IEEE/ACM transactions on audio, speech, and language processing [electronic resource] : a publication of the Signal Processing Society

Additional Titles	IEEE/ACM trans. audio, speech lang. process. (Online), Transactions on audio, speech, and language processing, Audio, speech, and language processing, and TASLP
Published	[Piscataway, NJ] : [Institute of Electrical and Electronics Engineers], [2014]-
Additional Creators	IEEE Signal Processing Society, Institute of Electrical and Electronics Engineers, and Association for Computing Machinery

Access Online

serialssolutions.com

Continues

IEEE transactions on audio, speech and language processing (Online)

Dates of Publication and/or Sequential Designation

Began in 2014

Subject(s)

Speech processing systems—Periodicals

Electro-acoustics—Periodicals

Genre(s)

Electronic journals

ISSN

2329-9304

Publication Frequency

Monthly

Technical Details

Mode of access: World Wide Web.

ASK A LIBRARIAN

View MARC record | catkey: 49457801

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368

Libraries Intranet (Staff Only)

Accessibility Help

Website Feedback

Policies and Guidelines

Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook

Twitter

Instagram

Online Journals, Books and more

All | Journals Only | Books Only | Videos Only

Title begins with | Search

Show more search options ▼

Showing results 1 through 1 of 1
for the search: "TC0001079974"

Refine Results: All | Journals Only | Books Only | Videos Only

Limit by: Peer Reviewed | Open Access

IEEE/ACM transactions on audio, speech, and language processing
ISSN: 2329-9290
Peer Reviewed
01/01/2014 to 01/31/2024 in ACM Digital Library
2014 to Present in IEEE Electronic Library (IEL)

Search inside this journal | Search

Back to Top

Copyright ©2025 The Pennsylvania State University.
All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Last Updated on 12-19-2025 © Ex Libris, a ProQuest Company, 2000-2025

Appendix ZHAO02

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion

Publisher: IEEE Cite This PDF

Guanlong Zhao ; Ricardo Gutierrez-Osuna All Authors

15 Cites in Papers

657 Full Text Views

Abstract

Document Sections

I. Introduction

II. Literature Review

III. Methods

IV. Experimental Setup

V. Results

Show Full Outline

Authors

Figures

References

Citations

Keywords

Metrics

More Like This

Footnotes

Abstract:
Accent conversion (AC) aims to transform non-native utterances to sound as if the speaker had a native accent. This can be achieved by mapping source speech spectra from a native speaker into the acoustic space of the target non-native speaker. In prior work, we proposed an AC approach that matches frames between the two speakers based on their acoustic similarity after compensating for differences in vocal tract length. In this paper, we propose a new approach that matches frames between the two speakers based on their phonetic (rather than acoustic) similarity. Namely, we map frames from the two speakers into a phonetic posteriorgram using speaker-independent acoustic models trained on native speech. We thoroughly evaluate the approach on a speech corpus containing multiple native and non-native speakers. The proposed algorithm outperforms the prior approach, improving ratings of acoustic quality (22% increase in mean opinion score) and native accent (69% preference) while retaining the voice quality of the non-native speaker. Furthermore, we show that the approach can be used in the reverse conversion direction, i.e., generating speech with a native speaker's voice quality and a non-native accent. Finally, we show that this approach can be applied to non-parallel training data, achieving the same accent conversion performance.

Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 27 , Issue: 10, October 2019)

Page(s): 1649 - 1660

DOI: 10.1109/TASLP.2019.2926754

Date of Publication: 04 July 2019

Publisher: IEEE

ISSN Information:

Funding Agency:

SECTION I.
Introduction

Learners who acquire a second language (L2) after a “critical period” [1] usually speak with a non-native accent. Having a non-native accent can often reduce the speaker's intelligibility [2] and may also lead to discriminatory attitudes [3], [4]. Therefore, non-native speakers have much to gain by improving their pronunciation. Several studies [5], [6] have shown that having a suitable native (L1) speaker to imitate – a so-called “golden speaker” with similar voice characteristics as the learner but with a native accent, can be beneficial in pronunciation training. Based on these findings, Felps *et al.* [7] suggested that such a “golden speaker” could be created by resynthesizing the non-native speaker's own voice with a native accent borrowed from a native reference speaker.

Traditional voice-conversion (VC) methods [8]– [11] cannot be used for this purpose since VC cannot decouple the speaker's voice quality from her or his accent, i.e., VC assumes that accent is part of the speaker's identity. In this work, we distinguish two concepts: *voice quality*, which focuses on the physical characteristics of the speaker's voice (e.g., vocal tract and glottal configuration, pitch range), and *speaker identity*, a combination of *voice quality* and other speaker characteristics (e.g., accent, speaking rate, intonation, word choice).

IPR2026-00272
Krisp Technologies EX1010 Page 240

To address the accent-and-voice-quality entanglement issue of traditional VC methods, Aryal and Gutierrez-Osuna [12] proposed a modified VC method where source frames (i.e., from the native reference speaker) and target frames (i.e., from the non-native speaker) were paired based on their acoustic similarity. In a first step, the authors applied vocal-tract length normalization (VTLN) to the source speech, so it matched the target speaker's vocal-tract length. Then, they paired each frame in the source corpus with the closest frame in the target corpus, and vice versa. Though VTLN did improve frame pairing compared to time alignment (i.e., the conventional approach in VC), vocal-tract length is just one of the potentially many differences between speakers, and it is too coarse to account for differences in pronunciation.

To address this issue, we present an approach that matches source and target frames based on their phonetic content. Leveraging advances in acoustic modeling [13], we extract phonetic information from phonetic posteriorgrams (PPGs) [14]. Namely, we compute the posteriorgram for each source and target speech frame through a speaker-independent acoustic model trained on a large corpus of native speech. Then, we use the symmetric Kullback-Leibler (KL) divergence [15] in posteriorgram space to match source and target frames. The result is a set of source-target frames that are paired based on their phonetic similarity, with which we train a Gaussian Mixture Model (GMM) to model the joint distribution of source and target Mel-Cepstral Coefficients (MCEPs). In a final step, we map source MCEPs into target MCEPs using maximum likelihood estimation of spectral parameter trajectories considering the global variance [8] of the target speaker. Our implementation is based on a conventional GMM spectral mapping method to ensure a fair comparison with the prior study [12], but our proposed frame matching method can be combined with any spectral mapping methods (e.g., neural networks, frequency warping) that take frame pairs as input.

Our approach differs from prior works on accent conversion, which modify speech features that carry accent information, such as prosody, formants, spectral envelopes, or articulatory gestures [7], [16]– [18]. Instead, we use a VC technique to capture the voice quality of the (target) non-native speaker while preserving the (source) native speaker's pronunciation characteristics – both segmental and prosodic. Unlike VC methods, however, we avoid the issue of time aligning source and target utterances, which is problematic when the target speaker is non-native. Our approach is related to that of Xie *et al.* [19], who used speaker-adaptive acoustic models to generate posteriorgrams for VC. Their method groups all target speaker training data into phonetic clusters in the posteriorgram space using symmetric KL divergence and K-means clustering. Then, each frame of the source speaker's corpus is mapped to the centroid of the closest target phonetic cluster. The final converted speech is generated from those closest cluster centroids using the maximum probability trajectory generation algorithm. In contrast with their frame clustering approach, we use PPGs to produce frame pairs between source and target speakers, and then we train a GMM using those frame pairs. A second major difference with their approach is that we use speaker-independent acoustic models trained on native speech to ensure that the PPGs only reflect native pronunciations, whereas their approach uses speaker-adaptive training, which would introduce non-native pronunciations into the acoustic models. Initial findings from this work were presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2018 [20]. That earlier conference paper presented preliminary listening test results that verified the effectiveness of the PPG-based frame-matching method. The present manuscript describes our method in detail and significantly expands the perceptual studies and data analyses, including an experimental comparison of the proposed method on parallel and non-parallel data.

The manuscript is organized as follows. Section II reviews prior work on accent conversion in the acoustic and articulatory domains, and discusses the connection between accent conversion and voice conversion. Section III describes the experimental methods used in the study, including the phonetic posteriorgrams, the proposed frame-pairing method for accent conversion, as well as the baseline systems we used for comparison. Section IV describes the experimental setup used for the study, including the speech corpus we used to train the acoustic model, and the native/non-native speech corpus we used for accent conversion. Section V presents results on three different experiments we used to evaluate various aspects of the algorithm. The paper concludes with a discussion of our findings and directions for future work.

SECTION II.
Literature Review

A. Algorithms for Accent Conversion

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 241

Foreign and non-native accents occur when speech deviates from the expected acoustic (e.g., formants) and prosodic (e.g., intonation, duration, and rate) norms of a language [7]. Therefore, prior work has focused on modifying certain speech characteristics to alter the perceived accent. In early work, Yan *et al.* [21] used a voice-morphing software to change the trajectories of formants, pitch, and duration to convert between three different English dialects (British, Australian, and General American English). The authors found that prosodic modifications produced noticeable differences on perceived accent, although not as significant as those produced by modifying formant trajectories. In the approach of Felps *et al.* [7], the spectral envelope of the non-native speech was replaced with that of the native speaker's, which had been normalized to the non-native speaker's vocal tract length with a piecewise linear warping function. Their results showed that the segmental correction was able to significantly reduce the foreign accentedness of the modified utterances. More recently, Jügler *et al.* [22] used PSOLA to correct the prosody of non-native German speech spoken by native French speakers. Prosodic (duration and pitch) corrections were performed at the syllable level, and the results showed a moderate but significant reduction in accentedness of the corrected speech.

A couple of studies also tried to blend native and non-native spectra to control the accent. Huckvale and Yanagisawa [17] blended the spectral envelope of non-native Japanese speech produced by an English Text-To-Speech (TTS) with its native counterpart through voice morphing to reduce the accent. Aryal *et al.* [18] decomposed the cepstrum into spectral slope and spectral detail, and then generated accent conversions by combining the spectral slope of the non-native speaker with a morph of the spectral detail of the native speaker. Though these spectra-blending methods can reduce non-native accents, they also tend to produce syntheses that are perceived as a “third speaker,” one who is different from either the source (native) or target (non-native) speaker. To tackle this problem, Aryal and Gutierrez-Osuna [12] adapted VC techniques to perform accent conversion. The authors used vocal-tract-length normalization (VTLN) before pairing acoustic frames between source (native) and target (non-native) speaker, then built a GMM using those frame pairs to perform VC. This method was able to reduce non-native accent significantly, while retaining the non-native speaker's voice quality; however, it required a relatively large set of parallel recordings from the two speakers, and VTLN only accounted for a subset of the speaker characteristics.

An alternative to using acoustic methods is to operate in the articulatory domain. Along these lines, Felps *et al.* [16] used an articulatory synthesizer based on unit-selection to replace mispronounced non-native diphones with those from the non-native corpus that matched the articulatory configuration of a reference utterance from a native speaker. Later, Aryal and Gutierrez-Osuna used GMMs [23] and DNNs [24] to build an articulatory synthesizer (i.e., a mapping from articulatory gestures into acoustics) for the non-native speaker, then drove the GMM/DNN with articulatory gestures from a native speaker. Methods based on articulatory data generate syntheses that sound more like the non-native (target) speaker than acoustic methods, since they effectively decouple linguistic information (e.g., articulatory gestures from a native [source] speaker) from voice quality (captured by the articulatory-to-acoustic synthesizer of the non-native speaker). However, articulatory methods are expensive and require specialized equipment to collect articulatory data, so they are impractical for pronunciation training.

B. Connection Between Accent and Voice Conversion

Accent conversion is closely related to the problem of voice conversion [25]. Voice conversion transforms a source speaker's speech into that of a (known) target speaker. The conversion aims to match the voice characteristics of the target speaker, which may include vocal tract configuration, glottal characteristics, pitch range, pronunciation, and speaking rate. Ideally, the only information retained from the source speech is its linguistic content, i.e., the words that were uttered. Popular methods for voice conversion include joint-density GMMs [8], frequency warping [26], [27], DNNs [28], [29], and sparse coding [11], [30]– [32]. Accent conversion modifies speech at a finer level of granularity, and seeks to combine the linguistic content and pronunciation of the source speaker with the voice quality of the target speaker. Therefore, accent conversion is a more challenging problem than voice conversion in the sense that, first, there is no ground truth for the output voice, and second, accent conversion needs to split the speech into voice quality (converted) and accent (preserved), whereas voice conversion jointly converts both.

SECTION III.
Methods

A. Phonetic Posteriorgrams

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 242

At its core, our proposed method relies on Phonetic Posteriorgrams (PPGs) to measure the similarity of speech frames across speakers. A phonetic posteriorgram is computed by segmenting speech into frames and computing the posterior probability that each frame belongs to a set of pre-defined phonetic units. As an example, Fig. 1 shows the PPG of the spoken word "air." In practice, it is advisable to include context when computing the PPG by concatenating each speech frame with its neighboring right and left frames. Moreover, phoneme labels are too coarse to describe the variety of speech sounds. Therefore, the dimensions in a phonetic posteriorgram are often associated with triphones, as we will see next.

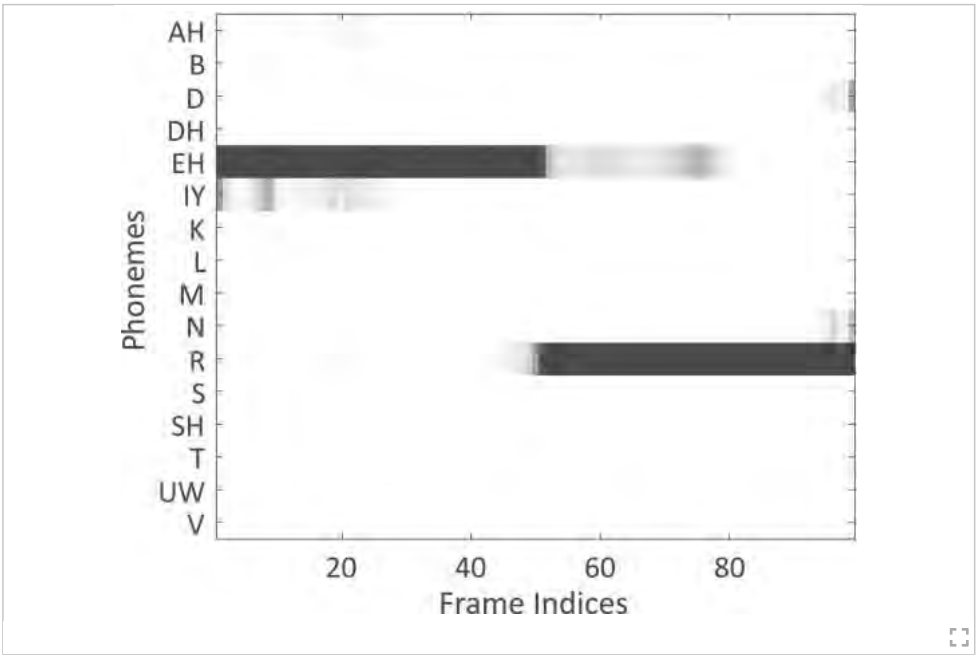


Fig. 1.
PPG for the word "air," whose phonetic transcription in ARPABET is "EH R." For visualization purposes, we used a subset of the ARPABET phoneme set and omitted phonemes that had small values.

Generally, the phonetic posteriorgram is computed from the acoustic model in an automatic speech recognizer (ASR). The acoustic model in ASR acts as a sequential classifier: given an input acoustic feature vector, the acoustic model assigns how likely it is that the vector belongs to each of a set of states/senones. In recent years, acoustic models based on DNNs have yielded state-of-the-art speech recognition accuracy [13]. The most advanced ASR systems can achieve Word Error Rates that are comparable to or better than expert human transcribers on specific tasks [33].

In this work, we compute phonetic posteriorgrams using a p-norm DNN [34] as the acoustic model. The input layer accepts a feature frame accompanied by its left and right neighbors; then the input is de-correlated by a fixed linear transformation [35]. The de-correlated features are then passed through N hidden layers, each employing the p-norm non-linearity ${\boldsymbol y}=\Vert {\boldsymbol x}\Vert _p=(\sum _i|x_i|^p)^{\frac{1}{p}}$, where ${\boldsymbol y}$ is one output dimension of a hidden layer and ${\boldsymbol x}$ represents a group of hidden neurons of that layer. Therefore, the number of output dimensions of each hidden layer is smaller than the number of hidden neurons. The output of the p-norm layer is then processed by a normalization layer to limit its standard deviation to one [34]. The output of the final hidden layer is fed into a softmax layer that produces more output nodes than the desired number of senones using a technique called "mixing-up" [34]. "Mixing-up" operates as follows. About halfway throu training, the dimension of the softmax layer is increased by letting each output senone's probability be a s over potentially multiple "mixture components." The mixture components are distributed using a power proportional to the senone class priors. The neural network then "group-sums" the output of the softmax layer according to the group assignment defined in the "mixing-up" step, resulting in the final output nodes that correspond to individual senones. Fig. 2 shows the overall structure of the p-norm deep neural network that we use in this work.

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 243

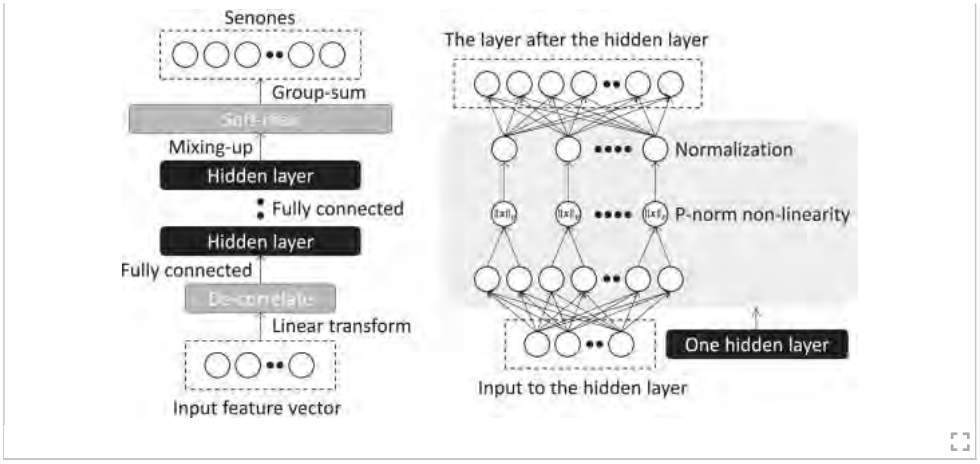


Fig. 2.
P-norm deep neural network structure for acoustic modeling.

During training, inputs to the p-norm DNN consist of stacked MFCC frames ${\boldsymbol X}$, whereas target outputs ${\boldsymbol Y}$ are senone labels obtained from force-alignment using an existing GMM-HMM speech recognizer. The training objective is the sum (across all frames of training data) of the log-probability of ${\boldsymbol Y}$ given ${\boldsymbol X}$: $\sum _i\log {p({\boldsymbol Y}_i \vert {\boldsymbol X}_i)}$. After the DNN is fine-tuned using Stochastic Gradient Descent [36], we compute the posterior probability of observing senone l given the speech frame ${\boldsymbol x}$ by doing a complete forward propagation, \begin{equation*} p({l\vert {\boldsymbol x}})=\sum _{g\in G}\frac{\exp {x_{g}^{^{\prime }}}}{\sum _{k}\exp {x_{k}^{^{\prime }}}}, \tag{1} \end{equation*}

View Source

where $x_{k}^{^{\prime }}$ is the output of the hidden layer that precedes the softmax layer, and G is the set of softmax outputs that are grouped into senone l during the “mixing-up” procedure. A PPG frame of ${\boldsymbol x}$ is constructed by forming a vector from all possible values of $p(l \vert {\boldsymbol x})$, see eq. (2).

B. Frame Pairing

Conventional voice conversion methods use time alignment to pair frames from source and target utterances. As such, a VC model trained from time-aligned frame pairs will retain the non-native speaker's accent. Instead, to perform accent conversion, the pairing must be based on the phonetic similarity between source and target frames. In this way, each native speech frame is associated with its most similar non-native counterpart in terms of pronunciation. If we train a spectral conversion model between these frame pairs, the pronunciation from the native speech data will be preserved and the spectral envelope of the native speaker will be modified to match the non-native speaker's voice quality.

1) Frame Pairing Based on Phonetic Similarity (AC-PPG)

We use PPGs to pair frames between the native and the non-native speaker. Our rationale is straightforward: if an ASR trained on native speech determines that a non-native speech segment ${\boldsymbol y}$ is close to the native speech production of a particular phoneme (or triphone, in our case), then it is reasonable to pair ${\boldsymbol y}$ with a native speech segment ${\boldsymbol x}$ with the same phonetic label; see Fig. 3 (a). Specifically, our approach works as follows. In a first step, we compute PPG frames for speech frames from the two speakers, \begin{equation*} \mathcal {L}_{{\boldsymbol x}_i}=[p(l_1 \vert {\boldsymbol x}_i), p(l_2 \vert {\boldsymbol x}_i), \ldots, p(l_V \vert {\boldsymbol x}_i)], \tag{2} \end{equation*}

View Source

where ${\boldsymbol x}_i$ is the acoustic feature vector of the i-th speech frame; $V={\lbrace l_1, l_2, \ldots, l_V\rbrace }$ is the predefined senone set; $p(l_j\vert {\boldsymbol x}_i)$ is the conditional

IPR2026-00272
Krisp Technologies EX1010 Page 244

probability that the speech frame belongs to senone l_j given ${\boldsymbol x}_i$; $\sum _{j} p(l_j \vert {\boldsymbol x}_i)=1$.

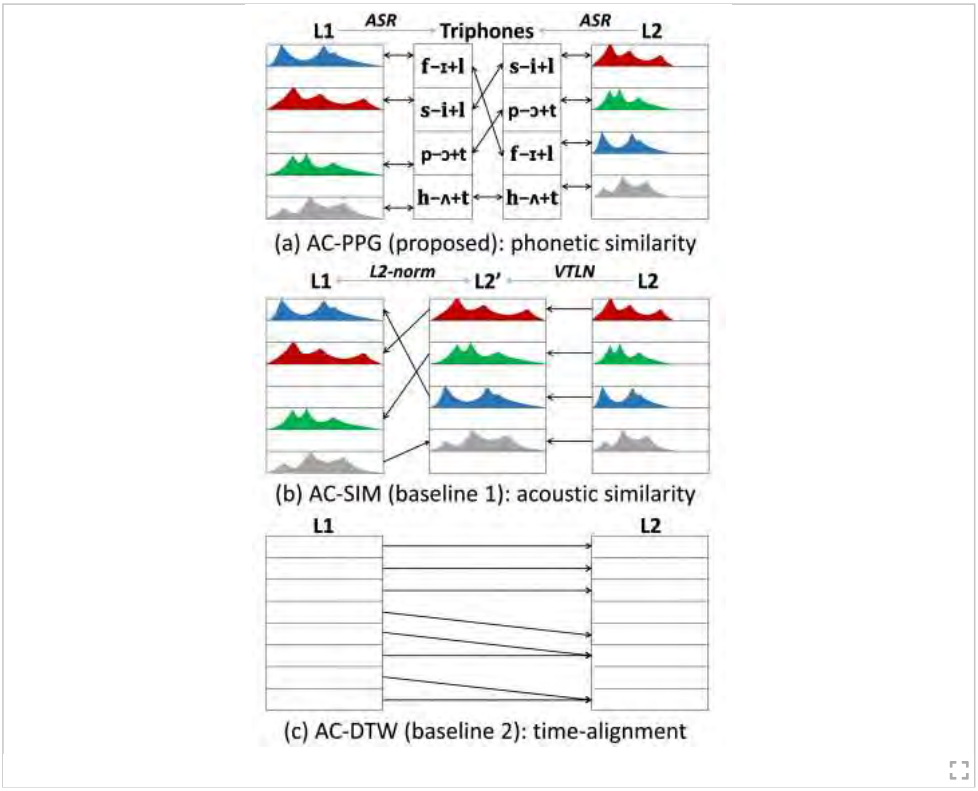


Fig. 3.
L1: native, L2: non-native. (a) AC-PPG: proposed AC algorithm that uses phonetic similarity. (b) AC-SIM: Baseline 1 that uses acoustic similarity through VTLN to pair frames [12]. (c) AC-DTW: Baseline 2; native and non-native frames are time-aligned following their ordering in the data.

Given posterior feature vectors $\mathcal {L}_{{\boldsymbol x}_i}$ and $\mathcal {L}_{{\boldsymbol x}_j}$, we calculate their distance using the symmetric KL divergence, \begin{equation*} D(\mathcal {L}_{{\boldsymbol x}_i }, \mathcal {L}_{{\boldsymbol x}_j})=(\mathcal {L}_{{\boldsymbol x}_i }-\mathcal {L}_{{\boldsymbol x}_j})\cdot (\log {\mathcal {L}_{{\boldsymbol x}_i}}-\log {\mathcal {L}_{{\boldsymbol x}_j}}). \tag{3} \end{equation*}

View Source

The symmetric KL divergence [15] is commonly used to compute the similarity between distributions, and here, each frame of the PPG functions like a distribution. For each source (i.e., native) frame ${\boldsymbol x}_i$ we find its closest target (i.e., non-native) frame ${\boldsymbol y}_{i}^{*}$, \begin{equation*} {\boldsymbol y}_{i}^{*}=\mathop{\rm arg\,min}_{\forall {\boldsymbol y}}D(\mathcal {L}_{{\boldsymbol x}_i}, \mathcal {L}_{{\boldsymbol y}}). \tag{4} \end{equation*}

View Source

PDF

Help

Likewise, for each non-native frame ${\boldsymbol y}_{i}$ we find its closest native frame ${\boldsymbol x}_{i}^{*}$, \begin{equation*} {\boldsymbol x}_{i}^{*}=\mathop{\rm arg\,min}_{\forall {\boldsymbol

IPR2026-00272
Krisp Technologies EX1010 Page 245

x}}D(\mathcal {L}_{{\boldsymbol x}}, \mathcal {L}_{{\boldsymbol y}_i}). \tag{5} \end{equation*}

View Source

Each frame pairing process only involves two speakers – the given native and non-native speakers. The frame pairing does not constrain the search space. Therefore, it is possible to pair multiple frames from one speaker with the same frame from the other speaker. In this case, we duplicate that frame multiple times. The resulting frame pairs are used to train a Gaussian Mixture Model (GMM).

2) Baseline Methods for Frame Pairing

We compared the proposed PPG-based method against two baseline techniques for frame pairing: the acoustic similarity method of Aryal and Gutierrez-Osuna [12], and dynamic time warping.

Baseline 1. *(AC-SIM):* Following [12], we measured acoustic similarity as the inverse of the L2-norm between native and non-native speaker frames, after normalizing the native speaker to match the vocal tract length of the non-native speaker; see Fig. 3 (b).

In a first step, we learn a VTLN transform to reduce physiological differences in vocal tract between the two speakers. For this purpose, we time-align parallel training utterances of the two speakers, each utterance represented as a sequence of MFCCs. Following Panchapagesan and Alwan [37], we then learn a linear transform between the MFCCs of both speakers using ridge regression: \begin{equation*} T^*=\mathop{\rm arg\,min}_{T} \Vert {\boldsymbol x}-T{\boldsymbol y}\Vert ^2 + \lambda \Vert T \Vert ^2, \tag{6} \end{equation*}

View Source

where ${\boldsymbol x}$ and ${\boldsymbol y}$ are vectors of MFCCs from the native and non-native speakers, respectively, and T^* is the VTLN transform. Next, for each native vector ${\boldsymbol x}_i$ we find its closest non-native vector ${\boldsymbol y}_j^*$ as: \begin{equation*} {\boldsymbol y}_j^*=\mathop{\rm arg\,min}_{\forall {{\boldsymbol y}}} \Vert {\boldsymbol x}_i-T^*{\boldsymbol y}\Vert ^2. \tag{7} \end{equation*}

View Source

We repeat the process for each non-native vector ${\boldsymbol y}_i$ to find its closest match ${\boldsymbol x}_j^*$: \begin{equation*} {\boldsymbol x}_j^*=\mathop{\rm arg\,min}_{\forall {{\boldsymbol x}}} \Vert {\boldsymbol x}-T^*{\boldsymbol y}_i\Vert ^2. \tag{8} \end{equation*}

View Source

The above process results in a lookup table where each native and non-native frame in the database is paired with the closest one from the other speaker.

Baseline 2. *(AC-DTW):* As our second baseline method, we use Dynamic Time Warping (DTW) [38] to time-align native and non-native frames, as illustrated in Fig. 3 (c).

We note that baselines 1 and 2 need parallel data for training, whereas the proposed method can operate on non-parallel data, as we shall see in Section V-C.

C. Spectral Conversion

To ensure a fair comparison between the three frame-pairing methods, we use a common spectral conversion technique to map a native source speaker's spectral features to match a non-native target speaker's voice quality. Following Toda *et al.* [8], we use a GMM to model the joint distribution of source and target frame pairs, and then use maximum likelihood parameter generation (MLPG) with global variance (GV) [39] to generate the converted speech for a given source utterance. Specifically, we use 2D-dimensional acoustic

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 246

features, $X_t=[x_t^\top,\Delta x_t^\top]^\top$ from the source speaker, and $Y_t=[y_t^\top,\Delta y_t^\top]^\top$ from the target speaker, consisting of D-dimensional static and dynamic features, where $(\cdot)^\top$ denotes the transpose. Given the paired source and target features, we train a GMM to model the joint probability density $p(X,Y \vert \theta)$ where θ denotes model parameters, estimated using Expectation-Maximization (EM): \begin{equation*} \theta =\text{EM}(\mathop{\rm arg\,max}_{\theta }p(X,Y\vert \theta)). \tag{9} \end{equation*}

View Source

When converting source static and dynamic feature vectors $X=[X_1^\top, X_2^\top, \ldots, X_T^\top]^\top$ to the target static feature vectors $y=[y_1^\top, y_2^\top, \ldots, y_T^\top]^\top$ – after the GMM is trained, we maximize the function below with respect to y, \begin{equation*} \hat{y}=\mathop{\rm arg\,max}_{y}\log {(p(Y \vert X, \theta)^{\omega }p(\nu (y)\vert \theta _{\nu }))}, \quad Y=Wy, \tag{10} \end{equation*}

View Source

where $p(Y \vert X, \theta)$ denotes the conditional probability density function (PDF) on the target static and dynamic feature vectors, and $p(\nu (y)\vert \theta _{\nu })$ represents the likelihood of a PDF on the global variance of the target feature vectors, which is represented as a separate GMM (one mixture) and trained using the EM algorithm as well. W is a matrix that appends dynamic features to the static features, and ω adjusts the relative importance between the two distributions and is set as the ratio of number of dimensions between vectors $\nu (y)$ and Y ($=1/2T$). We use a GMM instead of a DNN in this study to focus on low-resource accent conversion scenarios – in real pronunciation training applications, we generally have a limited amount of data from the non-native speakers.

D. Pitch Scaling

Previous studies [7], [17], [21] have shown that prosody modification is an essential part of accent conversion, and the pitch contour contains identity-related information. Since pitch modification is not the focus of this study, we follow the standard procedure [8] and use the pitch trajectory from the source (native) speaker, which captures native intonation patterns, then normalize it to match the pitch range of the target (non-native) speaker using mean and variance normalization in the $\log {F_0}$ space.

SECTION IV.
Experimental Setup

A. DNN Acoustic Model for Extracting PPG

To train the DNN acoustic model, we used Kaldi's Librispeech recipe.[1] The model is a p-norm DNN ($p=2$), as introduced in the method section, with five hidden layers. We extracted 13-dim MFCC vectors with a 7-frame context, passed the concatenated 91-dim (13 × 7) MFCCs through a Linear Discriminant Analysis (LDA) to generate a 40-dim input feature vector, then concatenated nine frames of such 40-dim LDA features as the final input to the DNN. The 360-dim (40 × 9) input features were de-correlated using a fixed linear transform. All hidden layers had 5,000 hidden neurons and output 500 activations because each p-norm non-linearity was computed over ten hidden neurons. Every hidden layer was fully-connected with the previous layer. Right after the last hidden layer was a softmax layer of 14,000 nodes; those nodes were then "group-summed" to produce the final output across senones (5,816 dimensions, which were obtained from state-tying on a phonetic decision tree built from the transcripts of the training data; see [for more details on how the decision tree was constructed). The DNN acoustic model was trained on Librispeech's [41] training set, a speech recognition corpus that contains 960 hours of native English spe the majority being American English. In the following experiments, the Librispeech corpus was used solely for building the acoustic model.

B. Speech Corpus for Accent Conversion

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 247

For the native speech synthesis corpus, we used two speakers from the CMU ARCTIC dataset [42]: BDL and CLB. Those recordings have a sampling rate of 16 KHz. For the non-native (L2) English speech synthesis corpus, we used five non-native speakers from the L2-ARCTIC corpus [2] [43]: two native Hindi speakers, two native Korean speakers; and one native Arabic speaker. Each non-native speaker produced the full ARCTIC dataset ($\sim$1100 utterances; around one hour of speech). The speech was recorded in a quiet room at 44.1 KHz. For the following experiments, we down-sampled all the non-native speech data to 16 KHz using sox. [3] The speaker demographic information is summarized in Table I. For the non-native speakers, their English proficiency level was measured in their TOEFL iBT scores [4] [46].

TABLE I Demographic Information of the Speakers

Speaker	Gender	Native Language	English Proficiency
BDL	M	English	Native
CLB	F	English	Native
RRBI	M	Hindi	91
TNI	F	Hindi	99
HKK	M	Korean	114
YKWK	M	Korean	N/A
ABA	M	Arabic	94-101

C. System Configuration

In what follows, we will refer to the proposed frame-pairing algorithm, baseline 1 (acoustic similarity), and baseline 2 (dynamic time warping) as **AC-PPG**, **AC-SIM**, and **AC-DTW**, respectively.

We used the TANDEM-STRAIGHT vocoder[5] [49] to decompose speech into aperiodicity (AP), F_0, and a 513-dim spectral envelope. Then, we computed 25-dim MFCCs [6] from the spectral envelopes to learn the VTLN transform and pair frames using acoustic similarity (AC-SIM); see Section III-B2. AC-DTW also used those MFCCs (excluding MFCC$_{0}$) to time-align a source speaker to a target speaker. AC-PPG used the 5816-dim PPGs extracted by the acoustic model to perform frame pairing.

We also computed 25-dim MCEPs from the spectral envelopes as the acoustic feature (excluding MCEP$_{0}$ since it is energy) to train the spectral conversion models (GMMs) and convert speech from the native speaker to the non-native speaker. MCEPs from the two speakers were frame paired using the three methods (AC-PPG, AC-SIM, AC-DTW) before being fed to the GMMs. Following Aryal and Gutierrez-Osuna [12], all GMMs had 128 mixture components with diagonal covariance matrices. Input features to the GMM include delta features, and therefore the joint feature vectors had 96 dimensions. Once we converted the native speaker's MCEPs to the non-native speaker's space, we reconstructed the spectrogram from the converted MCEPs (MCEP$_{0}$ being copied from the native speaker), and combined it with the native speaker's AP and normalized F_0 to synthesize speech using the TANDEM-STRAIGHT vocoder. The conversion pipeline is illustrated in Fig. 4.

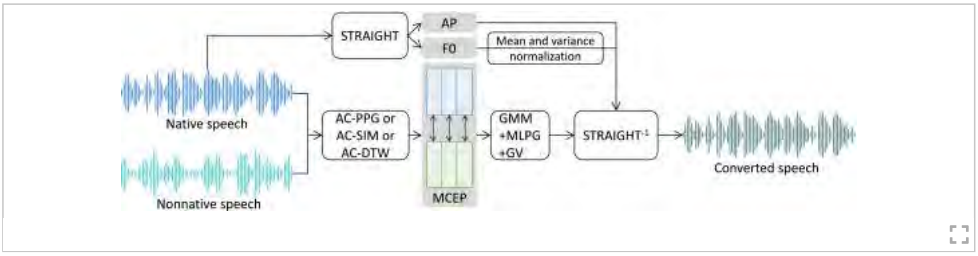


Fig. 4.
Accent conversion workflow; frame pairing can be AC-PPG, AC-SIM (baseline 1), or AC-DTW (baseline 2).

All experiments were conducted on a desktop running Windows 10 with an Intel Core i7-7700 K CPU@4.2 GHz, 16 GB of memory, and an NVidia GTX 1070 GPU. Most of the algorithms were implemented and run on Matlab v9.3, except for the acoustic model and PPGs, which were computed using Kaldi on Ubuntu 16.04.

IPR2026-00272
Krisp Technologies EX1010 Page 248

SECTION V.
Results

We conducted three sets of perceptual listening studies to evaluate different properties of the proposed frame-pairing algorithm. In the first experiment, we compared the approach against the two baseline systems by its ability to reduce perceived accents while matching the voice quality of the non-native speakers. In the second experiment, we evaluated whether the approach could also be used for the reverse purpose, i.e., to impart a non-native accent to a native speaker's voice. In the third and final experiment, we evaluated the approach to perform accent conversion using non-parallel speech corpora.

We recruited anonymous human participants from Amazon's Mechanical Turk platform[7] for our listening tests. Following Buchholz and Latorre [50], all listening tests included calibration trials designed to be easy to judge, and we used the participants' responses on those calibration trials to detect cheating behaviors. We excluded data from participants whose responses were below chance level on those calibration questions. All participants' calibration responses were excluded from the final analyses. In addition, and following [16], all human subjects passed a screening test that consisted of identifying various American English accents. We compensated participants for their time at an hourly rate of eight USD. In all experiments, the reference native and non-native English speech were resynthesized from their MCEPs using TANDEM-STRAIGHT to keep their acoustic quality comparable with the converted speech, which went through the same vocoder compression. When selecting testing samples, we always randomly draw from the available pools, i.e., we did not cherry-pick the audio clips. All test trials were randomly presented. For any listening tests that required pairwise comparisons, the presentation order within an utterance pair was counterbalanced. Unless otherwise noted, we used paired-sample t-tests for the analyses.

A. Experiment 1: Comparing AC-PPG Against Baselines

In this experiment, we considered five native to non-native speaker pairings for accent conversion: BDL to RRBI, BDL to HKK, BDL to YKWK, BDL to ABA, and CLB to TNI. For each speaker pair, we used 100 parallel utterances for training and 50 utterances for testing; there was no overlap between the two sets. We performed accent conversion on all 50 test utterances using models trained on each of the three frame-pairing algorithms, i.e., AC-PPG, AC-SIM, and AC-DTW.

Acoustic quality: We used a standard five-point (1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent) Mean Opinion Score (MOS) to rate the acoustic quality of the synthesized speech. Thirty listeners rated 150 test samples: 50 per system, 10 utterances per conversion direction. Results are shown in Fig. 5. The proposed method (AC-PPG) received a MOS rating of 2.99, which was significantly higher than either baseline: AC-SIM (2.45 MOS, 22% relative improvement; $t(29)=15.61$, $p\ll 0.001$; one-tail) and AC-DTW (2.55 MOS, 17% relative improvement; $t(29)=12.04$, $p\ll 0.001$; one-tail). These results suggest that the proposed algorithm can boost the acoustic quality of the converted speech using exactly the same training data without even having to modify the GMM training and spectral conversion methods.

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 249

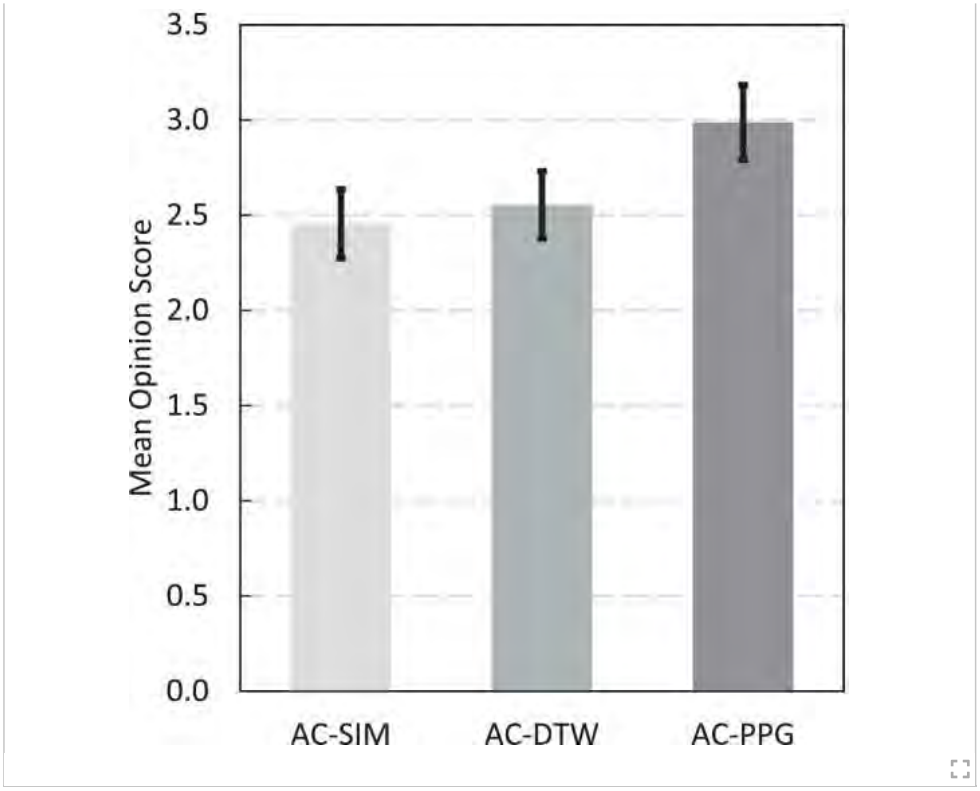


Fig. 5.
Mean Opinion Scores for the proposed method (AC-PPG) and the two baseline methods (AC-SIM, AC-DTW); the error bars show 95% confidence intervals.

Voice quality: Following our prior work [32], we used a voice similarity score (VSS) ranging from -7 (definitely different speakers) to +7 (definitely same speaker) to assess the speaker's voice quality. Twenty-six participants rated 150 utterance pairs: 50 pairs per system (25 AC-L1 and 25 AC-L2 pairs, each pair contained one AC and one L1 [native]/L2 [non-native] utterance), and ten pairs per conversion direction. Following Felps *et al.* [7], we played utterances in reverse to prevent the accent from interfering with the perception of voice quality. In each trial, listeners first answered whether both utterances were produced by the same speaker (+1) or different speakers (–1), and then rated their confidence level on a 7-point scale (1-Not at all confident, 7-Extremely confident). The VSS was then compiled by multiplying the response from the first question with the confidence rating. Results are summarized in Fig. 6. Overall, the three systems have similar VSS, and AC-L1 pairs received an average VSS between –3.29 to –3.62, indicating that listeners were "*confident*" that the AC utterances had a different voice quality from those of the native speaker. Likewise, AC-L2 pairs received an average VSS between 3.50 to 4.07, indicating that listeners were "*confident*" that the same speaker produced the AC and L2 utterances. When analyzing the AC-L1 pairs, we found no significant differences in VSS between AC-PPG and either baseline (AC-PPG:AC-SIM $t(25)=1.13$, $p=0.27$; AC-PPG:AC-DTW, $t(25)=1.95$, $p=0.06$; two-tail). These results suggest that the three methods are equivalent in terms of producing speech that is different from the native speaker. When analyzing AC-L2 pairs, we found no significant difference between AC-PPG and AC-SIM ($t(25)=0.42$, $p=0.68$, two-tail), suggesting that the new accent conversion algorithm did not sacrifice the speaker's voice quality. However, AC-DTW achieved a higher VSS (4.07) than AC-PPG (3.50); one-tail t-test ($t(25)=3.59$, $p\ll 0.05$). One possible explanation for this result is that listeners still picked up subtle cues of non-native accent in the AC-DTW speech samples, and used it to rate voice quality. Because AC-DTW only performs voice conversion, it retains some of the non-native speaker's accent. This residual non-native accent may have led listeners to rate samples from AC-DTW as more similar to the non-native speech, even though the recordings were played backwards. This explanation is consistent with prior studies [51], [52] showing that, even when speech is played backwards, native English speakers can still detect non-native English accents.

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 250

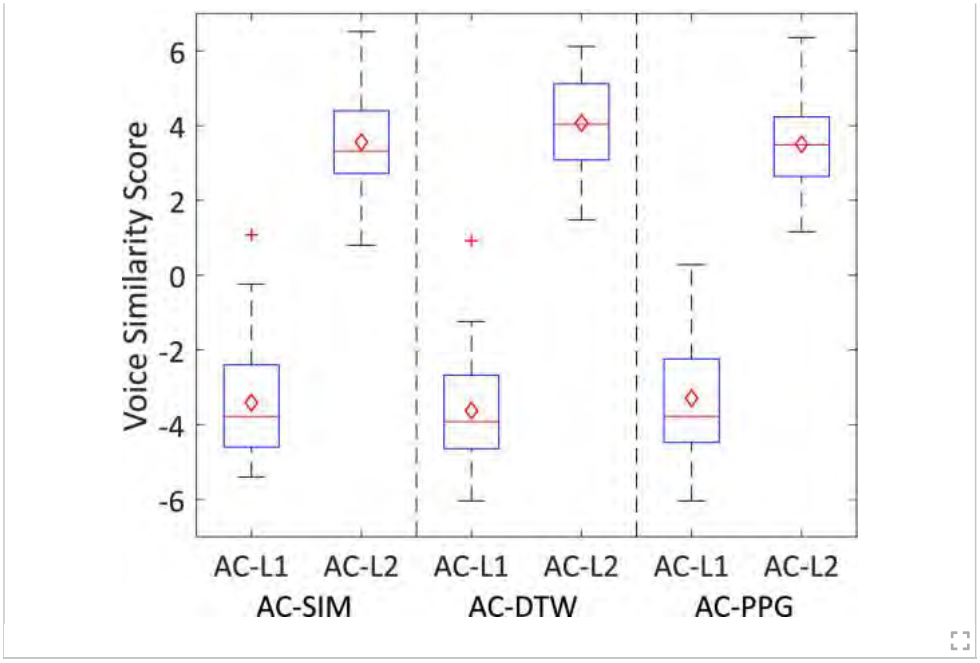


Fig. 6.

Voice quality results; AC-L1: VSS between AC and native (L1) speaker; AC-L2: VSS between AC and non-native (L2) speaker; the middle bars in the boxes show the median values and diamond markers ($\diamond$) show the mean values, the plus signs (+) indicate outliers, those notations apply to all boxplots in this paper.

Non-native accentedness: We used a preference test to determine if AC-PPG does indeed make the converted speech sound more native-like. Thirty native English speakers rated 150 utterance pairs: 50 pairs for each comparison: AC-PPG vs. AC-SIM, AC-PPG vs. AC-DTW, and AC-PPG vs. L2 (i.e., original utterances from the non-native speaker), ten pairs of utterances per conversion direction, each utterance pair was from the same sentence. Listeners were asked to choose the most native-like (least foreign) utterance from each pair, and then rate their confidence level using a seven-point scale (1-Not confident at all, 7-Very confident). Aryal and Gutierrez [12] had previously established that AC-SIM outperforms AC-DTW and L2 in this task; therefore, we omitted those comparisons in this study.

In a first analysis, we sought to determine if a particular system was preferred as "less-accented" and compared the *preference ratings* from the participants. Results are summarized in Fig. 7 (a). On average, listeners were very confident (mean: 98%, STD: 3%) that the AC-PPG conversions were more native-like than the original non-native utterances. More importantly, listeners were positive that AC-PPG outperformed both AC-SIM (mean: 69%, STD: 11%) and AC-DTW (mean: 72%, STD: 10%). All the above preference scores are statistically significant ($p\ll 0.001$; one-tail) compared with chance levels (50%). Since preference tests sometimes are too coarse and will mask out nuances in raters' attitudes, we further used the *confidence ratings* to compute a more detailed measurement – the cumulative confidence score (CCS) [53]. The CCS for each system in each comparison pair was computed as follows. We treat each response as if it were assigning a number of points to a system; for example, if a listener preferred the AC-PPG system and was "somewhat confident" (rated as three), then the AC-PPG system would receive three points. We then computed the average CCS that listeners allocated to each system. Therefore, the highest score a system can get is 350 points (7 × 50), within a comparison pair. Results are summarized in Fig. 7(b). As shown, all comparison pairs have the same trend as in the preference test, with AC-PPG performing significantly better than both baselines. All differences in CCS were statistically significant ($p\ll 0.001$, one-tail).

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 251

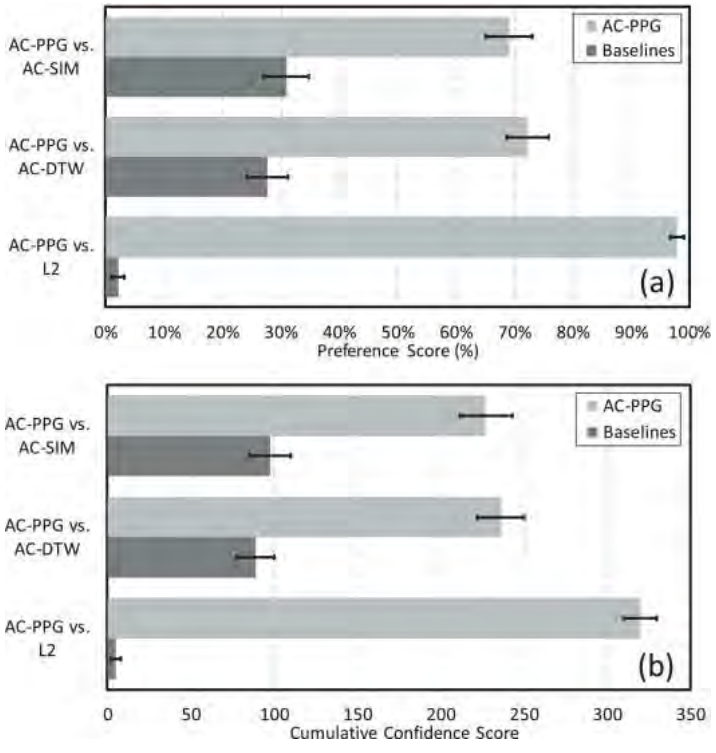


Fig. 7.
(a) Accent preference score with 95% confidence interval. (b) Cumulative confidence score for accentedness with 95% confidence interval.

B. Experiment 2: Native to Non-Native Conversion

In a second experiment, we evaluated whether AC-PPG can perform the accent conversion task in the opposite direction – creating a voice that has the native speaker's voice quality, but speaking with a non-native accent. Prior work [54] has tackled this problem from a Text-To-Speech perspective, so we wanted to determine if it could also be achieved through accent conversion. Accordingly, for this experiment we performed accent conversion in five directions that were from non-native to native English speakers, i.e., RRBI to BDL, HKK to BDL, YKWK to BDL, ABA to BDL, and TNI to CLB. The training and testing data for all speakers was identical as that used for *Experiment 1*.

In an initial listening test, we recruited 20 subjects to rate the non-native English accent of the converted speech using a nine-point Likert-scale rating test [2], where 1 corresponded to "no accent" and 9 to "very strong accent." For each conversion direction, we randomly picked five utterances, and we made sure that the final 25 (5 × 5) utterances for evaluation were from different elicitation sentences. To provide a reference, we also included the same set of sentences that were uttered by the native and non-native speaker in the test. Therefore, all listeners rated 75 (25 × 3) sentences. Given that our native speakers (BDL and CLB) spoke American English, before the test, we instructed listeners to consider that "*A 'foreign accent' is defined as an accent that is different from the General American English accent*." We also provided two samples of American accent English that were produced by native speakers not used in this study. All listeners were geographically located in the United States and all but one listeners self-reported to be native English speakers. The only listener whose native language is not English is a native Italian speaker who also speaks English and French, and since this participant passed our American accent pretest, we did not exclude this participant's responses. Results are summarized in Fig. 8. On average, listeners rated the native speech to be 1.4 points (closer to "no accent") and the non-native speech to be 6.4 points (closer to "very strong accent"). The accent-converted speech had an average rating of 6.2 points (closer to "very strong accent"), which was similar to the ratings of the non-native speech. No significant difference was found between accentedness ratings of non-native and AC speech ($t(19)=0.82$, $p=0.42$, two-tail). Therefore, this experiment indicates that our accent conversion approach was able to impart the non-native accent of the non-native speaker to utterances from a native speaker.

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 252

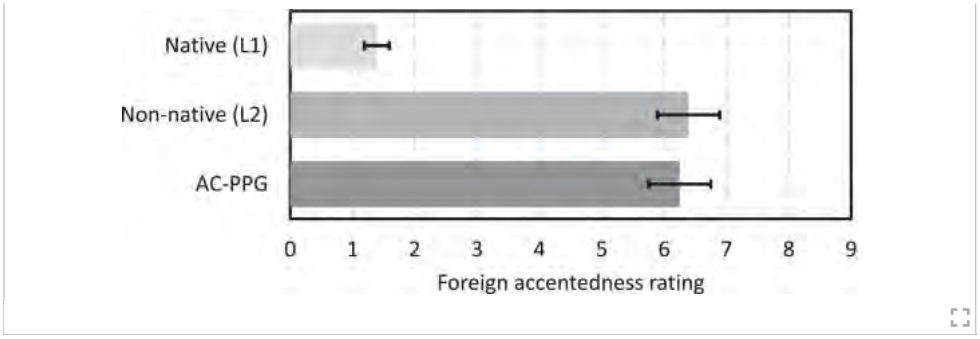


Fig. 8.
Foreign accentedness ratings for L1 (native English), L2 (non-native English), and AC speech; the error bars show 95% confidence intervals.

In a second listening test, we focused on evaluating whether the converted speech retained the voice quality of the original (native) speaker. Accordingly, we used the same VSS test as in *Experiment 1* to produce voice similarity scores between AC sentences and the original native/non-native sentences. Twenty listeners rated 50 utterance pairs, among which 25 were AC-L1, and the rest were AC-L2 pairs. As before, we randomized all presentation order and played the recordings in reverse. Results are summarized in Fig. 9. Listeners were "confident" that AC utterances had the voice quality of the native speakers (mean AC-L1 VSS score 2.87), and was different from the non-native speaker (mean AC-L2 VSS score -3.90) despite that they share the same accent. Considering the results from both listening tests in this experiment, we can conclude that AC-PPG is able to impart a non-native accent to native voices.

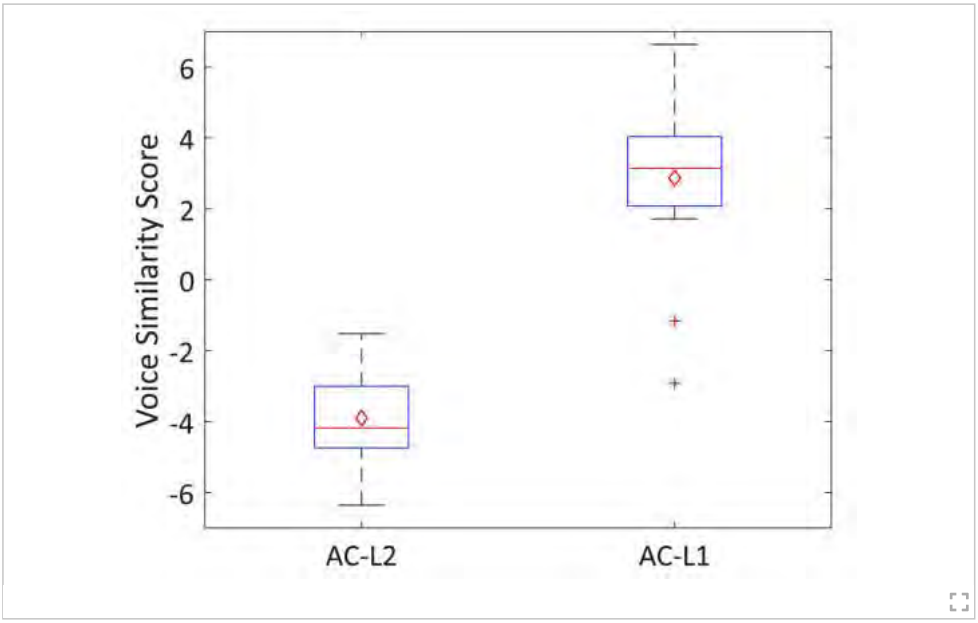


Fig. 9.
Voice similarity score for AC-L1 and AC-L2 comparisons.

C. Experiment 3: AC-PPG Using Non-Parallel Training Data

Our method does not impose timing constraints when pairing native and non-native speech frames: an acoustic frame from the native speaker is paired with a frame in the non-native speaker's training set by minimizing the symmetric KL divergence between their respective PPGs. Thus, in principle, our method removes the constraint that native and non-native speakers must produce the same set of utterances. This property is particularly useful for real-world applications because it allows more flexibility when recording training sentences. Therefore, in a third and final experiment we evaluated the AC performance by comparing two variants of our method:

- **AC-PPG-P**: the same system used in *Experiment 1*, i.e., using parallel sentences as the training data;

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 253

- **AC-PPG-NP**: a system that used non-parallel sentences. For this purpose, we randomly selected 100 native training utterances that were different from those in the non-native training or non-native test sentences. As a result, the native and non-native speakers never uttered any common sentence. All other configurations for this system were the same as AC-PPG-P.

The AC directions and test sentences were the same as those used in *Experiment 1*. For each system, we generated accent converted sentences from all 50 testing samples for evaluation.

In a first listening study, we used a preference test to determine which system yielded better acoustic quality. Twenty participants rated 50 utterance pairs – one from AC-PPG-P and the other from AC-PPG-NP, both utterances having the same linguistic content. We randomly selected 10 utterance pairs from each AC direction. For each pair, participants were asked to pick the utterance that has the best acoustic quality. The test allowed them to choose "no preference" as their response. Results are summarized in Fig. 10. The majority of the votes (40.3%) reflected no difference between the acoustic quality of the two systems ("no preference"), and both systems received a similar percentage of votes (29.7% for AC-PPG-P; 30.0% for AC-PPG-NP). We found no significant difference in terms of acoustic quality between using parallel or non-parallel data ($t(19)=0.11$, $p=0.91$, two-tail).

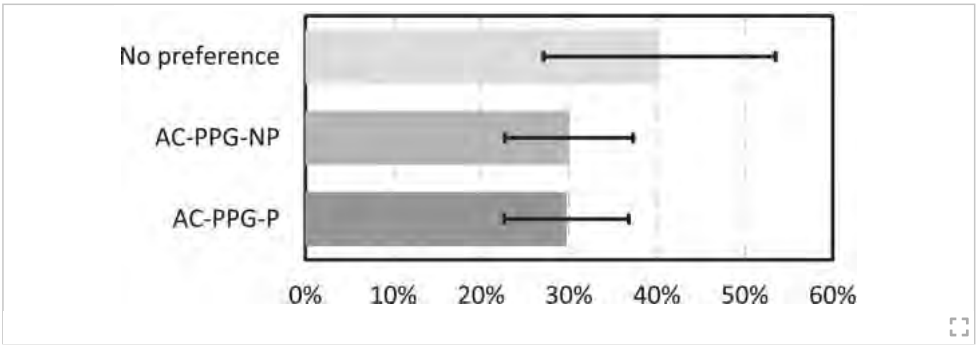


Fig. 10.
Preference scores for comparing the acoustic quality of AC-PPG-P and AC-PPG-NP; the error bars display the 95% confidence intervals.

In a second listening test, we investigated whether using non-parallel data would affect the non-native ratings of the converted speech. The experimental protocol was the same as the one we used in the acoustic quality experiment, except that in this case, for each AC-PPG-P and AC-PPG-NP utterance pair, we asked participants to select the one that had the "least foreign accent." Twenty participants rated 50 utterance pairs, 10 pairs for each AC direction. Results are summarized in Fig. 11. The vast majority of the votes (84.3%) indicated that there was no difference between the two systems. Furthermore, a t-test on the preference scores for AC-PPG-P (mean 7.7%) and AC-PPG-NP (mean 8.0%) revealed no significant differences ($t(19)=0.17$, $p=0.86$, two-tail).

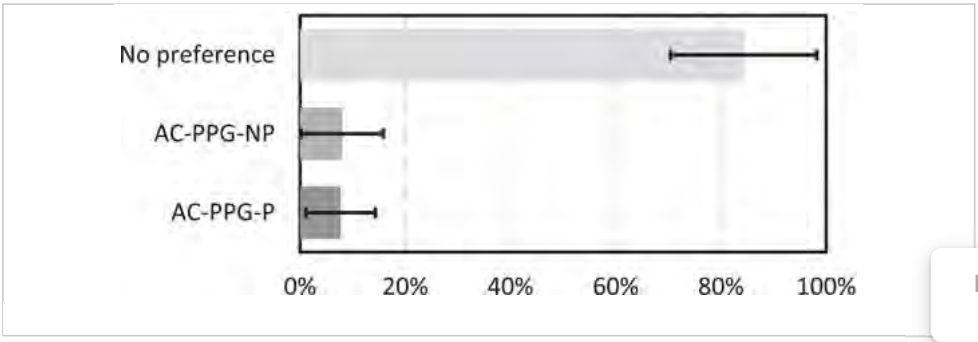


Fig. 11.
Preference scores for comparing foreign accentedness of AC-PPG-P and AC-PPG-NP; the error bars display the 95% confidence intervals.

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 254

Finally, we asked 21 listeners to rate AC-PPG-NP sentences in terms of voice quality. Each listener rated 50 converted utterances, where we randomly selected 10 utterances from all 5 conversion directions. The VSS scores are summarized in Fig. 12. The average AC-L1 VSS is -2.64 (std: 1.52), and 3.07 (std: 1.48) for AC-L2. Using a two-tail independent samples t-test assuming unequal variances, [8] we found no significant difference between the average VSS for the AC-PPG system in Fig. 6 and those in Fig. 12. For AC-L1, the test gave $t(43)=1.44$, $p=0.16$. For AC-L2, the test yielded $t(40)=1.06$, $p=0.29$. Thus, this experiment verified that using non-parallel data still allows our frame-pairing technique to preserve the non-native speaker's voice quality in the converted speech.

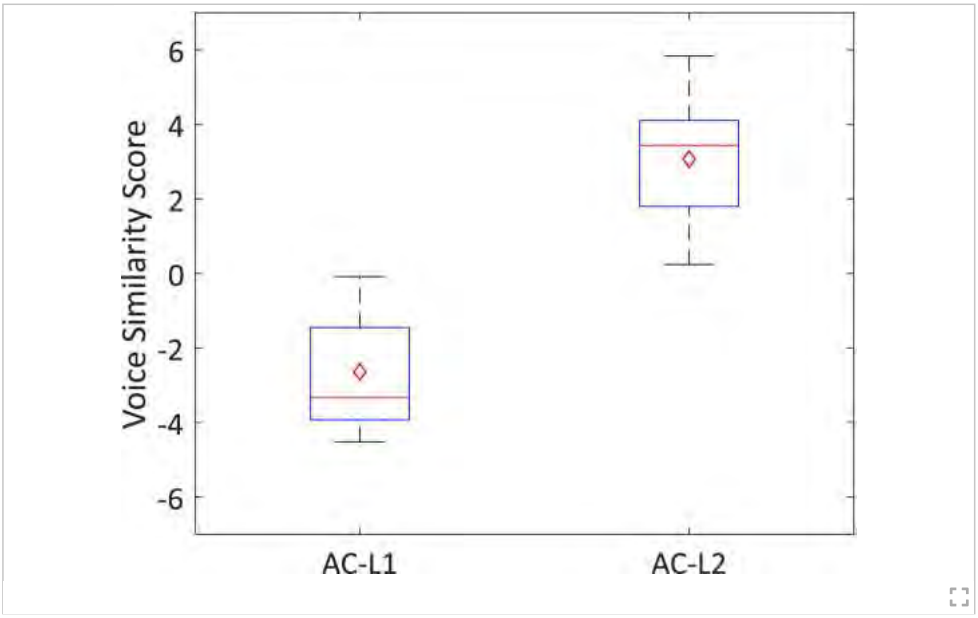


Fig. 12.
Voice similarity scores for AC-PPG-NP.

SECTION VI.
Discussion

In prior work [12], Aryal and Gutierrez-Osuna had shown that paring speech frames based on acoustic similarity (i.e., the AC-SIM baseline in our study), and then using the resulting frame pairs to train a voice conversion model could be used to create a voice that captured a native speaker's pronunciation and a non-native speaker's voice quality. Their method was able to achieve significantly better accentedness rating compared with pairing frames using DTW, though the results were based on a single pair of speakers. During our internal evaluations (results not shown) with multiple pairs of speakers and several set of non-native accents, we found that the speech generated by AC-SIM still contained noticeable mispronunciations. Since AC-SIM normalizes the vocal tract length difference between native and non-native speakers, we hypothesized that there remains a lot of other unattended speaker-dependent (SD) information in the VTLN-transformed acoustic feature space, which makes the resulting frame pairing not ideal. PPGs, on the other hand, are produced by speaker-independent (SI) acoustic models built for ASR. As a result, the most dominant information in PPGs is linguistic information. These analyses reinforced our intuition to use AC-PPG to eliminate the effects of SD cues in the frame pairing process.

The listening tests in *Experiment 1* show that the proposed frame pairing method can significantly reduce non-native accent ratings compared with two baselines. In terms of *voice similarity* between the non-native speaker and the converted speech, AC-PPG performs as well as AC-SIM. Although the speech generated by AC-DTW was rated more similar to the non-native speaker than AC-PPG, we suspected that it is hard to decouple the influence of *accent* and *voice quality* on the perceived *speaker identity* (refer to Section I for difference between *voice quality* and *speaker identity*). Listeners may have used the remaining foreign accent in the AC-DTW utterances to select the *speaker identity* of the utterances instead of their *voice quality*.

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 255

Therefore, an interesting future direction would be to design a new perceptual experiment protocol that can better decouple *voice quality* and *accent* in spoken sentences, compared with the current solution of playing audio in reverse.

Another interesting observation from *Experiment 1* is that despite using the same spectral conversion model as the two baseline systems, AC-PPG can significantly boost the acoustic quality of the synthesis. When comparing the speech syntheses from AC-PPG with the others, we did notice that there were fewer noises and artifacts. One possible explanation for this is that AC-PPG pairs frames with similar phonetic context. Therefore, frame pairs have similar spectral structures, making the statistical regression model for spectra estimation less likely to introduce odd shapes in the predicted spectral envelopes. Consequently, better spectral predictions lead to better synthesis quality. Future work could investigate if this property of AC-PPG generalizes to other statistical conversion models that take frame pairs as training input (e.g., deep neural networks [10], [29], direct waveform modification [9]).

Experiments 2 and *3* investigated other interesting aspects of the proposed frame-pairing method. *Experiment 2* verified that AC-PPG could also work in the opposite conversion direction – creating an artificial voice that has a native speaker's voice quality while speaking in a foreign accent. This artificial voice can be useful for generating materials for perceptual studies. For example, it can map speech from speakers that have different accents to the same voice quality, therefore removing the impact of voice quality when comparing differences in accents. *Experiment 3* verified that we could use non-parallel dataset to achieve the same accent conversion performance (measured in acoustic quality, accentedness, and voice quality) using AC-PPG. One possible reason why we could use non-parallel training data is that AC-PPG looks at a fine-grained context (95 ms in the current implementation),[9] and this context size is comparable with the duration of a vowel [55] or consonant [56] segment in American English. Therefore, as long as the two sets of training data from native and non-native speakers have a balanced phonetic distribution, the approach is indifferent to the actual word-level prompts. The non-parallel data constraint is much more relaxed than the widely used parallel constraint, making the proposed method applicable to real-world scenarios, where parallel data are scarce or tedious to obtain.

AC-PPG can run efficiently with careful optimization and GPU-based parallelization. In our experiments, it generally took no more than two minutes to compute the pairing between 100 training utterances ($\sim$5 minutes of speech) from the native and non-native speakers. Further reductions in computation time may be achieved via dimensionality reduction and clustering.

At present, our ratings of acoustic quality are on the low end of what state-of-the-art voice conversion systems can achieve [57]. This is largely due to the choice of voice conversion system used, i.e., a conventional GMM-based spectral conversion system as a case study, which was needed to ensure a fair comparison with our previous work [12]. Fortunately, our frame-pairing approach can be combined with other spectral conversion methods to produce higher quality speech synthesis. For example, instead of converting speech frame-by-frame, we could perform the conversion over a larger context (e.g., sequence to sequence conversion [58].) Using a larger conversion context is likely to increase the acoustic quality [59], [60]. More importantly, mispronunciations often occur at the segment level, which is beyond the scope of frame-level conversion, and contextual information has to be taken into consideration to accurately correct segmental pronunciation errors in accent conversion.

Another line of ongoing work in our group is to relax the non-parallel data constraint further to allow the use of cross-lingual training data. In preliminary experiments (not shown here), we successfully performed accent conversion using utterances recorded in the target speaker's native language to capture their voice quality.[10]

SECTION VII.
Conclusion

We have proposed a new frame-pairing method based on the phonetic similarity between acoustic frames. To measure phonetic similarity, we map source and target frames into a phonetic posteriorgram space using speaker-independent acoustic models trained on a native English corpus. Through a series of perceptual studies, we have shown that merely changing the frame pairing method can lead to significant improvement in acoustic quality and “nativeness,” while keeping the voice quality of the non-native speaker. Our results

PDF
Help

IPR2026-00272
Krisp Technologies EX1010 Page 256

also show that the approach works well across multiple non-native speakers with different native tongues. Additionally, the proposed algorithm does not need parallel data for training, which is ideal for real-world applications. Our approach only requires 5–10 minutes of speech data from the non-native speaker, making it practical for pronunciation training in realistic settings [63]. The implementation of the proposed system can be found at https://github.com/guanlongzhao/ppg-gmm.

ACKNOWLEDGMENT

The authors would like to thank A. Hair, S. Ding, and D. R. da Cunha Silva for reviewing early versions of the manuscript. The authors are grateful for the feedback from the reviewers, which helped improve the quality of this manuscript.

Authors

Figures

References

Citations

Keywords

Metrics

Footnotes

Back to Results | Next >

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

A public charity, IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

© Copyright 2025 IEEE - All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies.

PDF

Help

IPR2026-00272
Krisp Technologies EX1010 Page 257

Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion

Guanlong Zhao, *Student Member, IEEE*, and Ricardo Gutierrez-Osuna, *Senior Member, IEEE*

***Abstract*—Accent conversion (AC) aims to transform non-native utterances to sound as if the speaker had a native accent. This can be achieved by mapping source speech spectra from a native speaker into the acoustic space of the target non-native speaker. In prior work, we proposed an AC approach that matches frames between the two speakers based on their acoustic similarity after compensating for differences in vocal tract length. In this paper, we propose a new approach that matches frames between the two speakers based on their phonetic (rather than acoustic) similarity. Namely, we map frames from the two speakers into a phonetic posteriorgram using speaker-independent acoustic models trained on native speech. We thoroughly evaluate the approach on a speech corpus containing multiple native and non-native speakers. The proposed algorithm outperforms the prior approach, improving ratings of acoustic quality (22% increase in mean opinion score) and native accent (69% preference) while retaining the voice quality of the non-native speaker. Furthermore, we show that the approach can be used in the reverse conversion direction, i.e., generating speech with a native speaker's voice quality and a non-native accent. Finally, we show that this approach can be applied to non-parallel training data, achieving the same accent conversion performance.**

***Index Terms*—Accent conversion, voice conversion, acoustic model, posteriorgram.**

I. Introduction

LEARNERS who acquire a second language (L2) after a "critical period" [1] usually speak with a non-native accent. Having a non-native accent can often reduce the speaker's intelligibility [2] and may also lead to discriminatory attitudes [3], [4]. Therefore, non-native speakers have much to gain by improving their pronunciation. Several studies [5], [6] have shown that having a suitable native (L1) speaker to imitate – a so-called "golden speaker" with similar voice characteristics as the learner but with a native accent, can be beneficial in pronunciation training. Based on these findings, Felps *et al.* [7] suggested that such a "golden speaker" could be created by resynthesizing the non-native speaker's own voice with a native accent borrowed from a native reference speaker.

Manuscript received October 3, 2018; revised April 26, 2019; accepted June 24, 2019. Date of publication July 4, 2019; date of current version July 22, 2019. This work was supported by the NSF Awards 1619212 and 1623750. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Najim Dehak. *(Corresponding author: Guanlong Zhao.)*

The authors are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: gzhao@tamu.edu; rgutier@tamu.edu).

Digital Object Identifier 10.1109/TASLP.2019.2926754

Traditional voice-conversion (VC) methods [8]–[11] cannot be used for this purpose since VC cannot decouple the speaker's voice quality from her or his accent, i.e., VC assumes that accent is part of the speaker's identity. In this work, we distinguish two concepts: *voice quality*, which focuses on the physical characteristics of the speaker's voice (e.g., vocal tract and glottal configuration, pitch range), and *speaker identity*, a combination of *voice quality* and other speaker characteristics (e.g., accent, speaking rate, intonation, word choice).

To address the accent-and-voice-quality entanglement issue of traditional VC methods, Aryal and Gutierrez-Osuna [12] proposed a modified VC method where source frames (i.e., from the native reference speaker) and target frames (i.e., from the non-native speaker) were paired based on their acoustic similarity. In a first step, the authors applied vocal-tract length normalization (VTLN) to the source speech, so it matched the target speaker's vocal-tract length. Then, they paired each frame in the source corpus with the closest frame in the target corpus, and vice versa. Though VTLN did improve frame pairing compared to time alignment (i.e., the conventional approach in VC), vocal-tract length is just one of the potentially many differences between speakers, and it is too coarse to account for differences in pronunciation.

To address this issue, we present an approach that matches source and target frames based on their phonetic content. Leveraging advances in acoustic modeling [13], we extract phonetic information from phonetic posteriorgrams (PPGs) [14]. Namely, we compute the posteriorgram for each source and target speech frame through a speaker-independent acoustic model trained on a large corpus of native speech. Then, we use the symmetric Kullback-Leibler (KL) divergence [15] in posteriorgram space to match source and target frames. The result is a set of source-target frames that are paired based on their phonetic similarity, with which we train a Gaussian Mixture Model (GMM) to model the joint distribution of source and target Mel-Cepstral Coefficients (MCEPs). In a final step, we map source MCEPs into target MCEPs using maximum likelihood estimation of spectral parameter trajectories considering the global variance [8] of the target speaker. Our implementation is based on a conventional GMM spectral mapping method to ensure a fair comparison with the prior study [12], but our proposed frame matching method can be combined with any spectral mapping methods (e.g., neural networks, frequency warping) that take frame pairs as input.

Our approach differs from prior works on accent conversion, which modify speech features that carry accent information,

2329-9290 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 258

such as prosody, formants, spectral envelopes, or articulatory gestures [7], [16]–[18]. Instead, we use a VC technique to capture the voice quality of the (target) non-native speaker while preserving the (source) native speaker's pronunciation characteristics – both segmental and prosodic. Unlike VC methods, however, we avoid the issue of time aligning source and target utterances, which is problematic when the target speaker is non-native. Our approach is related to that of Xie *et al.* [19], who used speaker-adaptive acoustic models to generate posteriorgrams for VC. Their method groups all target speaker training data into phonetic clusters in the posteriorgram space using symmetric KL divergence and K-means clustering. Then, each frame of the source speaker's corpus is mapped to the centroid of the closest target phonetic cluster. The final converted speech is generated from those closest cluster centroids using the maximum probability trajectory generation algorithm. In contrast with their frame clustering approach, we use PPGs to produce frame pairs between source and target speakers, and then we train a GMM using those frame pairs. A second major difference with their approach is that we use speaker-independent acoustic models trained on native speech to ensure that the PPGs only reflect native pronunciations, whereas their approach uses speaker-adaptive training, which would introduce non-native pronunciations into the acoustic models. Initial findings from this work were presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2018 [20]. That earlier conference paper presented preliminary listening test results that verified the effectiveness of the PPG-based frame-matching method. The present manuscript describes our method in detail and significantly expands the perceptual studies and data analyses, including an experimental comparison of the proposed method on parallel and non-parallel data.

The manuscript is organized as follows. Section II reviews prior work on accent conversion in the acoustic and articulatory domains, and discusses the connection between accent conversion and voice conversion. Section III describes the experimental methods used in the study, including the phonetic posteriorgrams, the proposed frame-pairing method for accent conversion, as well as the baseline systems we used for comparison. Section IV describes the experimental setup used for the study, including the speech corpus we used to train the acoustic model, and the native/non-native speech corpus we used for accent conversion. Section V presents results on three different experiments we used to evaluate various aspects of the algorithm. The paper concludes with a discussion of our findings and directions for future work.

II. Literature Review

A. Algorithms for Accent Conversion

Foreign and non-native accents occur when speech deviates from the expected acoustic (e.g., formants) and prosodic (e.g., intonation, duration, and rate) norms of a language [7]. Therefore, prior work has focused on modifying certain speech characteristics to alter the perceived accent. In early work, Yan *et al.* [21] used a voice-morphing software to change the trajectories of formants, pitch, and duration to convert between three different English dialects (British, Australian, and General American English). The authors found that prosodic modifications produced noticeable differences on perceived accent, although not as significant as those produced by modifying formant trajectories. In the approach of Felps *et al.* [7], the spectral envelope of the non-native speech was replaced with that of the native speaker's, which had been normalized to the non-native speaker's vocal tract length with a piecewise linear warping function. Their results showed that the segmental correction was able to significantly reduce the foreign accentedness of the modified utterances. More recently, Jügler *et al.* [22] used PSOLA to correct the prosody of non-native German speech spoken by native French speakers. Prosodic (duration and pitch) corrections were performed at the syllable level, and the results showed a moderate but significant reduction in accentedness of the corrected speech.

A couple of studies also tried to blend native and non-native spectra to control the accent. Huckvale and Yanagisawa [17] blended the spectral envelope of non-native Japanese speech produced by an English Text-To-Speech (TTS) with its native counterpart through voice morphing to reduce the accent. Aryal *et al.* [18] decomposed the cepstrum into spectral slope and spectral detail, and then generated accent conversions by combining the spectral slope of the non-native speaker with a morph of the spectral detail of the native speaker. Though these spectra-blending methods can reduce non-native accents, they also tend to produce syntheses that are perceived as a "third speaker," one who is different from either the source (native) or target (non-native) speaker. To tackle this problem, Aryal and Gutierrez-Osuna [12] adapted VC techniques to perform accent conversion. The authors used vocal-tract-length normalization (VTLN) before pairing acoustic frames between source (native) and target (non-native) speaker, then built a GMM using those frame pairs to perform VC. This method was able to reduce non-native accent significantly, while retaining the non-native speaker's voice quality; however, it required a relatively large set of parallel recordings from the two speakers, and VTLN only accounted for a subset of the speaker characteristics.

An alternative to using acoustic methods is to operate in the articulatory domain. Along these lines, Felps *et al.* [16] used an articulatory synthesizer based on unit-selection to replace mispronounced non-native diphones with those from the non-native corpus that matched the articulatory configuration of a reference utterance from a native speaker. Later, Aryal and Gutierrez-Osuna used GMMs [23] and DNNs [24] to build an articulatory synthesizer (i.e., a mapping from articulatory gestures into acoustics) for the non-native speaker, then drove the GMM/DNN with articulatory gestures from a native speaker. Methods based on articulatory data generate syntheses that sound more like the non-native (target) speaker than acoustic methods, since they effectively decouple linguistic information (e.g., articulatory gestures from a native [source] speaker) from voice quality (captured by the articulatory-to-acoustic synthesizer of the non-native speaker). However, articulatory methods are expensive and require specialized equipment to collect articulatory data, so they are impractical for pronunciation training.

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 259

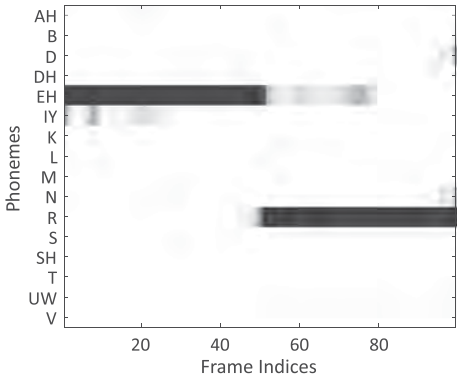


Fig. 1. PPG for the word "air," whose phonetic transcription in ARPABET is "EH R." For visualization purposes, we used a subset of the ARPABET phoneme set and omitted phonemes that had small values.

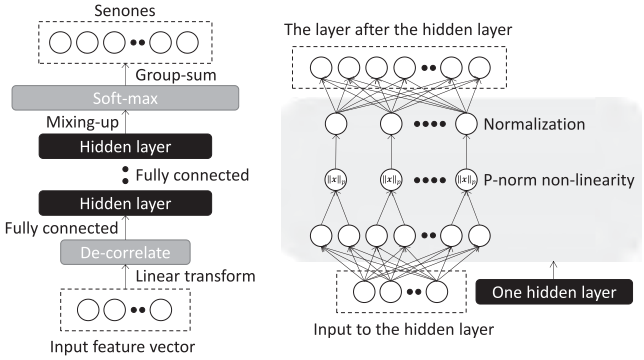


Fig. 2. P-norm deep neural network structure for acoustic modeling.

B. Connection Between Accent and Voice Conversion

Accent conversion is closely related to the problem of voice conversion [25]. Voice conversion transforms a source speaker's speech into that of a (known) target speaker. The conversion aims to match the voice characteristics of the target speaker, which may include vocal tract configuration, glottal characteristics, pitch range, pronunciation, and speaking rate. Ideally, the only information retained from the source speech is its linguistic content, i.e., the words that were uttered. Popular methods for voice conversion include joint-density GMMs [8], frequency warping [26], [27], DNNs [28], [29], and sparse coding [11], [30]–[32]. Accent conversion modifies speech at a finer level of granularity, and seeks to combine the linguistic content and pronunciation of the source speaker with the voice quality of the target speaker. Therefore, accent conversion is a more challenging problem than voice conversion in the sense that, first, there is no ground truth for the output voice, and second, accent conversion needs to split the speech into voice quality (converted) and accent (preserved), whereas voice conversion jointly converts both.

III. Methods

A. Phonetic Posteriorgrams

At its core, our proposed method relies on Phonetic Posteriorgrams (PPGs) to measure the similarity of speech frames across speakers. A phonetic posteriorgram is computed by segmenting speech into frames and computing the posterior probability that each frame belongs to a set of pre-defined phonetic units. As an example, Fig. 1 shows the PPG of the spoken word "air." In practice, it is advisable to include context when computing the PPG by concatenating each speech frame with its neighboring right and left frames. Moreover, phoneme labels are too coarse to describe the variety of speech sounds. Therefore, the dimensions in a phonetic posteriorgram are often associated with triphones, as we will see next.

Generally, the phonetic posteriorgram is computed from the acoustic model in an automatic speech recognizer (ASR). The acoustic model in ASR acts as a sequential classifier: given an input acoustic feature vector, the acoustic model assigns how likely it is that the vector belongs to each of a set of states/senones. In recent years, acoustic models based on DNNs have yielded state-of-the-art speech recognition accuracy [13]. The most advanced ASR systems can achieve Word Error Rates that are comparable to or better than expert human transcribers on specific tasks [33].

In this work, we compute phonetic posteriorgrams using a p-norm DNN [34] as the acoustic model. The input layer accepts a feature frame accompanied by its left and right neighbors; then the input is de-correlated by a fixed linear transformation [35]. The de-correlated features are then passed through N hidden layers, each employing the p-norm non-linearity $\boldsymbol{y} = \|\boldsymbol{x}\|_p = (\sum_i |x_i|^p)^{\frac{1}{p}}$, where $\boldsymbol{y}$ is one output dimension of a hidden layer and $\boldsymbol{x}$ represents a group of hidden neurons of that layer. Therefore, the number of output dimensions of each hidden layer is smaller than the number of hidden neurons. The output of the p-norm layer is then processed by a normalization layer to limit its standard deviation to one [34]. The output of the final hidden layer is fed into a softmax layer that produces more output nodes than the desired number of senones using a technique called "mixing-up" [34]. "Mixing-up" operates as follows. About halfway through training, the dimension of the softmax layer is increased by letting each output senone's probability be a sum over potentially multiple "mixture components." The mixture components are distributed using a power rule, proportional to the senone class priors. The neural network then "group-sums" the output of the softmax layer according to the group assignment defined in the "mixing-up" step, resulting in the final output nodes that correspond to individual senones. Fig. 2 shows the overall structure of the p-norm deep neural network that we use in this work.

During training, inputs to the p-norm DNN consist of stacked MFCC frames $\boldsymbol{X}$, whereas target outputs $\boldsymbol{Y}$ are senone labels obtained from force-alignment using an existing GMM-HMM speech recognizer. The training objective is the sum (across all frames of training data) of the log-probability of $\boldsymbol{Y}$ given $\boldsymbol{X}$: $\sum_i \log p(\boldsymbol{Y}_i|\boldsymbol{X}_i)$. After the DNN is fine-tuned using Stochastic Gradient Descent [36], we compute the posterior probability of observing senone l given the speech frame $\boldsymbol{x}$ by doing a complete

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 260

forward propagation,

$$p(l|\boldsymbol{x}) = \sum_{g \in G} \frac{\exp x_g^{'}}{\sum_k \exp x_k^{'}}, \tag{1}$$

where $x_k^{'}$ is the output of the hidden layer that precedes the softmax layer, and G is the set of softmax outputs that are grouped into senone l during the "mixing-up" procedure. A PPG frame of $\boldsymbol{x}$ is constructed by forming a vector from all possible values of $p(l|\boldsymbol{x})$, see eq. (2).

B. Frame Pairing

Conventional voice conversion methods use time alignment to pair frames from source and target utterances. As such, a VC model trained from time-aligned frame pairs will retain the non-native speaker's accent. Instead, to perform accent conversion, the pairing must be based on the phonetic similarity between source and target frames. In this way, each native speech frame is associated with its most similar non-native counterpart in terms of pronunciation. If we train a spectral conversion model between these frame pairs, the pronunciation from the native speech data will be preserved and the spectral envelope of the native speaker will be modified to match the non-native speaker's voice quality.

1) Frame Pairing Based on Phonetic Similarity (AC-PPG): We use PPGs to pair frames between the native and the non-native speaker. Our rationale is straightforward: if an ASR trained on native speech determines that a non-native speech segment $\boldsymbol{y}$ is close to the native speech production of a particular phoneme (or triphone, in our case), then it is reasonable to pair $\boldsymbol{y}$ with a native speech segment $\boldsymbol{x}$ with the same phonetic label; see Fig. 3 (a). Specifically, our approach works as follows. In a first step, we compute PPG frames for speech frames from the two speakers,

$$\mathcal{L}_{\boldsymbol{x}_i} = [p(l_1|\boldsymbol{x}_i), p(l_2|\boldsymbol{x}_i), \ldots, p(l_V|\boldsymbol{x}_i)], \tag{2}$$

where $\boldsymbol{x}_i$ is the acoustic feature vector of the i-th speech frame; $V = \{l_1, l_2, \ldots, l_V\}$ is the predefined senone set; $p(l_j|\boldsymbol{x}_i)$ is the conditional probability that the speech frame belongs to senone l_j given $\boldsymbol{x}_i$; $\sum_j p(l_j|\boldsymbol{x}_i) = 1$.

Given posterior feature vectors $\mathcal{L}_{\boldsymbol{x}_i}$ and $\mathcal{L}_{\boldsymbol{x}_j}$, we calculate their distance using the symmetric KL divergence,

$$D(\mathcal{L}_{\boldsymbol{x}_i}, \mathcal{L}_{\boldsymbol{x}_j}) = (\mathcal{L}_{\boldsymbol{x}_i} - \mathcal{L}_{\boldsymbol{x}_j}) \cdot (\log \mathcal{L}_{\boldsymbol{x}_i} - \log \mathcal{L}_{\boldsymbol{x}_j}). \tag{3}$$

The symmetric KL divergence [15] is commonly used to compute the similarity between distributions, and here, each frame of the PPG functions like a distribution. For each source (i.e., native) frame $\boldsymbol{x}_i$ we find its closest target (i.e., non-native) frame $\boldsymbol{y}_i^*$,

$$\boldsymbol{y}_i^* = \arg\min_{\forall \boldsymbol{y}} D(\mathcal{L}_{\boldsymbol{x}_i}, \mathcal{L}_{\boldsymbol{y}}). \tag{4}$$

Likewise, for each non-native frame $\boldsymbol{y}_i$ we find its closest native frame $\boldsymbol{x}_i^*$,

$$\boldsymbol{x}_i^* = \arg\min_{\forall \boldsymbol{x}} D(\mathcal{L}_{\boldsymbol{x}}, \mathcal{L}_{\boldsymbol{y}_i}). \tag{5}$$

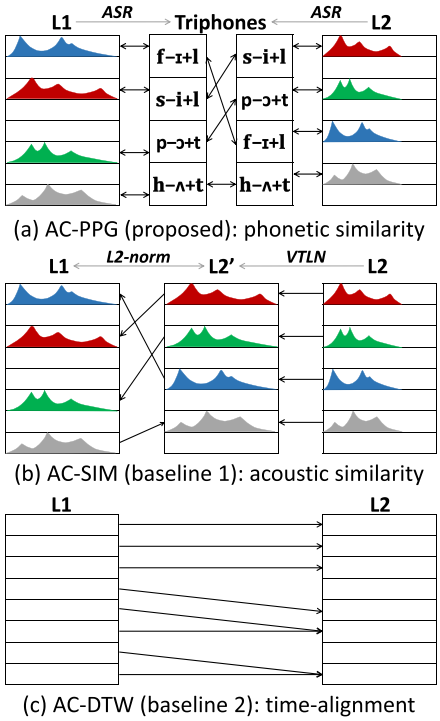


Fig. 3. L1: native, L2: non-native. (a) AC-PPG: proposed AC algorithm that uses phonetic similarity. (b) AC-SIM: Baseline 1 that uses acoustic similarity through VTLN to pair frames [12]. (c) AC-DTW: Baseline 2; native and non-native frames are time-aligned following their ordering in the data.

Each frame pairing process only involves two speakers – the given native and non-native speakers. The frame pairing does not constrain the search space. Therefore, it is possible to pair multiple frames from one speaker with the same frame from the other speaker. In this case, we duplicate that frame multiple times. The resulting frame pairs are used to train a Gaussian Mixture Model (GMM).

2) Baseline Methods for Frame Pairing: We compared the proposed PPG-based method against two baseline techniques for frame pairing: the acoustic similarity method of Aryal and Gutierrez-Osuna [12], and dynamic time warping.

Baseline 1*. (AC-SIM):* Following [12], we measured acoustic similarity as the inverse of the L2-norm between native and non-native speaker frames, after normalizing the native speaker to match the vocal tract length of the non-native speaker; see Fig. 3 (b).

In a first step, we learn a VTLN transform to reduce physiological differences in vocal tract between the two speakers. For this purpose, we time-align parallel training utterances of the two speakers, each utterance represented as a sequence of MFCCs. Following Panchapagesan and Alwan [37], we then learn a linear transform between the MFCCs of both speakers using ridge regression:

$$T^* = \arg\min_T \|\boldsymbol{x} - T\boldsymbol{y}\|^2 + \lambda \|T\|^2, \tag{6}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of MFCCs from the native and non-native speakers, respectively, and T^* is the VTLN transform.

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 261

Next, for each native vector $\boldsymbol{x}_i$ we find its closest non-native vector $\boldsymbol{y}_j^*$ as:

$$\boldsymbol{y}_j^* = \arg\min_{\forall \boldsymbol{y}} \|\boldsymbol{x}_i - T^*\boldsymbol{y}\|^2. \quad (7)$$

We repeat the process for each non-native vector $\boldsymbol{y}_i$ to find its closest match $\boldsymbol{x}_j^*$:

$$\boldsymbol{x}_j^* = \arg\min_{\forall \boldsymbol{x}} \|\boldsymbol{x} - T^*\boldsymbol{y}_i\|^2. \quad (8)$$

The above process results in a lookup table where each native and non-native frame in the database is paired with the closest one from the other speaker.

***Baseline 2**. (AC-DTW):* As our second baseline method, we use Dynamic Time Warping (DTW) [38] to time-align native and non-native frames, as illustrated in Fig. 3 (c).

We note that baselines 1 and 2 need parallel data for training, whereas the proposed method can operate on non-parallel data, as we shall see in Section V-C.

C. Spectral Conversion

To ensure a fair comparison between the three frame-pairing methods, we use a common spectral conversion technique to map a native source speaker's spectral features to match a non-native target speaker's voice quality. Following Toda *et al.* [8], we use a GMM to model the joint distribution of source and target frame pairs, and then use maximum likelihood parameter generation (MLPG) with global variance (GV) [39] to generate the converted speech for a given source utterance. Specifically, we use 2D-dimensional acoustic features, $X_t = [x_t^\top, \Delta x_t^\top]^\top$ from the source speaker, and $Y_t = [y_t^\top, \Delta y_t^\top]^\top$ from the target speaker, consisting of D-dimensional static and dynamic features, where $(\cdot)^\top$ denotes the transpose. Given the paired source and target features, we train a GMM to model the joint probability density $p(X, Y|\theta)$ where θ denotes model parameters, estimated using Expectation-Maximization (EM):

$$\theta = \text{EM}(\arg\max_{\theta} p(X, Y|\theta)). \quad (9)$$

When converting source static and dynamic feature vectors $X = [X_1^\top, X_2^\top, \ldots, X_T^\top]^\top$ to the target static feature vectors $y = [y_1^\top, y_2^\top, \ldots, y_T^\top]^\top$ – after the GMM is trained, we maximize the function below with respect to y,

$$\hat{y} = \arg\max_{y} \log\left(p(Y|X, \theta)^\omega p(\nu(y)|\theta_\nu)\right), \quad Y = Wy, \quad (10)$$

where $p(Y|X, \theta)$ denotes the conditional probability density function (PDF) on the target static and dynamic feature vectors, and $p(\nu(y)|\theta_\nu)$ represents the likelihood of a PDF on the global variance of the target feature vectors, which is represented as a separate GMM (one mixture) and trained using the EM algorithm as well. W is a matrix that appends dynamic features to the static features, and ω adjusts the relative importance between the two distributions and is set as the ratio of number of dimensions between vectors $\nu(y)$ and Y ($= 1/2T$). We use a GMM instead of a DNN in this study to focus on low-resource accent conversion scenarios – in real pronunciation training applications, we generally have a limited amount of data from the non-native speakers.

D. Pitch Scaling

Previous studies [7], [17], [21] have shown that prosody modification is an essential part of accent conversion, and the pitch contour contains identity-related information. Since pitch modification is not the focus of this study, we follow the standard procedure [8] and use the pitch trajectory from the source (native) speaker, which captures native intonation patterns, then normalize it to match the pitch range of the target (non-native) speaker using mean and variance normalization in the $\log F_0$ space.

IV. Experimental Setup

A. DNN Acoustic Model for Extracting PPG

To train the DNN acoustic model, we used Kaldi's Librispeech recipe.[1] The model is a p-norm DNN ($p = 2$), as introduced in the method section, with five hidden layers. We extracted 13-dim MFCC vectors with a 7-frame context, passed the concatenated 91-dim (13×7) MFCCs through a Linear Discriminant Analysis (LDA) to generate a 40-dim input feature vector, then concatenated nine frames of such 40-dim LDA features as the final input to the DNN. The 360-dim (40×9) input features were de-correlated using a fixed linear transform. All hidden layers had 5,000 hidden neurons and output 500 activations because each p-norm non-linearity was computed over ten hidden neurons. Every hidden layer was fully-connected with the previous layer. Right after the last hidden layer was a softmax layer of 14,000 nodes; those nodes were then "group-summed" to produce the final output across senones (5,816 dimensions, which were obtained from state-tying on a phonetic decision tree built from the transcripts of the training data; see [40] for more details on how the decision tree was constructed). The DNN acoustic model was trained on Librispeech's [41] training set, a speech recognition corpus that contains 960 hours of native English speech, the majority being American English. In the following experiments, the Librispeech corpus was used solely for building the acoustic model.

B. Speech Corpus for Accent Conversion

For the native speech synthesis corpus, we used two speakers from the CMU ARCTIC dataset [42]: BDL and CLB. Those recordings have a sampling rate of 16 KHz. For the non-native (L2) English speech synthesis corpus, we used five non-native speakers from the L2-ARCTIC corpus[2][43]: two native Hindi speakers, two native Korean speakers; and one native Arabic speaker. Each non-native speaker produced the full ARCTIC dataset (∼1100 utterances; around one hour of speech). The speech was recorded in a quiet room at 44.1 KHz. For the following experiments, we down-sampled all the non-native speech data to 16 KHz using sox.[3] The speaker demographic

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech
[2] https://psi.engr.tamu.edu/l2-arctic-corpus/
[3] http://sox.sourceforge.net/Main/HomePage

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 262

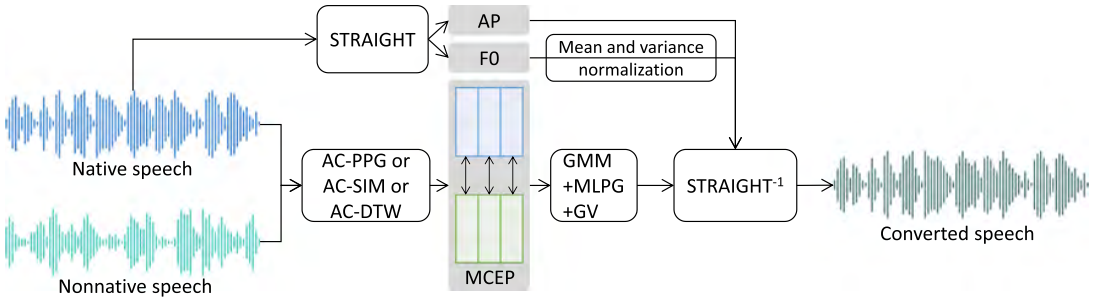


Fig. 4. Accent conversion workflow; frame pairing can be AC-PPG, AC-SIM (baseline 1), or AC-DTW (baseline 2).

TABLE I
DEMOGRAPHIC INFORMATION OF THE SPEAKERS

Speaker	Gender	Native Language	English Proficiency
BDL	M	English	Native
CLB	F	English	Native
RRBI	M	Hindi	91
TNI	F	Hindi	99
HKK	M	Korean	114
YKWK	M	Korean	N/A
ABA	M	Arabic	94-101

information is summarized in Table I. For the non-native speakers, their English proficiency level was measured in their TOEFL iBT scores[4] [46].

C. System Configuration

In what follows, we will refer to the proposed frame-pairing algorithm, baseline 1 (acoustic similarity), and baseline 2 (dynamic time warping) as **AC-PPG**, **AC-SIM**, and **AC-DTW**, respectively.

We used the TANDEM-STRAIGHT vocoder[5] [49] to decompose speech into aperiodicity (AP), F_0, and a 513-dim spectral envelope. Then, we computed 25-dim MFCCs[6] from the spectral envelopes to learn the VTLN transform and pair frames using acoustic similarity (AC-SIM); see Section III-B2. AC-DTW also used those MFCCs (excluding MFCC_0) to time-align a source speaker to a target speaker. AC-PPG used the 5816-dim PPGs extracted by the acoustic model to perform frame pairing.

We also computed 25-dim MCEPs from the spectral envelopes as the acoustic feature (excluding MCEP_0 since it is energy) to train the spectral conversion models (GMMs) and convert speech from the native speaker to the non-native speaker. MCEPs from the two speakers were frame paired using the three methods (AC-PPG, AC-SIM, AC-DTW) before being fed to the GMMs. Following Aryal and Gutierrez-Osuna [12], all GMMs had 128 mixture components with diagonal covariance matrices. Input features to the GMM include delta features, and therefore the joint feature vectors had 96 dimensions. Once we converted the native speaker's MCEPs to the non-native speaker's space, we reconstructed the spectrogram from the converted MCEPs (MCEP_0 being copied from the native speaker), and combined it with the native speaker's AP and normalized F_0 to synthesize speech using the TANDEM-STRAIGHT vocoder. The conversion pipeline is illustrated in Fig. 4.

All experiments were conducted on a desktop running Windows 10 with an Intel Core i7-7700 K CPU@4.2 GHz, 16 GB of memory, and an NVidia GTX 1070 GPU. Most of the algorithms were implemented and run on Matlab v9.3, except for the acoustic model and PPGs, which were computed using Kaldi on Ubuntu 16.04.

V. RESULTS

We conducted three sets of perceptual listening studies to evaluate different properties of the proposed frame-pairing algorithm. In the first experiment, we compared the approach against the two baseline systems by its ability to reduce perceived accents while matching the voice quality of the non-native speakers. In the second experiment, we evaluated whether the approach could also be used for the reverse purpose, i.e., to impart a non-native accent to a native speaker's voice. In the third and final experiment, we evaluated the approach to perform accent conversion using non-parallel speech corpora.

We recruited anonymous human participants from Amazon's Mechanical Turk platform[7] for our listening tests. Following Buchholz and Latorre [50], all listening tests included calibration trials designed to be easy to judge, and we used the participants' responses on those calibration trials to detect cheating behaviors. We excluded data from participants whose responses were below chance level on those calibration questions. All participants' calibration responses were excluded from the final analyses. In addition, and following [16], all human subjects passed a screening test that consisted of identifying various American English accents. We compensated participants for their time at an hourly rate of eight USD. In all experiments, the reference native and non-native English speech were resynthesized from their MCEPs using TANDEM-STRAIGHT to keep their acoustic quality comparable with the converted speech, which went through the same vocoder compression. When

[4]Speaker ABA only reported his IELTS [44] score (7.0). We converted it to a TOEFL iBT score following [45].

[5]We used the NDF F_0 extractor [47] instead of the default F_0 extractor that comes with TANDEM-STRAIGHT, because based on our experience and a prior study [48], the NDF F_0 extractor is more robust than the TANDEM-STRAIGHT default.

[6]We only used those MFCCs to generate the frame pairing lookup tables in **AC-SIM** and **AC-DTW** and discarded in other tasks.

[7]https://www.mturk.com/

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 263

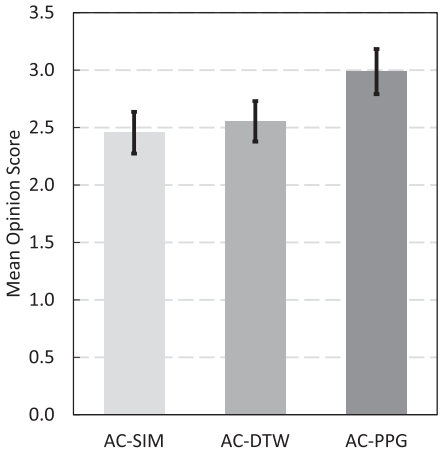


Fig. 5. Mean Opinion Scores for the proposed method (AC-PPG) and the two baseline methods (AC-SIM, AC-DTW); the error bars show 95% confidence intervals.

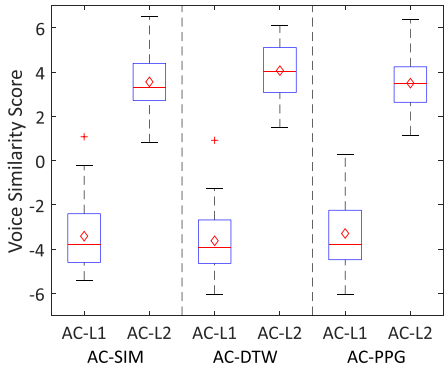


Fig. 6. Voice quality results; AC-L1: VSS between AC and native (L1) speaker; AC-L2: VSS between AC and non-native (L2) speaker; the middle bars in the boxes show the median values and diamond markers (⋄) show the mean values, the plus signs (+) indicate outliers, those notations apply to all boxplots in this paper.

selecting testing samples, we always randomly draw from the available pools, i.e., we did not cherry-pick the audio clips. All test trials were randomly presented. For any listening tests that required pairwise comparisons, the presentation order within an utterance pair was counterbalanced. Unless otherwise noted, we used paired-sample t-tests for the analyses.

A. Experiment 1: Comparing AC-PPG Against Baselines

In this experiment, we considered five native to non-native speaker pairings for accent conversion: BDL to RRBI, BDL to HKK, BDL to YKWK, BDL to ABA, and CLB to TNI. For each speaker pair, we used 100 parallel utterances for training and 50 utterances for testing; there was no overlap between the two sets. We performed accent conversion on all 50 test utterances using models trained on each of the three frame-pairing algorithms, i.e., AC-PPG, AC-SIM, and AC-DTW.

Acoustic quality: We used a standard five-point (1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent) Mean Opinion Score (MOS) to rate the acoustic quality of the synthesized speech. Thirty listeners rated 150 test samples: 50 per system, 10 utterances per conversion direction. Results are shown in Fig. 5. The proposed method (AC-PPG) received a MOS rating of 2.99, which was significantly higher than either baseline: AC-SIM (2.45 MOS, 22% relative improvement; $t(29) = 15.61$, $p \ll 0.001$; one-tail) and AC-DTW (2.55 MOS, 17% relative improvement; $t(29) = 12.04$, $p \ll 0.001$; one-tail). These results suggest that the proposed algorithm can boost the acoustic quality of the converted speech using exactly the same training data without even having to modify the GMM training and spectral conversion methods.

Voice quality: Following our prior work [32], we used a voice similarity score (VSS) ranging from -7 (definitely different speakers) to +7 (definitely same speaker) to assess the speaker's voice quality. Twenty-six participants rated 150 utterance pairs: 50 pairs per system (25 AC-L1 and 25 AC-L2 pairs, each pair contained one AC and one L1 [native]/L2 [non-native] utterance), and ten pairs per conversion direction. Following Felps *et al.* [7], we played utterances in reverse to prevent the accent from interfering with the perception of voice quality. In each trial, listeners first answered whether both utterances were produced by the same speaker (+1) or different speakers (-1), and then rated their confidence level on a 7-point scale (1-Not at all confident, 7-Extremely confident). The VSS was then compiled by multiplying the response from the first question with the confidence rating. Results are summarized in Fig. 6. Overall, the three systems have similar VSS, and AC-L1 pairs received an average VSS between -3.29 to -3.62, indicating that listeners were "*confident*" that the AC utterances had a different voice quality from those of the native speaker. Likewise, AC-L2 pairs received an average VSS between 3.50 to 4.07, indicating that listeners were "*confident*" that the same speaker produced the AC and L2 utterances. When analyzing the AC-L1 pairs, we found no significant differences in VSS between AC-PPG and either baseline (AC-PPG:AC-SIM $t(25) = 1.13$, $p = 0.27$; AC-PPG:AC-DTW, $t(25) = 1.95$, $p = 0.06$; two-tail). These results suggest that the three methods are equivalent in terms of producing speech that is different from the native speaker. When analyzing AC-L2 pairs, we found no significant difference between AC-PPG and AC-SIM ($t(25) = 0.42$, $p = 0.68$, two-tail), suggesting that the new accent conversion algorithm did not sacrifice the speaker's voice quality. However, AC-DTW achieved a higher VSS (4.07) than AC-PPG (3.50); one-tail t-test ($t(25) = 3.59$, $p \ll 0.05$). One possible explanation for this result is that listeners still picked up subtle cues of non-native accent in the AC-DTW speech samples, and used it to rate voice quality. Because AC-DTW only performs voice conversion, it retains some of the non-native speaker's accent. This residual non-native accent may have led listeners to rate samples from AC-DTW as more similar to the non-native speech, even though the recordings were played backwards. This explanation is consistent with prior studies [51], [52] showing that, even when speech is played backwards, native English speakers can still detect non-native English accents.

Non-native accentedness: We used a preference test to determine if AC-PPG does indeed make the converted speech

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 264

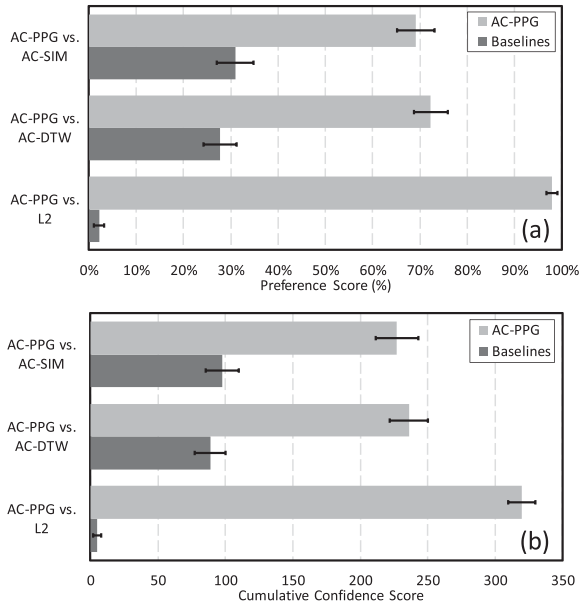


Fig. 7. (a) Accent preference score with 95% confidence interval. (b) Cumulative confidence score for accentedness with 95% confidence interval.

sound more native-like. Thirty native English speakers rated 150 utterance pairs: 50 pairs for each comparison: AC-PPG vs. AC-SIM, AC-PPG vs. AC-DTW, and AC-PPG vs. L2 (i.e., original utterances from the non-native speaker), ten pairs of utterances per conversion direction, each utterance pair was from the same sentence. Listeners were asked to choose the most native-like (least foreign) utterance from each pair, and then rate their confidence level using a seven-point scale (1-Not confident at all, 7-Very confident). Aryal and Gutierrez [12] had previously established that AC-SIM outperforms AC-DTW and L2 in this task; therefore, we omitted those comparisons in this study.

In a first analysis, we sought to determine if a particular system was preferred as "less-accented" and compared the *preference ratings* from the participants. Results are summarized in Fig. 7 (a). On average, listeners were very confident (mean: 98%, STD: 3%) that the AC-PPG conversions were more native-like than the original non-native utterances. More importantly, listeners were positive that AC-PPG outperformed both AC-SIM (mean: 69%, STD: 11%) and AC-DTW (mean: 72%, STD: 10%). All the above preference scores are statistically significant ($p \ll 0.001$; one-tail) compared with chance levels (50%). Since preference tests sometimes are too coarse and will mask out nuances in raters' attitudes, we further used the *confidence ratings* to compute a more detailed measurement – the cumulative confidence score (CCS) [53]. The CCS for each system in each comparison pair was computed as follows. We treat each response as if it were assigning a number of points to a system; for example, if a listener preferred the AC-PPG system and was "somewhat confident" (rated as three), then the AC-PPG system would receive three points. We then computed the average CCS that listeners allocated to each system. Therefore, the highest score a system can get is 350 points (7×50), within a comparison pair.

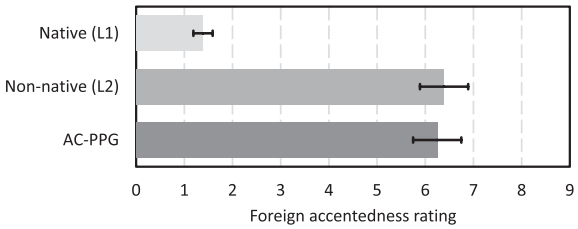


Fig. 8. Foreign accentedness ratings for L1 (native English), L2 (non-native English), and AC speech; the error bars show 95% confidence intervals.

Results are summarized in Fig. 7(b). As shown, all comparison pairs have the same trend as in the preference test, with AC-PPG performing significantly better than both baselines. All differences in CCS were statistically significant ($p \ll 0.001$, one-tail).

B. Experiment 2: Native to Non-Native Conversion

In a second experiment, we evaluated whether AC-PPG can perform the accent conversion task in the opposite direction – creating a voice that has the native speaker's voice quality, but speaking with a non-native accent. Prior work [54] has tackled this problem from a Text-To-Speech perspective, so we wanted to determine if it could also be achieved through accent conversion. Accordingly, for this experiment we performed accent conversion in five directions that were from non-native to native English speakers, i.e., RRBI to BDL, HKK to BDL, YKWK to BDL, ABA to BDL, and TNI to CLB. The training and testing data for all speakers was identical as that used for *Experiment 1*.

In an initial listening test, we recruited 20 subjects to rate the non-native English accent of the converted speech using a nine-point Likert-scale rating test [2], where 1 corresponded to "no accent" and 9 to "very strong accent." For each conversion direction, we randomly picked five utterances, and we made sure that the final 25 (5×5) utterances for evaluation were from different elicitation sentences. To provide a reference, we also included the same set of sentences that were uttered by the native and non-native speaker in the test. Therefore, all listeners rated 75 (25×3) sentences. Given that our native speakers (BDL and CLB) spoke American English, before the test, we instructed listeners to consider that "*A 'foreign accent' is defined as an accent that is different from the General American English accent*." We also provided two samples of American accent English that were produced by native speakers not used in this study. All listeners were geographically located in the United States and all but one listeners self-reported to be native English speakers. The only listener whose native language is not English is a native Italian speaker who also speaks English and French, and since this participant passed our American accent pretest, we did not exclude this participant's responses. Results are summarized in Fig. 8. On average, listeners rated the native speech to be 1.4 points (closer to "no accent") and the non-native speech to be 6.4 points (closer to "very strong accent"). The accent-converted speech had an average rating of 6.2 points (closer to "very strong accent"), which was similar to the ratings of the non-native speech. No significant difference was found

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 265

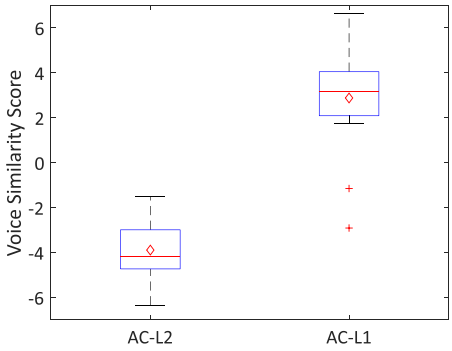


Fig. 9. Voice similarity score for AC-L1 and AC-L2 comparisons.

between accentedness ratings of non-native and AC speech ($t(19) = 0.82$, $p = 0.42$, two-tail). Therefore, this experiment indicates that our accent conversion approach was able to impart the non-native accent of the non-native speaker to utterances from a native speaker.

In a second listening test, we focused on evaluating whether the converted speech retained the voice quality of the original (native) speaker. Accordingly, we used the same VSS test as in *Experiment 1* to produce voice similarity scores between AC sentences and the original native/non-native sentences. Twenty listeners rated 50 utterance pairs, among which 25 were AC-L1, and the rest were AC-L2 pairs. As before, we randomized all presentation order and played the recordings in reverse. Results are summarized in Fig. 9. Listeners were "confident" that AC utterances had the voice quality of the native speakers (mean AC-L1 VSS score 2.87), and was different from the non-native speaker (mean AC-L2 VSS score -3.90) despite that they share the same accent. Considering the results from both listening tests in this experiment, we can conclude that AC-PPG is able to impart a non-native accent to native voices.

C. Experiment 3: AC-PPG Using Non-Parallel Training Data

Our method does not impose timing constraints when pairing native and non-native speech frames: an acoustic frame from the native speaker is paired with a frame in the non-native speaker's training set by minimizing the symmetric KL divergence between their respective PPGs. Thus, in principle, our method removes the constraint that native and non-native speakers must produce the same set of utterances. This property is particularly useful for real-world applications because it allows more flexibility when recording training sentences. Therefore, in a third and final experiment we evaluated the AC performance by comparing two variants of our method:

- **AC-PPG-P**: the same system used in *Experiment 1,* i.e., using parallel sentences as the training data;
- **AC-PPG-NP**: a system that used non-parallel sentences. For this purpose, we randomly selected 100 native training utterances that were different from those in the non-native training or non-native test sentences. As a result, the native and non-native speakers never uttered any common sentence. All other configurations for this system were the same as AC-PPG-P.

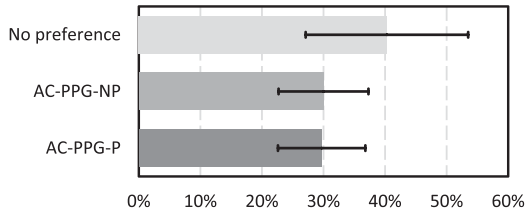


Fig. 10. Preference scores for comparing the acoustic quality of AC-PPG-P and AC-PPG-NP; the error bars display the 95% confidence intervals.

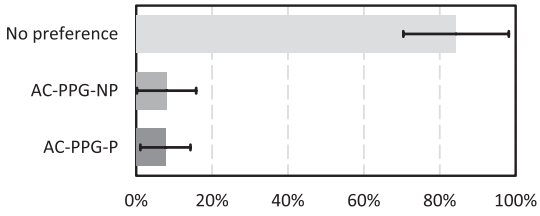


Fig. 11. Preference scores for comparing foreign accentedness of AC-PPG-P and AC-PPG-NP; the error bars display the 95% confidence intervals.

The AC directions and test sentences were the same as those used in *Experiment 1*. For each system, we generated accent converted sentences from all 50 testing samples for evaluation.

In a first listening study, we used a preference test to determine which system yielded better acoustic quality. Twenty participants rated 50 utterance pairs – one from AC-PPG-P and the other from AC-PPG-NP, both utterances having the same linguistic content. We randomly selected 10 utterance pairs from each AC direction. For each pair, participants were asked to pick the utterance that has the best acoustic quality. The test allowed them to choose "no preference" as their response. Results are summarized in Fig. 10. The majority of the votes (40.3%) reflected no difference between the acoustic quality of the two systems ("no preference"), and both systems received a similar percentage of votes (29.7% for AC-PPG-P; 30.0% for AC-PPG-NP). We found no significant difference in terms of acoustic quality between using parallel or non-parallel data ($t(19) = 0.11$, $p = 0.91$, two-tail).

In a second listening test, we investigated whether using non-parallel data would affect the non-native ratings of the converted speech. The experimental protocol was the same as the one we used in the acoustic quality experiment, except that in this case, for each AC-PPG-P and AC-PPG-NP utterance pair, we asked participants to select the one that had the "least foreign accent." Twenty participants rated 50 utterance pairs, 10 pairs for each AC direction. Results are summarized in Fig. 11. The vast majority of the votes (84.3%) indicated that there was no difference between the two systems. Furthermore, a t-test on the preference scores for AC-PPG-P (mean 7.7%) and AC-PPG-NP (mean 8.0%) revealed no significant differences ($t(19) = 0.17$, $p = 0.86$, two-tail).

Finally, we asked 21 listeners to rate AC-PPG-NP sentences in terms of voice quality. Each listener rated 50 converted utterances, where we randomly selected 10 utterances from all 5 conversion directions. The VSS scores are summarized in Fig. 12. The average AC-L1 VSS is -2.64 (std: 1.52), and 3.07 (std: 1.48)

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 266

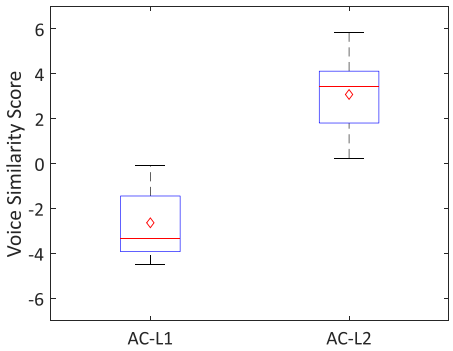


Fig. 12. Voice similarity scores for AC-PPG-NP.

for AC-L2. Using a two-tail independent samples t-test assuming unequal variances,[8] we found no significant difference between the average VSS for the AC-PPG system in Fig. 6 and those in Fig. 12. For AC-L1, the test gave $t(43) = 1.44$, $p = 0.16$. For AC-L2, the test yielded $t(40) = 1.06$, $p = 0.29$. Thus, this experiment verified that using non-parallel data still allows our frame-pairing technique to preserve the non-native speaker's voice quality in the converted speech.

VI. Discussion

In prior work [12], Aryal and Gutierrez-Osuna had shown that paring speech frames based on acoustic similarity (i.e., the AC-SIM baseline in our study), and then using the resulting frame pairs to train a voice conversion model could be used to create a voice that captured a native speaker's pronunciation and a non-native speaker's voice quality. Their method was able to achieve significantly better accentedness rating compared with pairing frames using DTW, though the results were based on a single pair of speakers. During our internal evaluations (results not shown) with multiple pairs of speakers and several set of non-native accents, we found that the speech generated by AC-SIM still contained noticeable mispronunciations. Since AC-SIM normalizes the vocal tract length difference between native and non-native speakers, we hypothesized that there remains a lot of other unattended speaker-dependent (SD) information in the VTLN-transformed acoustic feature space, which makes the resulting frame pairing not ideal. PPGs, on the other hand, are produced by speaker-independent (SI) acoustic models built for ASR. As a result, the most dominant information in PPGs is linguistic information. These analyses reinforced our intuition to use AC-PPG to eliminate the effects of SD cues in the frame pairing process.

The listening tests in *Experiment 1* show that the proposed frame pairing method can significantly reduce the non-native accent ratings compared with two baselines. In terms of *voice similarity* between the non-native speaker and the converted speech, AC-PPG performs as well as AC-SIM. Although the speech generated by AC-DTW was rated more similar to the non-native speaker than AC-PPG, we suspected that it is hard to decouple the influence of *accent* and *voice quality* on the perceived *speaker identity* (refer to Section I for difference between *voice quality* and *speaker identity*). Listeners may have used the remaining foreign accent in the AC-DTW utterances to select the *speaker identity* of the utterances instead of their *voice quality*. Therefore, an interesting future direction would be to design a new perceptual experiment protocol that can better decouple *voice quality* and *accent* in spoken sentences, compared with the current solution of playing audio in reverse.

Another interesting observation from *Experiment 1* is that despite using the same spectral conversion model as the two baseline systems, AC-PPG can significantly boost the acoustic quality of the synthesis. When comparing the speech syntheses from AC-PPG with the others, we did notice that there were fewer noises and artifacts. One possible explanation for this is that AC-PPG pairs frames with similar phonetic context. Therefore, frame pairs have similar spectral structures, making the statistical regression model for spectra estimation less likely to introduce odd shapes in the predicted spectral envelopes. Consequently, better spectral predictions lead to better synthesis quality. Future work could investigate if this property of AC-PPG generalizes to other statistical conversion models that take frame pairs as training input (e.g., deep neural networks [10], [29], direct waveform modification [9]).

Experiments 2 and *3* investigated other interesting aspects of the proposed frame-pairing method. *Experiment 2* verified that AC-PPG could also work in the opposite conversion direction – creating an artificial voice that has a native speaker's voice quality while speaking in a foreign accent. This artificial voice can be useful for generating materials for perceptual studies. For example, it can map speech from speakers that have different accents to the same voice quality, therefore removing the impact of voice quality when comparing differences in accents. *Experiment 3* verified that we could use non-parallel dataset to achieve the same accent conversion performance (measured in acoustic quality, accentedness, and voice quality) using AC-PPG. One possible reason why we could use non-parallel training data is that AC-PPG looks at a fine-grained context (95 ms in the current implementation),[9] and this context size is comparable with the duration of a vowel [55] or consonant [56] segment in American English. Therefore, as long as the two sets of training data from native and non-native speakers have a balanced phonetic distribution, the approach is indifferent to the actual word-level prompts. The non-parallel data constraint is much more relaxed than the widely used parallel constraint, making the proposed method applicable to real-world scenarios, where parallel data are scarce or tedious to obtain.

AC-PPG can run efficiently with careful optimization and GPU-based parallelization. In our experiments, it generally took no more than two minutes to compute the pairing between 100 training utterances (∼5 minutes of speech) from the native and

[8]The two groups we are comparing have 26 (AC-PPG in Fig. 6) and 21 (AC-PPG-NP) subjects respectively, therefore, it is not reasonable to assume that they have the same variance.

[9]Each frame of PPG feature looks at a larger context than the analysis window (25 ms), because the input to the acoustic model consists of nine frames of adjacent LDA feature, and each frame was computed from seven consecutive MFCC feature vectors (25 ms). Therefore, the total context for a frame of PPG feature is $9 + 7 - 1 = 15$ consecutive analysis windows, which converts to 95 ms under a 5 ms window shift.

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 267

non-native speakers. Further reductions in computation time may be achieved via dimensionality reduction and clustering.

At present, our ratings of acoustic quality are on the low end of what state-of-the-art voice conversion systems can achieve [57]. This is largely due to the choice of voice conversion system used, i.e., a conventional GMM-based spectral conversion system as a case study, which was needed to ensure a fair comparison with our previous work [12]. Fortunately, our frame-pairing approach can be combined with other spectral conversion methods to produce higher quality speech synthesis. For example, instead of converting speech frame-by-frame, we could perform the conversion over a larger context (e.g., sequence to sequence conversion [58].) Using a larger conversion context is likely to increase the acoustic quality [59], [60]. More importantly, mispronunciations often occur at the segment level, which is beyond the scope of frame-level conversion, and contextual information has to be taken into consideration to accurately correct segmental pronunciation errors in accent conversion.

Another line of ongoing work in our group is to relax the non-parallel data constraint further to allow the use of cross-lingual training data. In preliminary experiments (not shown here), we successfully performed accent conversion using utterances recorded in the target speaker's native language to capture their voice quality.[10]

VII. Conclusion

We have proposed a new frame-pairing method based on the phonetic similarity between acoustic frames. To measure phonetic similarity, we map source and target frames into a phonetic posteriorgram space using speaker-independent acoustic models trained on a native English corpus. Through a series of perceptual studies, we have shown that merely changing the frame pairing method can lead to significant improvement in acoustic quality and "nativeness," while keeping the voice quality of the non-native speaker. Our results also show that the approach works well across multiple non-native speakers with different native tongues. Additionally, the proposed algorithm does not need parallel data for training, which is ideal for real-world applications. Our approach only requires 5–10 minutes of speech data from the non-native speaker, making it practical for pronunciation training in realistic settings [63]. The implementation of the proposed system can be found at `https://github.com/guanlongzhao/ppg-gmm`.

Acknowledgment

The authors would like to thank A. Hair, S. Ding, and D. R. da Cunha Silva for reviewing early versions of the manuscript. The authors are grateful for the feedback from the reviewers, which helped improve the quality of this manuscript.

[10] In these preliminary experiments, we used native Brazilian Portuguese speakers from the SID dataset [61] as the target speakers. Since Portuguese share some phonological similarities with English [62], we used the acoustic model used in this study directly to produce the PPGs from native Portuguese speech. For future work and more general cases (e.g., languages from the Sino-Tibetan family), we have to include senones from the target speaker's native tongues in the acoustic modeling process.

References

[1] E. H. Lenneberg, "The biological foundations of language," *Hospital Pract.*, vol. 2, no. 12, pp. 59–67, 1967.
[2] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Lang. Learn.*, vol. 45, no. 1, pp. 73–97, 1995.
[3] D. L. Rubin and K. A. Smith, "Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants," *Int. J. Intercultural Relations*, vol. 14, no. 3, pp. 337–353, 1990.
[4] S. Lev-Ari and B. Keysar, "Why don't we believe non-native speakers? The influence of accent on credibility," *J. Exp. Social Psychol.*, vol. 46, no. 6, pp. 1093–1096, 2010.
[5] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors–in search of the golden speaker," *Speech Commun.*, vol. 37, no. 3–4, pp. 161–173, 2002.
[6] M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann, "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis," in *Proc. 11th Australian Int. Conf. Speech Sci. Technol.*, 2006, pp. 24–29.
[7] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Commun.*, vol. 51, no. 10, pp. 920–932, 2009.
[8] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
[9] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. Interspeech*, 2014, pp. 2514–2518.
[10] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 3, pp. 580–587, Mar. 2015.
[11] C. Liberatore, S. Aryal, Z. Wang, S. Polsley, and R. Gutierrez-Osuna, "SABR: Sparse, anchor-based representation of the speech signal," in *Proc. Interspeech*, 2015, pp. 608–612.
[12] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 7879–7883.
[13] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[14] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Nov. 2009, pp. 421–426.
[15] I. J. Taneja, "On generalized information measures and their applications," in *Advances in Electronics and Electron Physics*, vol. 76. Amsterdam, The Netherlands: Elsevier, 1989, pp. 327–413.
[16] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 8, pp. 2301–2312, Oct. 2012.
[17] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," in *Proc. 6th ISCA Workshop Speech Synthesis*, 2007, pp. 64–70.
[18] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *Proc. Interspeech*, 2013, pp. 3077–3081.
[19] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. Interspeech*, 2016, pp. 287–291.
[20] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2018, pp. 5314–5318.
[21] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis and synthesis of formant spaces of British, Australian, and American accents," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 676–689, Feb. 2007.
[22] J. Jügler, F. Zimmerer, J. Trouvain, and B. Möbius, "The perceptual effect of L1 prosody transplantation on L2 speech: The case of French accented German," in *Proc. Interspeech*, 2016, pp. 67–71.
[23] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *J. Acoust. Soc. Amer.*, vol. 137, no. 1, pp. 433–446, 2015.
[24] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Comput. Speech Lang.*, vol. 36, pp. 260–273, 2016.

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 268

[25] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, no. 88, pp. 65–82, 2017.
[26] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 922–931, Jul. 2010.
[27] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1313–1323, May 2012.
[28] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using speaker-dependent conditional restricted Boltzmann machine," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, p. 8, 2015.
[29] L. Sun, S. Kang, K. Li, and H. M. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4869–4873.
[30] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2012, pp. 313–317.
[31] Z. Wu, E. S. Chng, and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools Appl.*, vol. 74, no. 22, pp. 9943–9958, 2015.
[32] G. Zhao and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5525–5529.
[33] W. Xiong *et al.*, "Toward human parity in conversational speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2410–2423, Dec. 2017.
[34] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 215–219.
[35] S. P. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. Interspeech*, 2013, pp. 109–113.
[36] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Y. Lechevallier and G. Saporta, Eds., 2010, pp. 177–186.
[37] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 42–64, 2009.
[38] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining*, 1994, pp. 359–370.
[39] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 1, pp. 9–12.
[40] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Workshop Human Lang. Technol.*, 1994, pp. 307–312.
[41] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[42] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA ITRW Speech Synthesis*, 2004, pp. 223–224.
[43] G. Zhao *et al.*, "L2-ARCTIC: A non-native English speech corpus," in *Proc. Interspeech*, 2018, pp. 2783–2787.
[44] M. Chalhoub-Deville and C. E. Turner, "What to look for in ESL admission tests: Cambridge Certificate Exams, IELTS, and TOEFL," *System*, vol. 28, no. 4, pp. 523–539, 2000.
[45] ETS, "Linking TOEFL iBT scores to IELTS scores - A research report," Educational Testing Service, Princeton, NJ, USA, Tech. Rep., 2010. [Online]. Available: https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf
[46] Y. Cho and B. Bridgeman, "Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities," *Lang. Testing*, vol. 29, no. 3, pp. 421–442, 2012.
[47] H. Kawahara, A. de Cheveign, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech*, 2005, pp. 537–540.
[48] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.
[49] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 3933–3936.
[50] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. Interspeech*, 2011, pp. 3053–3056.
[51] "The detection of foreign accent in backwards speech," in *Proc. 15th Int. Congr. Phonetic Sci.*, 2010, pp. 535–538.
[52] M. J. Munro, T. M. Derwing, and C. S. Burgess, "Detection of nonnative speaker status from content-masked speech," *Speech Commun.*, vol. 52, pp. 626–637, 2010.
[53] J. Yamagishi, private communication, Calgary, AB, Canada, Apr. 2018.
[54] G. E. Henter, J. Lorenzo-Trueba, X. Wang, M. Kondo, and J. Yamagishi, "Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2018, pp. 4799–4803.
[55] N. Umeda, "Vowel duration in American English," *J. Acoust. Soc. Amer.*, vol. 58, no. 2, pp. 434–445, 1975.
[56] N. Umeda, "Consonant duration in American English," *J. Acoust. Soc. Amer.*, vol. 61, no. 3, pp. 846–858, 1977.
[57] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018, pp. 195–202.
[58] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds, 2014, pp. 3104–3112.
[59] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech*, 2017, pp. 1283–1287.
[60] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *Proc. Interspeech*, 2017, pp. 1268–1272.
[61] I. M. Quintanilha, L. W. P. Biscainho, and S. L. Netto, "Towards an end-to-end speech recognizer for Portuguese using deep neural networks," in *Proc. 35th Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, Sao Pedro, Brazil, 2017.
[62] M. M. Azevedo, "A contrastive phonology of Portuguese and English," *Modern Lang. J.*, vol. 66, no. 2, 1982, Art. no. 222.
[63] S. Ding *et al.*, "Golden speaker builder: an interactive online tool for L2 learners to build pronunciation models," in *Proc. Pronunciation 2nd Lang. Learn. Teaching*, 2017, pp. 25–26.

Authors' photographs and biographies not available at the time of publication.

Authorized licensed use limited to: Penn State University. Downloaded on December 19,2025 at 19:38:48 UTC from IEEE Xplore. Restrictions apply.

IPR2026-00272
Krisp Technologies EX1010 Page 269

Appendix CUTURI01

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

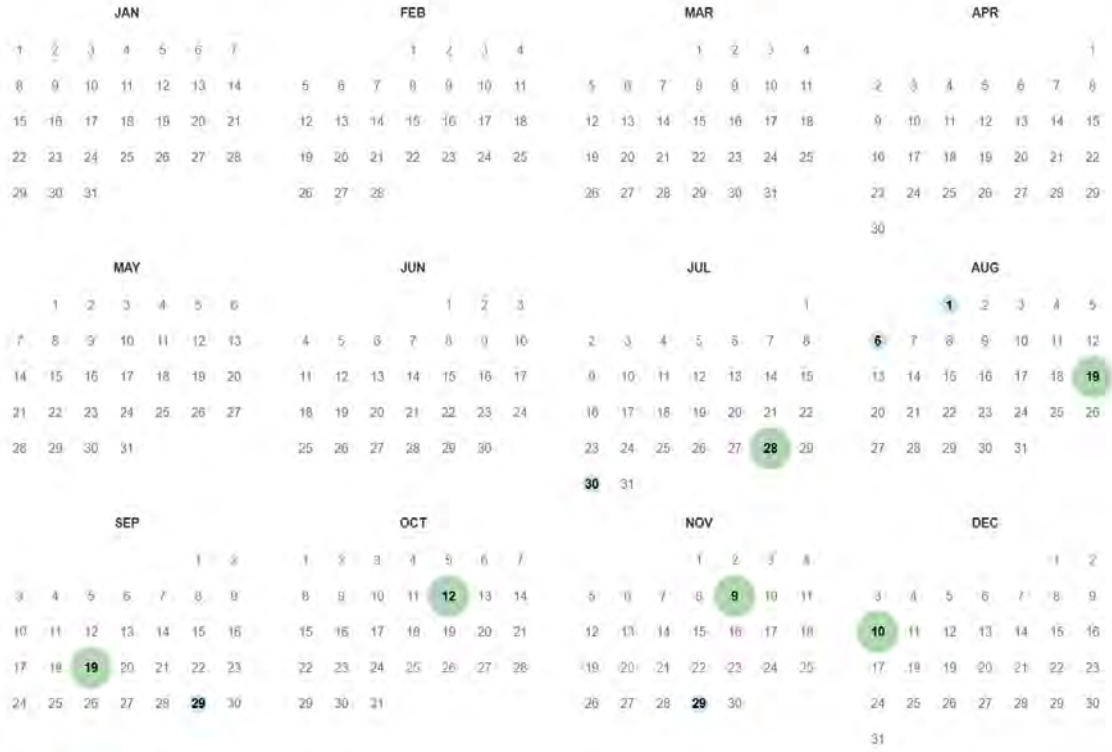
DONATE INTERNET ARCHIVE WayBackMachine Explore more than 1 trillion web pages saved over time

http://proceedings.mlr.press/v70

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **160 times** between July 28, 2017 and October 13, 2025.





Note

This calendar view maps the number of times **http://proceedings.mlr.press/v70** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ. Green indicates redirects (3xx).

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20170728161753/http://proceedings.mlr.press/v70 December 19, 2025
INTERNET ARCHIVE Wayback Machine http://proceedings.mlr.press/v70 Go JUN JUL SEP
160 captures 28 Jul 2017 - 15 Oct 2025 2016 28 2017 2018 About this capture

PMLR Proceedings of Machine Learning Research

Volume 70: International Conference on Machine Learning, 6-11 August 2017, International Convention Centre, Sydney, Australia

[edit]

Editors: Doina Precup, Yee Whye Teh

[bib][citeproc]

Filter Authors: Filter Titles:

Uncovering Causality from Multivariate Hawkes Integrated Cumulants
Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, Jean-François Muzy ; PMLR 70:1-10
[abs] [Download PDF][Supplementary PDF]

A Unified Maximum Likelihood Approach for Estimating Symmetric Properties of Discrete Distributions
Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Ananda Theertha Suresh ; PMLR 70:11-21
[abs] [Download PDF][Supplementary PDF]

Constrained Policy Optimization
Joshua Achiam, David Held, Aviv Tamar, Pieter Abbeel ; PMLR 70:22-31
[abs] [Download PDF][Supplementary PDF]

The Price of Differential Privacy for Online Learning
Naman Agarwal, Karan Singh ; PMLR 70:32-40
[abs] [Download PDF][Supplementary PDF]

Local Bayesian Optimization of Motor Skills
Riad Akrour, Dmitry Sorokin, Jan Peters, Gerhard Neumann ; PMLR 70:41-50
[abs] [Download PDF]

Connected Subgraph Detection with Mirror Descent on SDPs
Cem Aksoylar, Lorenzo Orecchia, Venkatesh Saligrama ; PMLR 70:51-59
[abs] [Download PDF][Supplementary PDF]

Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis
Ahmed M. Alaa, Scott Hu, Mihaela Schaar ; PMLR 70:60-69
[abs] [Download PDF][Supplementary PDF]

A Semismooth Newton Method for Fast, Generic Convex Programming
Alnur Ali, Eric Wong, J. Zico Kolter ; PMLR 70:70-79
[abs] [Download PDF][Supplementary PDF]

Learning Continuous Semantic Representations of Symbolic Expressions
Miltiadis Allamanis, Pankajan Chanthirasegaran, Pushmeet Kohli, Charles Sutton ; PMLR 70:80-88
[abs] [Download PDF][Supplementary PDF]

Natasha: Faster Non-Convex Stochastic Optimization via Strongly Non-Convex Parameter
Zeyuan Allen-Zhu ; PMLR 70:89-97
[abs] [Download PDF]

Doubly Accelerated Methods for Faster CCA and Generalized Eigendecomposition
Zeyuan Allen-Zhu, Yuanzhi Li ; PMLR 70:98-106
[abs] [Download PDF]

Faster Principal Component Regression and Stable Matrix Chebyshev Approximation
Zeyuan Allen-Zhu, Yuanzhi Li ; PMLR 70:107-115
[abs] [Download PDF]

Follow the Compressed Leader: Faster Online Learning of Eigenvectors and Faster MMWU
Zeyuan Allen-Zhu, Yuanzhi Li ; PMLR 70:116-125
[abs] [Download PDF]

Near-Optimal Design of Experiments via Regret Minimization
Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, Yining Wang ; PMLR 70:126-135
[abs] [Download PDF]

OptNet: Differentiable Optimization as a Layer in Neural Networks
Brandon Amos, J. Zico Kolter ; PMLR 70:136-145
[abs] [Download PDF][Supplementary PDF]

Input Convex Neural Networks
Brandon Amos, Lei Xu, J. Zico Kolter ; PMLR 70:146-155
[abs] [Download PDF][Supplementary PDF]

An Efficient, Sparsity-Preserving, Online Algorithm for Low-Rank Approximation

David Anderson, Ming Gu ; PMLR 70:156-165

[abs] [Download PDF][Supplementary PDF]

Modular Multitask Reinforcement Learning with Policy Sketches

Jacob Andreas, Dan Klein, Sergey Levine ; PMLR 70:166-175

[abs] [Download PDF][Supplementary PDF]

Averaged-DQN: Variance Reduction and Stabilization for Deep Reinforcement Learning

Oron Anschel, Nir Baram, Nahum Shimkin ; PMLR 70:176-185

[abs] [Download PDF][Supplementary PDF]

A Simple Multi-Class Boosting Framework with Theoretical Guarantees and Empirical Proficiency

Ron Appel, Pietro Perona ; PMLR 70:186-194

[abs] [Download PDF][Supplementary PDF]

Deep Voice: Real-time Neural Text-to-Speech

Sercan Ö. Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi ; PMLR 70:195-204

[abs] [Download PDF][Supplementary PDF]

Oracle Complexity of Second-Order Methods for Finite-Sum Problems

Yossi Arjevani, Ohad Shamir ; PMLR 70:205-213

[abs] [Download PDF][Supplementary PDF]

Wasserstein Generative Adversarial Networks

Martin Arjovsky, Soumith Chintala, Léon Bottou ; PMLR 70:214-223

[abs] [Download PDF][Supplementary PDF]

Generalization and Equilibrium in Generative Adversarial Nets (GANs)

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, Yi Zhang ; PMLR 70:224-232

[abs] [Download PDF][Supplementary PDF]

A Closer Look at Memorization in Deep Networks

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, Simon Lacoste-Julien ; PMLR 70:233-242

[abs] [Download PDF]

An Alternative Softmax Operator for Reinforcement Learning

Kavosh Asadi, Michael L. Littman ; PMLR 70:243-252

[abs] [Download PDF]

Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, Amir Zandieh ; PMLR 70:253-262

[abs] [Download PDF][Supplementary PDF]

Minimax Regret Bounds for Reinforcement Learning

Mohammad Gheshlaghi Azar, Ian Osband, Rémi Munos ; PMLR 70:263-272

[abs] [Download PDF][Supplementary PDF]

Learning the Structure of Generative Models without Labeled Data

Stephen H. Bach, Bryan He, Alexander Ratner, Christopher Ré ; PMLR 70:273-282

[abs] [Download PDF][Supplementary PDF]

Uniform Deviation Bounds for k-Means Clustering

Olivier Bachem, Mario Lucic, S. Hamed Hassani, Andreas Krause ; PMLR 70:283-291

[abs] [Download PDF][Supplementary PDF]

Distributed and Provably Good Seedings for k-Means in Constant Rounds

Olivier Bachem, Mario Lucic, Andreas Krause ; PMLR 70:292-300

[abs] [Download PDF]

Learning Algorithms for Active Learning

Philip Bachman, Alessandro Sordoni, Adam Trischler ; PMLR 70:301-310

[abs] [Download PDF]

Improving Viterbi is Hard: Better Runtimes Imply Faster Clique Algorithms

Arturs Backurs, Christos Tzamos ; PMLR 70:311-321

[abs] [Download PDF][Supplementary PDF]

Differentially Private Clustering in High-Dimensional Euclidean Spaces

Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, Hongyang Zhang ; PMLR 70:322-331

[abs] [Download PDF][Supplementary PDF]

Strongly-Typed Agents are Guaranteed to Interact Safely

David Balduzzi ; PMLR 70:332-341

[abs] [Download PDF][Supplementary PDF]

The Shattered Gradients Problem: If resnets are the answer, then what is the question?

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, Brian McWilliams ; PMLR 70:342-350

[abs] [Download PDF][Supplementary PDF]

Neural Taylor Approximations: Convergence and Exploration in Rectifier Networks

David Balduzzi, Brian McWilliams, Tony Butler-Yeoman ; PMLR 70:351-360

[abs] [Download PDF][Supplementary PDF]

Spectral Learning from a Single Trajectory under Finite-State Policies

Borja Balle, Odalric-Ambrym Maillard ; PMLR 70:361-370

[abs] [Download PDF][Supplementary PDF]

Lost Relatives of the Gumbel Trick

Matej Balog, Nilesh Tripuraneni, Zoubin Ghahramani, Adrian Weller ; PMLR 70:371-379

[abs] [Download PDF][Supplementary PDF]

Dynamic Word Embeddings

Robert Bamler, Stephan Mandt ; PMLR 70:380-389

[abs] [Download PDF][Supplementary PDF]

End-to-End Differentiable Adversarial Imitation Learning

Nir Baram, Oron Anschel, Itai Caspi, Shie Mannor ; PMLR 70:390-399

[abs] [Download PDF]

Emulating the Expert: Inverse Optimization through Online Learning

Andreas Bärmann, Sebastian Pokutta, Oskar Schneider ; PMLR 70:400-410

[abs] [Download PDF]

Unimodal Probability Distributions for Deep Ordinal Classification

Christopher Beckham, Christopher Pal ; PMLR 70:411-419

[abs] [Download PDF]

Globally Induced Forest: A Prepruning Compression Scheme

Jean-Michel Begon, Arnaud Joly, Pierre Geurts ; PMLR 70:420-428

[abs] [Download PDF][Supplementary PDF]

End-to-End Learning for Structured Prediction Energy Networks

David Belanger, Bishan Yang, Andrew McCallum ; PMLR 70:429-439

[abs] [Download PDF][Supplementary PDF]

Learning to Discover Sparse Graphical Models

Eugene Belilovsky, Kyle Kastner, Gael Varoquaux, Matthew B. Blaschko ; PMLR 70:440-448

[abs] [Download PDF][Supplementary PDF]

A Distributional Perspective on Reinforcement Learning

Marc G. Bellemare, Will Dabney, Rémi Munos ; PMLR 70:449-458

[abs] [Download PDF][Supplementary PDF]

Neural Optimizer Search with Reinforcement Learning

Irwan Bello, Barret Zoph, Vijay Vasudevan, Quoc V. Le ; PMLR 70:459-468

[abs] [Download PDF]

Learning Texture Manifolds with the Periodic Spatial GAN

Urs Bergmann, Nikolay Jetchev, Roland Vollgraf ; PMLR 70:469-477

[abs] [Download PDF]

Differentially Private Learning of Undirected Graphical Models Using Collective Graphical Models

Garrett Bernstein, Ryan McKenna, Tao Sun, Daniel Sheldon, Michael Hay, Gerome Miklau ; PMLR 70:478-487

[abs] [Download PDF][Supplementary PDF]

Efficient Online Bandit Multiclass Learning with $\tilde{O}(\sqrt{T})$ Regret

Alina Beygelzimer, Francesco Orabona, Chicheng Zhang ; PMLR 70:488-497

[abs] [Download PDF][Supplementary PDF]

Guarantees for Greedy Maximization of Non-submodular Functions with Applications

Andrew An Bian, Joachim M. Buhmann, Andreas Krause, Sebastian Tschiatschek ; PMLR 70:498-507

[abs] [Download PDF][Supplementary PDF]

Robust Submodular Maximization: A Non-Uniform Partitioning Approach

Ilija Bogunovic, Slobodan Mitrović, Jonathan Scarlett, Volkan Cevher ; PMLR 70:508-516

[abs] [Download PDF][Supplementary PDF]

Unsupervised Learning by Predicting Noise

Piotr Bojanowski, Armand Joulin ; PMLR 70:517-526

[abs] [Download PDF]

Adaptive Neural Networks for Efficient Inference

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, Venkatesh Saligrama ; PMLR 70:527-536

[abs] [Download PDF]

Compressed Sensing using Generative Models

Ashish Bora, Ajil Jalal, Eric Price, Alexandros G. Dimakis ; PMLR 70:537-546

[abs] [Download PDF][Supplementary PDF]

Programming with a Differentiable Forth Interpreter

Matko Bošnjak, Tim Rocktäschel, Jason Naradowsky, Sebastian Riedel ; PMLR 70:547-556

[abs] [Download PDF]

Practical Gauss-Newton Optimisation for Deep Learning

Aleksandar Botev, Hippolyt Ritter, David Barber ; PMLR 70:557-565

[abs] [Download PDF][Supplementary PDF]

Lazifying Conditional Gradient Algorithms

Gábor Braun, Sebastian Pokutta, Daniel Zink ; PMLR 70:566-575

[abs] [Download PDF][Supplementary PDF]

Clustering High Dimensional Dynamic Data Streams

Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, Lin F. Yang ; PMLR 70:576-585

[abs] [Download PDF][Supplementary PDF]

On the Sampling Problem for Kernel Quadrature

François-Xavier Briol, Chris J. Oates, Jon Cockayne, Wilson Ye Chen, Mark Girolami ; PMLR 70:586-595

[abs] [Download PDF][Supplementary PDF]

Reduced Space and Faster Convergence in Imperfect-Information Games via Pruning

Noam Brown, Tuomas Sandholm ; PMLR 70:596-604

[abs] [Download PDF][Supplementary PDF]

Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs

Alon Brutzkus, Amir Globerson ; PMLR 70:605-614

[abs] [Download PDF][Supplementary PDF]

Deep Tensor Convolution on Multicores

David Budden, Alexander Matveev, Shibani Santurkar, Shraman Ray Chaudhuri, Nir Shavit ; PMLR 70:615-624

[abs] [Download PDF][Supplementary PDF]

Multi-objective Bandits: Optimizing the Generalized Gini Index

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Shie Mannor ; PMLR 70:625-634

[abs] [Download PDF][Supplementary PDF]

Priv'IT: Private and Sample Efficient Identity Testing

Bryan Cai, Constantinos Daskalakis, Gautam Kamath ; PMLR 70:635-644

[abs] [Download PDF][Supplementary PDF]

Second-Order Kernel Online Convex Optimization with Adaptive Sketching

Daniele Calandriello, Alessandro Lazaric, Michal Valko ; PMLR 70:645-653

[abs] [Download PDF][Supplementary PDF]

"Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

Yair Carmon, John C. Duchi, Oliver Hinder, Aaron Sidford ; PMLR 70:654-663

[abs] [Download PDF][Supplementary PDF]

Sliced Wasserstein Kernel for Persistence Diagrams

Mathieu Carrière, Marco Cuturi, Steve Oudot ; PMLR 70:664-673

[abs] [Download PDF]

Multiple Clustering Views from Multiple Uncertain Experts

Yale Chang, Junxiang Chen, Michael H. Cho, Peter J. Castaldi, Edwin K. Silverman, Jennifer G. Dy ; PMLR 70:674-683

[abs] [Download PDF][Supplementary PDF]

Uncertainty Assessment and False Discovery Rate Control in High-Dimensional Granger Causal Inference

Aditya Chaudhry, Pan Xu, Quanquan Gu ; PMLR 70:684-693

[abs] [Download PDF][Supplementary PDF]

Active Heteroscedastic Regression

Kamalika Chaudhuri, Prateek Jain, Nagarajan Natarajan ; PMLR 70:694-702

[abs] [Download PDF][Supplementary PDF]

Combining Model-Based and Model-Free Updates for Trajectory-Centric Reinforcement Learning

Yevgen Chebotar, Karol Hausman, Marvin Zhang, Gaurav Sukhatme, Stefan Schaal, Sergey Levine ; PMLR 70:703-711

[abs] [Download PDF][Supplementary PDF]

Robust Structured Estimation with Single-Index Models

Sheng Chen, Arindam Banerjee ; PMLR 70:712-721

[abs] [Download PDF][Supplementary PDF]

Adaptive Multiple-Arm Identification

Jiecao Chen, Xi Chen, Qin Zhang, Yuan Zhou ; PMLR 70:722-730

[abs] [Download PDF]

Dueling Bandits with Weak Regret

Bangrui Chen, Peter I. Frazier ; PMLR 70:731-739

[abs] [Download PDF][Supplementary PDF]

Strong NP-Hardness for Sparse Optimization with Concave Penalty Functions

Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, Hao Yin ; PMLR 70:740-747

[abs] [Download PDF][Supplementary PDF]

Learning to Learn without Gradient Descent by Gradient Descent

Yutian Chen, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matt Botvinick, Nando Freitas ; PMLR 70:748-756

[abs] [Download PDF]

Identification and Model Testing in Linear Structural Equation Models using Auxiliary Variables

Bryant Chen, Daniel Kumor, Elias Bareinboim ; PMLR 70:757-766

[abs] [Download PDF]

Toward Efficient and Accurate Covariance Matrix Estimation on Compressed Data

Xixian Chen, Michael R. Lyu, Irwin King ; PMLR 70:767-776

[abs] [Download PDF][Supplementary PDF]

Online Partial Least Square Optimization: Dropping Convexity for Better Efficiency and Scalability

Zhehui Chen, Lin F. Yang, Chris Junchi Li, Tuo Zhao ; PMLR 70:777-786

[abs] [Download PDF][Supplementary PDF]

Learning to Aggregate Ordinal Labels by Maximizing Separating Width

Guangyong Chen, Shengyu Zhang, Di Lin, Hui Huang, Pheng Ann Heng ; PMLR 70:787-796

[abs] [Download PDF]

Nearly Optimal Robust Matrix Completion

Yeshwanth Cherapanamjeri, Kartik Gupta, Prateek Jain ; PMLR 70:797-805

[abs] [Download PDF][Supplementary PDF]

Algorithms for ℓ_p Low-Rank Approximation

Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, David P. Woodruff ; PMLR 70:806-814

[abs] [Download PDF]

MEC: Memory-efficient Convolution for Deep Neural Network

Minsik Cho, Daniel Brand ; PMLR 70:815-824

[abs] [Download PDF]

On Relaxing Determinism in Arithmetic Circuits

Arthur Choi, Adnan Darwiche ; PMLR 70:825-833

[abs] [Download PDF]

Improving Stochastic Policy Gradients in Continuous Control with Deep Reinforcement Learning using the Beta Distribution

Po-Wei Chou, Daniel Maturana, Sebastian Scherer ; PMLR 70:834-843

[abs] [Download PDF][Supplementary PDF]

On Kernelized Multi-armed Bandits

Sayak Ray Chowdhury, Aditya Gopalan ; PMLR 70:844-853

[abs] [Download PDF][Supplementary PDF]

Parseval Networks: Improving Robustness to Adversarial Examples

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, Nicolas Usunier ; PMLR 70:854-863

[abs] [Download PDF]

Deep Latent Dirichlet Allocation with Topic-Layer-Adaptive Stochastic Gradient Riemannian MCMC

Yulai Cong, Bo Chen, Hongwei Liu, Mingyuan Zhou ; PMLR 70:864-873

[abs] [Download PDF][Supplementary PDF]

AdaNet: Adaptive Structural Learning of Artificial Neural Networks

Corinna Cortes, Xavier Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, Scott Yang ; PMLR 70:874-883

[abs] [Download PDF][Supplementary PDF]

Random Feature Expansions for Deep Gaussian Processes

Kurt Cutajar, Edwin V. Bonilla, Pietro Michiardi, Maurizio Filippone ; PMLR 70:884-893

[abs] [Download PDF][Supplementary PDF]

Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi, Mathieu Blondel ; PMLR 70:894-903

[abs] [Download PDF][Supplementary PDF]

Understanding Synthetic Gradients and Decoupled Neural Interfaces

Wojciech Marian Czarnecki, Grzegorz Świrszcz, Max Jaderberg, Simon Osindero, Oriol Vinyals, Koray Kavukcuoglu ; PMLR 70:904-912

[abs] [Download PDF][Supplementary PDF]

Stochastic Generative Hashing

Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, Le Song ; PMLR 70:913-922

[abs] [Download PDF][Supplementary PDF]

Logarithmic Time One-Against-Some

Hal Daumé III, Nikos Karampatziakis, John Langford, Paul Mineiro ; PMLR 70:923-932

[abs] [Download PDF]

Language Modeling with Gated Convolutional Networks

Yann N. Dauphin, Angela Fan, Michael Auli, David Grangier ; PMLR 70:933-941

[abs] [Download PDF]

An Infinite Hidden Markov Model With Similarity-Biased Transitions

Colin Reimer Dawson, Chaofan Huang, Clayton T. Morrison ; PMLR 70:942-950

[abs] [Download PDF][Supplementary PDF]

Distributed Batch Gaussian Process Optimization

Erik A. Daxberger, Bryan Kian Hsiang Low ; PMLR 70:951-960

[abs] [Download PDF][Supplementary PDF]

Consistency Analysis for Binary Classification Revisited

Krzysztof Dembczyński, Wojciech Kotłowski, Oluwasanmi Koyejo, Nagarajan Natarajan ; PMLR 70:961-969

[abs] [Download PDF]

iSurvive: An Interpretable, Event-time Prediction Model for mHealth

Walter H. Dempsey, Alexander Moreno, Christy K. Scott, Michael L. Dennis, David H. Gustafson, Susan A. Murphy, James M. Rehg ; PMLR 70:970-979

[abs] [Download PDF][Supplementary PDF]

Image-to-Markup Generation with Coarse-to-Fine Attention

Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, Alexander M. Rush ; PMLR 70:980-989

[abs] [Download PDF]

RobustFill: Neural Program Learning under Noisy I/O

Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, Pushmeet Kohli ; PMLR 70:990-998

[abs] [Download PDF][Supplementary PDF]

Being Robust (in High Dimensions) Can Be Practical

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, Alistair Stewart ; PMLR 70:999-1008

[abs] [Download PDF][Supplementary PDF]

Probabilistic Path Hamiltonian Monte Carlo

Vu Dinh, Arman Bilge, Cheng Zhang, Frederick A. Matsen IV ; PMLR 70:1009-1018

[abs] [Download PDF][Supplementary PDF]

Sharp Minima Can Generalize For Deep Nets

Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio ; PMLR 70:1019-1028

[abs] [Download PDF][Supplementary PDF]

A Divergence Bound for Hybrids of MCMC and Variational Inference and an Application to Langevin Dynamics and SGVI

Justin Domke ; PMLR 70:1029-1038

[abs] [Download PDF][Supplementary PDF]

Dance Dance Convolution

Chris Donahue, Zachary C. Lipton, Julian McAuley ; PMLR 70:1039-1048

[abs] [Download PDF]

Stochastic Variance Reduction Methods for Policy Evaluation

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, Dengyong Zhou ; PMLR 70:1049-1058

[abs] [Download PDF][Supplementary PDF]

Rule-Enhanced Penalized Regression by Column Generation using Rectangular Maximum Agreement

Jonathan Eckstein, Noam Goldberg, Ai Kagawa ; PMLR 70:1059-1067

[abs] [Download PDF]

Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, Karen Simonyan ; PMLR 70:1068-1077

[abs] [Download PDF]

Statistical Inference for Incomplete Ranking Data: The Case of Rank-Dependent Coarsening

Mohsen Ahmadi Fahandar, Eyke Hüllermeier, Inés Couso ; PMLR 70:1078-1087

[abs] [Download PDF][Supplementary PDF]

Maximum Selection and Ranking under Noisy Comparisons

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, Ananda Theertha Suresh ; PMLR 70:1088-1096

[abs] [Download PDF][Supplementary PDF]

Fake News Mitigation via Point Process Based Intervention

Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, Hongyuan Zha ; PMLR 70:1097-1106

[abs] [Download PDF][Supplementary PDF]

Regret Minimization in Behaviorally-Constrained Zero-Sum Games

Gabriele Farina, Christian Kroer, Tuomas Sandholm ; PMLR 70:1107-1116

[abs] [Download PDF]

Coresets for Vector Summarization with Applications to Network Graphs

Dan Feldman, Sedat Ozer, Daniela Rus ; PMLR 70:1117-1125

[abs] [Download PDF]

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Chelsea Finn, Pieter Abbeel, Sergey Levine ; PMLR 70:1126-1135

[abs] [Download PDF][Supplementary PDF]

Input Switched Affine Networks: An RNN Architecture Designed for Interpretability

Jakob N. Foerster, Justin Gilmer, Jascha Sohl-Dickstein, Jan Chorowski, David Sussillo ; PMLR 70:1136-1145

[abs] [Download PDF]

Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip H. S. Torr, Pushmeet Kohli, Shimon Whiteson ; PMLR 70:1146-1155

[abs] [Download PDF]

Counterfactual Data-Fusion for Online Reinforcement Learners

Andrew Forney, Judea Pearl, Elias Bareinboim ; PMLR 70:1156-1164

[abs] [Download PDF][Supplementary PDF]

Forward and Reverse Gradient-Based Hyperparameter Optimization

Luca Franceschi, Michele Donini, Paolo Frasconi, Massimiliano Pontil ; PMLR 70:1165-1173

[abs] [Download PDF][Supplementary PDF]

Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier

Joseph Futoma, Sanjay Hariharan, Katherine Heller ; PMLR 70:1174-1182

[abs] [Download PDF]

Deep Bayesian Active Learning with Image Data

Yarin Gal, Riashat Islam, Zoubin Ghahramani ; PMLR 70:1183-1192

[abs] [Download PDF]

Local-to-Global Bayesian Network Structure Learning

Tian Gao, Kshitij Fadnis, Murray Campbell ; PMLR 70:1193-1202

[abs] [Download PDF]

Communication-efficient Algorithms for Distributed Stochastic Principal Component Analysis

Dan Garber, Ohad Shamir, Nathan Srebro ; PMLR 70:1203-1212

[abs] [Download PDF][Supplementary PDF]

Differentiable Programs with Neural Libraries

Alexander L. Gaunt, Marc Brockschmidt, Nate Kushman, Daniel Tarlow ; PMLR 70:1213-1222

[abs] [Download PDF][Supplementary PDF]

Zonotope Hit-and-run for Efficient Sampling from Projection DPPs

Guillaume Gautier, Rémi Bardenet, Michal Valko ; PMLR 70:1223-1232

[abs] [Download PDF][Supplementary PDF]

No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis

Rong Ge, Chi Jin, Yi Zheng ; PMLR 70:1233-1242

[abs] [Download PDF][Supplementary PDF]

Convolutional Sequence to Sequence Learning

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin ; PMLR 70:1243-1252

[abs] [Download PDF][Supplementary PDF]

On Context-Dependent Clustering of Bandits

Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, Evans Etrue ; PMLR 70:1253-1262

[abs] [Download PDF]

Neural Message Passing for Quantum Chemistry

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, George E. Dahl ; PMLR 70:1263-1272

[abs] [Download PDF][Supplementary PDF]

Convex Phase Retrieval without Lifting via PhaseMax

Tom Goldstein, Christoph Studer ; PMLR 70:1273-1281

[abs] [Download PDF]

Preferential Bayesian Optimization

Javier González, Zhenwen Dai, Andreas Damianou, Neil D. Lawrence ; PMLR 70:1282-1291

[abs] [Download PDF]

Measuring Sample Quality with Kernels

Jackson Gorham, Lester Mackey ; PMLR 70:1292-1301

[abs] [Download PDF][Supplementary PDF]

Efficient softmax approximation for GPUs

Grave, Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou ; PMLR 70:1302-1310

[abs] [Download PDF]

Automated Curriculum Learning for Neural Networks

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, Koray Kavukcuoglu ; PMLR 70:1311-1320

[abs] [Download PDF]

On Calibration of Modern Neural Networks

Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger ; PMLR 70:1321-1330

[abs] [Download PDF][Supplementary PDF]

ProtoNN: Compressed and Accurate kNN for Resource-scarce Devices

Chirag Gupta, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udupa, Manik Varma, Prateek Jain ; PMLR 70:1331-1340

[abs] [Download PDF][Supplementary PDF]

Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs

Michael Gygli, Mohammad Norouzi, Anelia Angelova ; PMLR 70:1341-1351

[abs] [Download PDF]

Reinforcement Learning with Deep Energy-Based Policies

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, Sergey Levine ; PMLR 70:1352-1361

[abs] [Download PDF][Supplementary PDF]

DeepBach: a Steerable Model for Bach Chorales Generation

Gaëtan Hadjeres, François Pachet, Frank Nielsen ; PMLR 70:1362-1371

[abs] [Download PDF]

Consistent On-Line Off-Policy Evaluation

Assaf Hallak, Shie Mannor ; PMLR 70:1372-1383

[abs] [Download PDF][Supplementary PDF]

Faster Greedy MAP Inference for Determinantal Point Processes

Insu Han, Prabhanjan Kambadur, Kyoungsoo Park, Jinwoo Shin ; PMLR 70:1384-1393

[abs] [Download PDF][Supplementary PDF]

Data-Efficient Policy Evaluation Through Behavior Policy Search

Josiah P. Hanna, Philip S. Thomas, Peter Stone, Scott Niekum ; PMLR 70:1394-1403

[abs] [Download PDF][Supplementary PDF]

Joint Dimensionality Reduction and Metric Learning: A Geometric Take

Mehrtash Harandi, Mathieu Salzmann, Richard Hartley ; PMLR 70:1404-1413

[abs] [Download PDF][Supplementary PDF]

Deep IV: A Flexible Approach for Counterfactual Prediction

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, Matt Taddy ; PMLR 70:1414-1423

[abs] [Download PDF][Supplementary PDF]

Robust Guarantees of Stochastic Greedy Algorithms

Avinatan Hassidim, Yaron Singer ; PMLR 70:1424-1432

[abs] [Download PDF]

Efficient Regret Minimization in Non-Convex Games

Elad Hazan, Karan Singh, Cyril Zhang ; PMLR 70:1433-1441

[abs] [Download PDF][Supplementary PDF]

Kernelized Support Tensor Machines

Lifang He, Chun-Ta Lu, Guixiang Ma, Shen Wang, Linlin Shen, Philip S. Yu, Ann B. Ragin ; PMLR 70:1442-1451

[abs] [Download PDF]

The Sample Complexity of Online One-Class Collaborative Filtering

Reinhard Heckel, Kannan Ramchandran ; PMLR 70:1452-1460

[abs] [Download PDF][Supplementary PDF]

Warped Convolutions: Efficient Invariance to Spatial Transformations

João F. Henriques, Andrea Vedaldi ; PMLR 70:1461-1469

[abs] [Download PDF]

Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space

José Miguel Hernández-Lobato, James Requeima, Edward O. Pyzer-Knapp, Alán Aspuru-Guzik ; PMLR 70:1470-1479

[abs] [Download PDF]

DARLA: Improving Zero-Shot Transfer in Reinforcement Learning

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, Alexander Lerchner ; PMLR 70:1480-1490

[abs] [Download PDF][Supplementary PDF]

SPLICE: Fully Tractable Hierarchical Extension of ICA with Pooling

Jun-ichiro Hirayama, Aapo Hyvärinen, Motoaki Kawanabe ; PMLR 70:1491-1500

[abs] [Download PDF][Supplementary PDF]

Multilevel Clustering via Wasserstein Means

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, Dinh Phung ; PMLR 70:1501-1509

[abs] [Download PDF][Supplementary PDF]

Learning Deep Latent Gaussian Models with Markov Chain Monte Carlo

Matthew D. Hoffman ; PMLR 70:1510-1519

[abs] [Download PDF][Supplementary PDF]

Minimizing Trust Leaks for Robust Sybil Detection

János Höner, Shinichi Nakajima, Alexander Bauer, Klaus-Robert Müller, Nico Görnitz ; PMLR 70:1520-1528

[abs] [Download PDF]

Prox-PDA: The Proximal Primal-Dual Algorithm for Fast Distributed Nonconvex Optimization and Learning Over Networks

Mingyi Hong, Davood Hajinezhad, Ming-Min Zhao ; PMLR 70:1529-1538

[abs] [Download PDF][Supplementary PDF]

Analysis and Optimization of Graph Decompositions by Lifted Multicuts

Andrea Horňáková, Jan-Hendrik Lange, Bjoern Andres ; PMLR 70:1539-1548

[abs] [Download PDF]

Dissipativity Theory for Nesterov's Accelerated Method

Bin Hu, Laurent Lessard ; PMLR 70:1549-1557

[abs] [Download PDF][Supplementary PDF]

Learning Discrete Representations via Information Maximizing Self-Augmented Training

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, Masashi Sugiyama ; PMLR 70:1558-1567

[abs] [Download PDF][Supplementary PDF]

State-Frequency Memory Recurrent Neural Networks

Hao Hu, Guo-Jun Qi ; PMLR 70:1568-1577

[abs] [Download PDF]

Deep Generative Models for Relational Data with Side Information

Changwei Hu, Piyush Rai, Lawrence Carin ; PMLR 70:1578-1586

[abs] [Download PDF][Supplementary PDF]

Toward Controlled Generation of Text

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing ; PMLR 70:1587-1596

[abs] [Download PDF][Supplementary PDF]

Tensor Decomposition with Smoothness

Masaaki Imaizumi, Kohei Hayashi ; PMLR 70:1597-1606

[abs] [Download PDF][Supplementary PDF]

Variational Inference for Sparse and Undirected Models

John Ingraham, Debora Marks ; PMLR 70:1607-1616

[abs] [Download PDF][Supplementary PDF]

Fairness in Reinforcement Learning

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Aaron Roth ; PMLR 70:1617-1626

[abs] [Download PDF][Supplementary PDF]

Decoupled Neural Interfaces using Synthetic Gradients

Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, Koray Kavukcuoglu ; PMLR 70:1627-1635

[abs] [Download PDF][Supplementary PDF]

Scalable Generative Models for Multi-label Learning with Missing Labels

Vikas Jain, Nirbhay Modhe, Piyush Rai ; PMLR 70:1636-1644

[abs] [Download PDF]

Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E. Turner, Douglas Eck ; PMLR 70:1645-1654

[abs] [Download PDF]

Bayesian Optimization with Tree-structured Dependencies

Rodolphe Jenatton, Cedric Archambeau, Javier González, Matthias Seeger ; PMLR 70:1655-1664

[abs] [Download PDF][Supplementary PDF]

Simultaneous Learning of Trees and Representations for Extreme Classification and Density Estimation

Yacine Jernite, Anna Choromanska, David Sontag ; PMLR 70:1665-1674

[abs] [Download PDF][Supplementary PDF]

From Patches to Images: A Nonparametric Generative Model

Geng Ji, Michael C. Hughes, Erik B. Sudderth ; PMLR 70:1675-1683

[abs] [Download PDF][Supplementary ZIP]

Density Level Set Estimation on Manifolds with DBSCAN

Heinrich Jiang ; PMLR 70:1684-1693

[abs] [Download PDF][Supplementary PDF]

Uniform Convergence Rates for Kernel Density Estimation

Heinrich Jiang ; PMLR 70:1694-1703

[abs] [Download PDF][Supplementary PDF]

Contextual Decision Processes with low Bellman rank are PAC-Learnable

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, Robert E. Schapire ; PMLR 70:1704-1713

[abs] [Download PDF][Supplementary PDF]

Efficient Nonmyopic Active Search

Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, Roman Garnett ; PMLR 70:1714-1723

[abs] [Download PDF][Supplementary PDF]

How to Escape Saddle Points Efficiently

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, Michael I. Jordan ; PMLR 70:1724-1732

[abs] [Download PDF][Supplementary PDF]

Tunable Efficient Unitary Neural Networks (EUNN) and their application to RNNs

Li Jing, Yichen Shen, Tena Dubcek, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, Marin Soljačić ; PMLR 70:1733-1741

[abs] [Download PDF]

An Adaptive Test of Independence with Analytic Kernel Embeddings

Wittawat Jitkrittum, Zoltán Szabó, Arthur Gretton ; PMLR 70:1742-1751

[abs] [Download PDF][Supplementary PDF]

StingyCD: Safely Avoiding Wasteful Updates in Coordinate Descent

Tyler B. Johnson, Carlos Guestrin ; PMLR 70:1752-1760

[abs] [Download PDF][Supplementary PDF]

Differentially Private Chi-squared Test by Unit Circle Mechanism

Kazuya Kakizaki, Kazuto Fukuchi, Jun Sakuma ; PMLR 70:1761-1770

[abs] [Download PDF][Supplementary PDF]

Video Pixel Networks

Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, Koray Kavukcuoglu ; PMLR 70:1771-1779

[abs] [Download PDF][Supplementary ZIP]

Adaptive Feature Selection: Computationally Efficient Online Sparse Linear Regression under RIP

Satyen Kale, Zohar Karnin, Tengyuan Liang, Dávid Pál ; PMLR 70:1780-1788

[abs] [Download PDF][Supplementary PDF]

Recursive Partitioning for Personalization using Observational Data

Nathan Kallus ; PMLR 70:1789-1798

[abs] [Download PDF][Supplementary PDF]

Multi-fidelity Bayesian Optimisation with Continuous Approximations

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, Barnabás Póczos ; PMLR 70:1799-1808

[abs] [Download PDF][Supplementary PDF]

Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, Dileep George ; PMLR 70:1809-1818

[abs] [Download PDF][Supplementary PDF]

Learning in POMDPs with Monte Carlo Tree Search

Sammie Katt, Frans A. Oliehoek, Christopher Amato ; PMLR 70:1819-1827

[abs] [Download PDF][Supplementary PDF]

Meritocratic Fairness for Cross-Population Selection

Michael Kearns, Aaron Roth, Zhiwei Steven Wu ; PMLR 70:1828-1836

[abs] [Download PDF]

On Approximation Guarantees for Greedy Low Rank Optimization

Rajiv Khanna, Ethan R. Elenberg, Alexandros G. Dimakis, Joydeep Ghosh, Sahand Negahban ; PMLR 70:1837-1846

[abs] [Download PDF][Supplementary PDF]

Graph-based Isometry Invariant Representation Learning

Renata Khasanova, Pascal Frossard ; PMLR 70:1847-1856

[abs] [Download PDF][Supplementary PDF]

Learning to Discover Cross-Domain Relations with Generative Adversarial Networks

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, Jiwon Kim ; PMLR 70:1857-1865

[abs] [Download PDF][Supplementary ZIP]

SplitNet: Learning to Semantically Split Deep Networks for Parameter Reduction and Model Parallelization

Juyong Kim, Yookoon Park, Gunhee Kim, Sung Ju Hwang ; PMLR 70:1866-1874

[abs] [Download PDF][Supplementary PDF]

Cost-Optimal Learning of Causal Graphs

Murat Kocaoglu, Alex Dimakis, Sriram Vishwanath ; PMLR 70:1875-1884

[abs] [Download PDF][Supplementary PDF]

Understanding Black-box Predictions via Influence Functions

Pang Wei Koh, Percy Liang ; PMLR 70:1885-1894

[abs] [Download PDF][Supplementary PDF]

Sub-sampled Cubic Regularization for Non-convex Optimization

Jonas Moritz Kohler, Aurelien Lucchi ; PMLR 70:1895-1904

[abs] [Download PDF][Supplementary PDF]

PixelCNN Models with Auxiliary Variables for Natural Image Modeling

Alexander Kolesnikov, Christoph H. Lampert ; PMLR 70:1905-1914

[abs] [Download PDF]

Active Learning for Cost-Sensitive Classification

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, John Langford ; PMLR 70:1915-1924

[abs] [Download PDF][Supplementary PDF]

Evaluating Bayesian Models with Posterior Dispersion Indices

Alp Kucukelbir, Yixin Wang, David M. Blei ; PMLR 70:1925-1934

[abs] [Download PDF][Supplementary ZIP]

Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things

Ashish Kumar, Saurabh Goyal, Manik Varma ; PMLR 70:1935-1944

[abs] [Download PDF]

Grammar Variational Autoencoder

Matt J. Kusner, Brooks Paige, José Miguel Hernández-Lobato ; PMLR 70:1945-1954

[abs] [Download PDF][Supplementary PDF]

Co-clustering through Optimal Transport

Charlotte Laclau, Ievgen Redko, Basarab Matei, Younès Bennani, Vincent Brault ; PMLR 70:1955-1964

[abs] [Download PDF][Supplementary PDF]

Conditional Accelerated Lazy Stochastic Gradient Descent

Guanghui Lan, Sebastian Pokutta, Yi Zhou, Daniel Zink ; PMLR 70:1965-1974

[abs] [Download PDF][Supplementary PDF]

Consistent k-Clustering

Silvio Lattanzi, Sergei Vassilvitskii ; PMLR 70:1975-1984

[abs] [Download PDF]

Deep Spectral Clustering Learning

Marc T. Law, Raquel Urtasun, Richard S. Zemel ; PMLR 70:1985-1994

[abs] [Download PDF][Supplementary ZIP]

Coordinated Multi-Agent Imitation Learning

Hoang M. Le, Yisong Yue, Peter Carr, Patrick Lucey ; PMLR 70:1995-2003

[abs] [Download PDF][Supplementary PDF]

Bayesian inference on random simple graphs with power law degree distributions

Juho Lee, Creighton Heaukulani, Zoubin Ghahramani, Lancelot F. James, Seungjin Choi ; PMLR 70:2004-2013

[abs] [Download PDF][Supplementary PDF]

Confident Multiple Choice Learning

Kimin Lee, Changho Hwang, KyoungSoo Park, Jinwoo Shin ; PMLR 70:2014-2023

[abs] [Download PDF][Supplementary PDF]

Deriving Neural Architectures from Sequence and Graph Kernels

Tao Lei, Wengong Jin, Regina Barzilay, Tommi Jaakkola ; PMLR 70:2024-2033

[abs] [Download PDF]

Doubly Greedy Primal-Dual Coordinate Descent for Sparse Empirical Risk Minimization

Qi Lei, Ian En-Hsu Yen, Chao-yuan Wu, Inderjit S. Dhillon, Pradeep Ravikumar ; PMLR 70:2034-2042

[abs] [Download PDF][Supplementary PDF]

Learning to Align the Source Code to the Compiled Object Code

Dor Levy, Lior Wolf ; PMLR 70:2043-2051

[abs] [Download PDF][Supplementary PDF]

Dropout Inference in Bayesian Neural Networks with Alpha-divergences

Yingzhen Li, Yarin Gal ; PMLR 70:2052-2061

[abs] [Download PDF][Supplementary PDF]

Provable Alternating Gradient Descent for Non-negative Matrix Factorization with Strong Correlations

Yuanzhi Li, Yingyu Liang ; PMLR 70:2062-2070

[abs] [Download PDF][Supplementary PDF]

Provably Optimal Algorithms for Generalized Linear Contextual Bandits

Lihong Li, Yu Lu, Dengyong Zhou ; PMLR 70:2071-2080

[abs] [Download PDF][Supplementary PDF]

Fast k-Nearest Neighbour Search via Prioritized DCI

Ke Li, Jitendra Malik ; PMLR 70:2081-2090

[abs] [Download PDF][Supplementary PDF]

Forest-type Regression with General Losses and Robust Forest

Alexander Hanbo Li, Andrew Martin ; PMLR 70:2091-2100

[abs] [Download PDF]

Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms

Qianxiao Li, Cheng Tai, Weinan E ; PMLR 70:2101-2110

[abs] [Download PDF][Supplementary PDF]

Convergence Analysis of Proximal Gradient with Momentum for Nonconvex Optimization

Qunwei Li, Yi Zhou, Yingbin Liang, Pramod K. Varshney ; PMLR 70:2111-2119

[abs] [Download PDF][Supplementary PDF]

Exact MAP Inference by Avoiding Fractional Vertices

Erik M. Lindgren, Alexandros G. Dimakis, Adam Klivans ; PMLR 70:2120-2129

[abs] [Download PDF]

Leveraging Union of Subspace Structure to Improve Constrained Clustering

John Lipor, Laura Balzano ; PMLR 70:2130-2139

[abs] [Download PDF][Supplementary PDF]

Zero-Inflated Exponential Family Embeddings

Li-Ping Liu, David M. Blei ; PMLR 70:2140-2148

[abs] [Download PDF]

Iterative Machine Teaching

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, Le Song ; PMLR 70:2149-2158

[abs] [Download PDF][Supplementary PDF]

Algorithmic Stability and Hypothesis Complexity

Tongliang Liu, Gábor Lugosi, Gergely Neu, Dacheng Tao ; PMLR 70:2159-2167

[abs] [Download PDF]

Analogical Inference for Multi-relational Embeddings

Hanxiao Liu, Yuexin Wu, Yiming Yang ; PMLR 70:2168-2178

[abs] [Download PDF]

Dual Iterative Hard Thresholding: From Non-convex Sparse Minimization to Non-smooth Concave Maximization

Bo Liu, Xiao-Tong Yuan, Lezi Wang, Qingshan Liu, Dimitris N. Metaxas ; PMLR 70:2179-2187

[abs] [Download PDF][Supplementary PDF]

Gram-CTC: Automatic Unit Selection and Target Decomposition for Sequence Labelling

Hairong Liu, Zhenyao Zhu, Xiangang Li, Sanjeev Satheesh ; PMLR 70:2188-2197

[abs] [Download PDF]

Learning Infinite Layer Networks Without the Kernel Trick

Roi Livni, Daniel Carmon, Amir Globerson ; PMLR 70:2198-2207

[abs] [Download PDF]

Deep Transfer Learning with Joint Adaptation Networks

Mingsheng Long, Han Zhu, Jianmin Wang, Michael I. Jordan ; PMLR 70:2208-2217

[abs] [Download PDF]

Multiplicative Normalizing Flows for Variational Bayesian Neural Networks

Christos Louizos, Max Welling ; PMLR 70:2218-2227

[abs] [Download PDF]

How Close Are the Eigenvectors of the Sample and Actual Covariance Matrices?

Andreas Loukas ; PMLR 70:2228-2237

[abs] [Download PDF]

Learning Deep Architectures via Generalized Whitened Neural Networks

Ping Luo ; PMLR 70:2238-2246

[abs] [Download PDF]

Learning Gradient Descent: Better Generalization and Longer Horizons

Kaifeng Lv, Shunhua Jiang, Jian Li ; PMLR 70:2247-2255

[abs] [Download PDF]

Spherical Structured Feature Maps for Kernel Approximation

Yueming Lyu ; PMLR 70:2256-2264

[abs] [Download PDF][Supplementary PDF]

Stochastic Gradient MCMC Methods for Hidden Markov Models

Yi-An Ma, Nicholas J. Foti, Emily B. Fox ; PMLR 70:2265-2274

[abs] [Download PDF][Supplementary PDF]

Self-Paced Co-training

Fan Ma, Deyu Meng, Qi Xie, Zina Li, Xuanyi Dong ; PMLR 70:2275-2284

[abs] [Download PDF][Supplementary PDF]

Interactive Learning from Policy-Dependent Human Feedback

James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, Michael L. Littman ; PMLR 70:2285-2294

[abs] [Download PDF]

A Laplacian Framework for Option Discovery in Reinforcement Learning

Marlos C. Machado, Marc G. Bellemare, Michael Bowling ; PMLR 70:2295-2304

[abs] [Download PDF][Supplementary PDF]

Frame-based Data Factorizations

Sebastian Mair, Ahcène Boubekki, Ulf Brefeld ; PMLR 70:2305-2313

[abs] [Download PDF]

Global optimization of Lipschitz functions

Cédric Malherbe, Nicolas Vayatis ; PMLR 70:2314-2323

[abs] [Download PDF][Supplementary ZIP]

On Mixed Memberships and Symmetric Nonnegative Matrix Factorizations

Xueyu Mao, Purnamrita Sarkar, Deepayan Chakrabarti ; PMLR 70:2324-2333

[abs] [Download PDF][Supplementary PDF]

Bayesian Models of Data Streams with Hierarchical Power Priors

Andrés Masegosa, Thomas D. Nielsen, Helge Langseth, Darío Ramos-López, Antonio Salmerón, Anders L. Madsen ; PMLR 70:2334-2343

[abs] [Download PDF][Supplementary PDF]

Just Sort It! A Simple and Effective Approach to Active Preference Learning

Lucas Maystre, Matthias Grossglauser ; PMLR 70:2344-2353

[abs] [Download PDF][Supplementary PDF]

ChoiceRank: Identifying Preferences from Node Traffic in Networks

Lucas Maystre, Matthias Grossglauser ; PMLR 70:2354-2362

[abs] [Download PDF][Supplementary ZIP]

Deciding How to Decide: Dynamic Routing in Artificial Neural Networks

Mason McGill, Pietro Perona ; PMLR 70:2363-2372

[abs] [Download PDF][Supplementary ZIP]

Risk Bounds for Transferring Representations With and Without Fine-Tuning

Daniel McNamara, Maria-Florina Balcan ; PMLR 70:2373-2381

[abs] [Download PDF]

Nonnegative Matrix Factorization for Time Series Recovery From a Few Temporal Aggregates

Jiali Mei, Yohann De Castro, Yannig Goude, Georges Hébrail ; PMLR 70:2382-2390

[abs] [Download PDF]

Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks

Lars Mescheder, Sebastian Nowozin, Andreas Geiger ; PMLR 70:2391-2400

[abs] [Download PDF][Supplementary PDF]

Efficient Orthogonal Parametrisation of Recurrent Neural Networks Using Householder Reflections

Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, James Bailey ; PMLR 70:2401-2409

[abs] [Download PDF][Supplementary PDF]

Discovering Discrete Latent Topics with Neural Variational Inference

Yishu Miao, Edward Grefenstette, Phil Blunsom ; PMLR 70:2410-2419

[abs] [Download PDF][Supplementary PDF]

Variational Boosting: Iteratively Refining Posterior Approximations

Andrew C. Miller, Nicholas J. Foti, Ryan P. Adams ; PMLR 70:2420-2429

[abs] [Download PDF][Supplementary PDF]

Device Placement Optimization with Reinforcement Learning

Azalia Mirhoseini, Hieu Pham, Quoc V. Le, Benoit Steiner, Rasmus Larsen, Yuefeng Zhou, Naveen Kumar, Mohammad Norouzi, Samy Bengio, Jeff Dean ; PMLR 70:2430-2439

[abs] [Download PDF]

Tight Bounds for Approximate Carathéodory and Beyond

Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, Sam Chiu-wai Wong ; PMLR 70:2440-2448

[abs] [Download PDF][Supplementary PDF]

Deletion-Robust Submodular Maximization: Data Summarization with "the Right to be Forgotten"

Baharan Mirzasoleiman, Amin Karbasi, Andreas Krause ; PMLR 70:2449-2458

[abs] [Download PDF][Supplementary PDF]

Prediction and Control with Temporal Segment Models

Nikhil Mishra, Pieter Abbeel, Igor Mordatch ; PMLR 70:2459-2468

[abs] [Download PDF][Supplementary PDF]

Improving Gibbs Sampler Scan Quality with DoGS

Ioannis Mitliagkas, Lester Mackey ; PMLR 70:2469-2477

[abs] [Download PDF][Supplementary PDF]

Differentially Private Submodular Maximization: Data Summarization in Disguise

Marko Mitrovic, Mark Bun, Andreas Krause, Amin Karbasi ; PMLR 70:2478-2487

[abs] [Download PDF][Supplementary PDF]

Active Learning for Top-K Rank Aggregation from Noisy Comparisons

Soheil Mohajer, Changho Suh, Adel Elmahdy ; PMLR 70:2488-2497

[abs] [Download PDF][Supplementary PDF]

Variational Dropout Sparsifies Deep Neural Networks

Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov ; PMLR 70:2498-2507

[abs] [Download PDF]

Regularising Non-linear Models Using Feature Side-information

Amina Mollaysa, Pablo Strasser, Alexandros Kalousis ; PMLR 70:2508-2517

[abs] [Download PDF][Supplementary PDF]

Coupling Distributed and Symbolic Execution for Natural Language Queries

Lili Mou, Zhengdong Lu, Hang Li, Zhi Jin ; PMLR 70:2518-2526

[abs] [Download PDF]

McGan: Mean and Covariance Feature Matching GAN

Youssef Mroueh, Tom Sercu, Vaibhava Goel ; PMLR 70:2527-2535

[abs] [Download PDF][Supplementary PDF]

Sequence to Better Sequence: Continuous Revision of Combinatorial Structures

Jonas Mueller, David Gifford, Tommi Jaakkola ; PMLR 70:2536-2544

[abs] [Download PDF][Supplementary PDF]

Variants of RMSProp and Adagrad with Logarithmic Regret Bounds

Mahesh Chandra Mukkamala, Matthias Hein ; PMLR 70:2545-2553

[abs] [Download PDF][Supplementary PDF]

Meta Networks

Tsendsuren Munkhdalai, Hong Yu ; PMLR 70:2554-2563

[abs] [Download PDF][Supplementary PDF]

Understanding the Representation and Computation of Multilayer Perceptrons: A Case Study in Speech Recognition

Tasha Nagamine, Nima Mesgarani ; PMLR 70:2564-2573

[abs] [Download PDF]

Adaptive Sampling Probabilities for Non-Smooth Optimization

Hongseok Namkoong, Aman Sinha, Steve Yadlowsky, John C. Duchi ; PMLR 70:2574-2583

[abs] [Download PDF][Supplementary PDF]

Delta Networks for Optimized Recurrent Network Computation

Daniel Neil, Jun Haeng Lee, Tobi Delbruck, Shih-Chii Liu ; PMLR 70:2584-2593

[abs] [Download PDF]

Post-Inference Prior Swapping

Willie Neiswanger, Eric Xing ; PMLR 70:2594-2602

[abs] [Download PDF][Supplementary PDF]

The Loss Surface of Deep and Wide Neural Networks

Quynh Nguyen, Matthias Hein ; PMLR 70:2603-2612

[abs] [Download PDF][Supplementary PDF]

SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient

Lam M. Nguyen, Jie Liu, Katya Scheinberg, Martin Takáč ; PMLR 70:2613-2621

[abs] [Download PDF][Supplementary PDF]

Composing Tree Graphical Models with Persistent Homology Features for Clustering Mixed-Type Data

Xiuyan Ni, Novi Quadrianto, Yusu Wang, Chao Chen ; PMLR 70:2622-2631

[abs] [Download PDF]

Multichannel End-to-end Speech Recognition

Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R. Hershey ; PMLR 70:2632-2641

[abs] [Download PDF]

Conditional Image Synthesis with Auxiliary Classifier GANs

Augustus Odena, Christopher Olah, Jonathon Shlens ; PMLR 70:2642-2651

[abs] [Download PDF][Supplementary PDF]

Nyström Method with Kernel K-means++ Samples as Landmarks

Dino Oglic, Thomas Gärtner ; PMLR 70:2652-2660

[abs] [Download PDF][Supplementary PDF]

Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning

Junhyuk Oh, Satinder Singh, Honglak Lee, Pushmeet Kohli ; PMLR 70:2661-2670

[abs] [Download PDF][Supplementary PDF]

The Statistical Recurrent Unit

Junier B. Oliva, Barnabás Póczos, Jeff Schneider ; PMLR 70:2671-2680

[abs] [Download PDF]

Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, John Vian ; PMLR 70:2681-2690

[abs] [Download PDF][Supplementary PDF]

Algebraic Variety Models for High-Rank Matrix Completion

Greg Ongie, Rebecca Willett, Robert D. Nowak, Laura Balzano ; PMLR 70:2691-2700

[abs] [Download PDF]

Why is Posterior Sampling Better than Optimism for Reinforcement Learning?

Ian Osband, Benjamin Van Roy ; PMLR 70:2701-2710

[abs] [Download PDF][Supplementary PDF]

Bidirectional Learning for Time-series Models with Hidden Units

Takayuki Osogami, Hiroshi Kajino, Taro Sekiyama ; PMLR 70:2711-2720

[abs] [Download PDF][Supplementary PDF]

Count-Based Exploration with Neural Density Models

Georg Ostrovski, Marc G. Bellemare, Aäron Oord, Rémi Munos ; PMLR 70:2721-2730

[abs] [Download PDF][Supplementary PDF]

Dictionary Learning Based on Sparse Distribution Tomography

Pedram Pad, Farnood Salehi, Elisa Celis, Patrick Thiran, Michael Unser ; PMLR 70:2731-2740

[abs] [Download PDF]

Stochastic Bouncy Particle Sampler

Ari Pakman, Dar Gilboa, David Carlson, Liam Paninski ; PMLR 70:2741-2750

[abs] [Download PDF][Supplementary PDF]

A Birth-Death Process for Feature Allocation

Konstantina Palla, David Knowles, Zoubin Ghahramani ; PMLR 70:2751-2759

[abs] [Download PDF][Supplementary PDF]

Prediction under Uncertainty in Sparse Spectrum Gaussian Processes with Applications to Filtering and Control

Yunpeng Pan, Xinyan Yan, Evangelos A. Theodorou, Byron Boots ; PMLR 70:2760-2768

[abs] [Download PDF][Supplementary ZIP]

Clustering by Sum of Norms: Stochastic Incremental Algorithm, Convergence and Cluster Recovery

Ashkan Panahi, Devdatt Dubhashi, Fredrik D. Johansson, Chiranjib Bhattacharyya ; PMLR 70:2769-2777

[abs] [Download PDF][Supplementary PDF]

Curiosity-driven Exploration by Self-supervised Prediction

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell ; PMLR 70:2778-2787

[abs] [Download PDF][Supplementary ZIP]

Asynchronous Distributed Variational Gaussian Process for Regression

Hao Peng, Shandian Zhe, Xiao Zhang, Yuan Qi ; PMLR 70:2788-2797

[abs] [Download PDF][Supplementary ZIP]

Geometry of Neural Network Loss Surfaces via Random Matrix Theory

Jeffrey Pennington, Yasaman Bahri ; PMLR 70:2798-2806

[abs] [Download PDF][Supplementary PDF]

Multi-task Learning with Labeled and Unlabeled Tasks

Anastasia Pentina, Christoph H. Lampert ; PMLR 70:2807-2816

[abs] [Download PDF][Supplementary PDF]

Robust Adversarial Reinforcement Learning

Lerrel Pinto, James Davidson, Rahul Sukthankar, Abhinav Gupta ; PMLR 70:2817-2826

[abs] [Download PDF]

Neural Episodic Control

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, Charles Blundell ; PMLR 70:2827-2836

[abs] [Download PDF][Supplementary PDF]

Online and Linear-Time Attention by Enforcing Monotonic Alignments

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, Douglas Eck ; PMLR 70:2837-2846

[abs] [Download PDF][Supplementary PDF]

On the Expressive Power of Deep Neural Networks

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, Jascha Sohl-Dickstein ; PMLR 70:2847-2854

[abs] [Download PDF][Supplementary PDF]

Estimating the unseen from multiple populations

Aditi Raghunathan, Gregory Valiant, James Zou ; PMLR 70:2855-2863

[abs] [Download PDF][Supplementary PDF]

Coherence Pursuit: Fast, Simple, and Robust Subspace Recovery

Mostafa Rahmani, George Atia ; PMLR 70:2864-2873

[abs] [Download PDF]

Innovation Pursuit: A New Approach to the Subspace Clustering Problem

Mostafa Rahmani, George Atia ; PMLR 70:2874-2882

[abs] [Download PDF]

High Dimensional Bayesian Optimization with Elastic Gaussian Process

Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, Svetha Venkatesh ; PMLR 70:2883-2891

[abs] [Download PDF]

Equivariance Through Parameter-Sharing

Siamak Ravanbakhsh, Jeff Schneider, Barnabás Póczos ; PMLR 70:2892-2901

[abs] [Download PDF][Supplementary PDF]

Large-Scale Evolution of Image Classifiers

Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V. Le, Alexey Kurakin ; PMLR 70:2902-2911

[abs] [Download PDF][Supplementary PDF]

Parallel Multiscale Autoregressive Density Estimation

Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, Nando Freitas ; PMLR 70:2912-2921

[abs] [Download PDF][Supplementary ZIP]

Real-Time Adaptive Image Compression

Oren Rippel, Lubomir Bourdev ; PMLR 70:2922-2930

[abs] [Download PDF][Supplementary PDF]

Active Learning for Accurate Estimation of Linear Models

Carlos Riquelme, Mohammad Ghavamzadeh, Alessandro Lazaric ; PMLR 70:2931-2939

[abs] [Download PDF][Supplementary PDF]

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study

Samuel Ritter, David G. T. Barrett, Adam Santoro, Matt M. Botvinick ; PMLR 70:2940-2949

[abs] [Download PDF]

Pain-Free Random Differential Privacy with Sensitivity Sampling

Benjamin I. P. Rubinstein, Francesco Aldà ; PMLR 70:2950-2959

[abs] [Download PDF]

Enumerating Distinct Decision Trees

Salvatore Ruggieri ; PMLR 70:2960-2968

[abs] [Download PDF]

Bayesian Boolean Matrix Factorisation

Tammo Rukat, Chris C. Holmes, Michalis K. Titsias, Christopher Yau ; PMLR 70:2969-2978

[abs] [Download PDF][Supplementary PDF]

Depth-Width Tradeoffs in Approximating Natural Functions with Neural Networks

Itay Safran, Ohad Shamir ; PMLR 70:2979-2987

[abs] [Download PDF][Supplementary PDF]

Asymmetric Tri-training for Unsupervised Domain Adaptation

Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada ; PMLR 70:2988-2997

[abs] [Download PDF][Supplementary PDF]

Semi-Supervised Classification Based on Classification from Positive and Unlabeled Data

Tomoya Sakai, Marthinus Christoffel Plessis, Gang Niu, Masashi Sugiyama ; PMLR 70:2998-3006

[abs] [Download PDF][Supplementary PDF]

Analytical Guarantees on Numerical Precision of Deep Neural Networks

Charbel Sakr, Yongjune Kim, Naresh Shanbhag ; PMLR 70:3007-3016

[abs] [Download PDF][Supplementary PDF]

Hierarchy Through Composition with Multitask LMDPs

Andrew M. Saxe, Adam C. Earle, Benjamin Rosman ; PMLR 70:3017-3026

[abs] [Download PDF][Supplementary PDF]

Optimal Algorithms for Smooth and Strongly Convex Distributed Optimization in Networks

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, Laurent Massoulié ; PMLR 70:3027-3036

[abs] [Download PDF][Supplementary PDF]

Adapting Kernel Representations Online Using Submodular Maximization

Matthew Schlegel, Yangchen Pan, Jiecao Chen, Martha White ; PMLR 70:3037-3046

[abs] [Download PDF][Supplementary PDF]

Developing Bug-Free Machine Learning Systems With Formal Mathematics

Daniel Selsam, Percy Liang, David L. Dill ; PMLR 70:3047-3056

[abs] [Download PDF]

Identifying Best Interventions through Online Importance Sampling

Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, Sanjay Shakkottai ; PMLR 70:3057-3066

[abs] [Download PDF][Supplementary PDF]

Failures of Gradient-Based Deep Learning

Shai Shalev-Shwartz, Ohad Shamir, Shaked Shammah ; PMLR 70:3067-3075

[abs] [Download PDF][Supplementary PDF]

Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit, Fredrik D. Johansson, David Sontag ; PMLR 70:3076-3085

[abs] [Download PDF][Supplementary PDF]

Online Learning with Local Permutations and Delayed Feedback

Ohad Shamir, Liran Szlak ; PMLR 70:3086-3094

[abs] [Download PDF][Supplementary PDF]

Orthogonalized ALS: A Theoretically Principled Tensor Decomposition Algorithm for Practical Use

Vatsal Sharan, Gregory Valiant ; PMLR 70:3095-3104

[abs] [Download PDF][Supplementary PDF]

Differentially Private Ordinary Least Squares

Or Sheffet ; PMLR 70:3105-3114

[abs] [Download PDF][Supplementary PDF]

On the Iteration Complexity of Support Recovery via Hard Thresholding Pursuit

Jie Shen, Ping Li ; PMLR 70:3115-3124

[abs] [Download PDF][Supplementary PDF]

GSOS: Gauss-Seidel Operator Splitting Algorithm for Multi-Term Nonsmooth Convex Composite Optimization

Li Shen, Wei Liu, Ganzhao Yuan, Shiqian Ma ; PMLR 70:3125-3134

[abs] [Download PDF]

World of Bits: An Open-Domain Platform for Web-Based Agents

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, Percy Liang ; PMLR 70:3135-3144

[abs] [Download PDF]

Learning Important Features Through Propagating Activation Differences

Avanti Shrikumar, Peyton Greenside, Anshul Kundaje ; PMLR 70:3145-3153

[abs] [Download PDF][Supplementary PDF]

Optimal Densification for Fast and Accurate Minwise Hashing

Anshumali Shrivastava ; PMLR 70:3154-3163

[abs] [Download PDF][Supplementary PDF]

Bottleneck Conditional Density Estimation

Rui Shu, Hung H. Bui, Mohammad Ghavamzadeh ; PMLR 70:3164-3172

[abs] [Download PDF][Supplementary PDF]

Attentive Recurrent Comparators

Pranav Shyam, Shubham Gupta, Ambedkar Dukkipati ; PMLR 70:3173-3181

[abs] [Download PDF]

Gradient Boosted Decision Trees for High Dimensional Sparse Output

Si Si, Huan Zhang, S. Sathiya Keerthi, Dhruv Mahajan, Inderjit S. Dhillon, Cho-Jui Hsieh ; PMLR 70:3182-3190

[abs] [Download PDF][Supplementary PDF]

The Predictron: End-To-End Learning and Planning

David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, Thomas Degris ; PMLR 70:3191-3199

[abs] [Download PDF][Supplementary PDF]

Fractional Langevin Monte Carlo: Exploring Levy Driven Stochastic Differential Equations for Markov Chain Monte Carlo

Umut Şimşekli ; PMLR 70:3200-3209

[abs] [Download PDF]

Nonparanormal Information Estimation

Shashank Singh, Barnabás Póczos ; PMLR 70:3210-3219

[abs] [Download PDF][Supplementary PDF]

High-Dimensional Structured Quantile Regression

Vidyashankar Sivakumar, Arindam Banerjee ; PMLR 70:3220-3229

[abs] [Download PDF]

Robust Budget Allocation via Continuous Submodular Functions

Matthew Staib, Stefanie Jegelka ; PMLR 70:3230-3240

[abs] [Download PDF][Supplementary PDF]

Probabilistic Submodular Maximization in Sub-Linear Time

Serban Stan, Morteza Zadimoghaddam, Andreas Krause, Amin Karbasi ; PMLR 70:3241-3250

[abs] [Download PDF][Supplementary PDF]

Approximate Steepest Coordinate Descent

Sebastian U. Stich, Anant Raj, Martin Jaggi ; PMLR 70:3251-3259

[abs] [Download PDF][Supplementary PDF]

Ordinal Graphical Models: A Tale of Two Approaches

Arun Sai Suggala, Eunho Yang, Pradeep Ravikumar ; PMLR 70:3260-3269

[abs] [Download PDF][Supplementary PDF]

Tensor Balancing on Statistical Manifold

Mahito Sugiyama, Hiroyuki Nakahara, Koji Tsuda ; PMLR 70:3270-3279

[abs] [Download PDF]

Safety-Aware Algorithms for Adversarial Contextual Bandit

Wen Sun, Debadeepta Dey, Ashish Kapoor ; PMLR 70:3280-3288

[abs] [Download PDF][Supplementary PDF]

Relative Fisher Information and Natural Gradient for Learning Large Modular Models

Ke Sun, Frank Nielsen ; PMLR 70:3289-3298

[abs] [Download PDF][Supplementary PDF]

meProp: Sparsified Back Propagation for Accelerated Deep Learning with Reduced Overfitting

Xu Sun, Xuancheng Ren, Shuming Ma, Houfeng Wang ; PMLR 70:3299-3308

[abs] [Download PDF]

Deeply AggreVaTeD: Differentiable Imitation Learning for Sequential Prediction

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, J. Andrew Bagnell ; PMLR 70:3309-3318

[abs] [Download PDF][Supplementary PDF]

Axiomatic Attribution for Deep Networks

Mukund Sundararajan, Ankur Taly, Qiqi Yan ; PMLR 70:3319-3328

[abs] [Download PDF]

Distributed Mean Estimation with Limited Communication

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, H. Brendan McMahan ; PMLR 70:3329-3337

[abs] [Download PDF][Supplementary PDF]

Selective Inference for Sparse High-Order Interaction Models

Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, Ichiro Takeuchi ; PMLR 70:3338-3347

[abs] [Download PDF][Supplementary PDF]

Coherent Probabilistic Forecasts for Hierarchical Time Series

Souhaib Ben Taieb, James W. Taylor, Rob J. Hyndman ; PMLR 70:3348-3357

[abs] [Download PDF][Supplementary PDF]

Partitioned Tensor Factorizations for Learning Mixed Membership Models

Zilong Tan, Sayan Mukherjee ; PMLR 70:3358-3367

[abs] [Download PDF]

Gradient Coding: Avoiding Stragglers in Distributed Learning

Rashish Tandon, Qi Lei, Alexandros G. Dimakis, Nikos Karampatziakis ; PMLR 70:3368-3376

[abs] [Download PDF][Supplementary PDF]

Gradient Projection Iterative Sketch for Large-Scale Constrained Least-Squares

Junqi Tang, Mohammad Golbabaee, Mike E. Davies ; PMLR 70:3377-3386

[abs] [Download PDF][Supplementary PDF]

Neural Networks and Rational Functions

Matus Telgarsky ; PMLR 70:3387-3393

[abs] [Download PDF][Supplementary PDF]

Stochastic DCA for the Large-sum of Non-convex Functions Problem and its Application to Group Variable Selection in Classification

Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, Bach Tran ; PMLR 70:3394-3403

[abs] [Download PDF]

An Analytical Formula of Population Gradient for two-layered ReLU network and its Applications in Convergence and Critical Point Analysis

Yuandong Tian ; PMLR 70:3404-3413

[abs] [Download PDF][Supplementary PDF]

Evaluating the Variance of Likelihood-Ratio Gradient Estimators

Seiya Tokui, Issei Sato ; PMLR 70:3414-3423

[abs] [Download PDF][Supplementary PDF]

Accelerating Eulerian Fluid Simulation With Convolutional Networks

Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, Ken Perlin ; PMLR 70:3424-3433

[abs] [Download PDF][Supplementary ZIP]

Boosted Fitted Q-Iteration

Samuele Tosatto, Matteo Pirotta, Carlo D'Eramo, Marcello Restelli ; PMLR 70:3434-3443

[abs] [Download PDF][Supplementary PDF]

Diameter-Based Active Learning

Christopher Tosh, Sanjoy Dasgupta ; PMLR 70:3444-3452

[abs] [Download PDF]

Magnetic Hamiltonian Monte Carlo

Nilesh Tripuraneni, Mark Rowland, Zoubin Ghahramani, Richard Turner ; PMLR 70:3453-3461

[abs] [Download PDF][Supplementary PDF]

Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs

Rakshit Trivedi, Hanjun Dai, Yichen Wang, Le Song ; PMLR 70:3462-3471

[abs] [Download PDF][Supplementary PDF]

Hyperplane Clustering via Dual Principal Component Pursuit

Manolis C. Tsakiris, René Vidal ; PMLR 70:3472-3481

[abs] [Download PDF]

Breaking Locality Accelerates Block Gauss-Seidel

Stephen Tu, Shivaram Venkataraman, Ashia C. Wilson, Alex Gittens, Michael I. Jordan, Benjamin Recht ; PMLR 70:3482-3491

[abs] [Download PDF]

Multilabel Classification with Group Testing and Codes

Shashanka Ubaru, Arya Mazumdar ; PMLR 70:3492-3501

[abs] [Download PDF][Supplementary PDF]

Learning Stable Stochastic Nonlinear Dynamical Systems

Jonas Umlauft, Sandra Hirche ; PMLR 70:3502-3510

[abs] [Download PDF]

Learning Determinantal Point Processes with Moments and Cycles

John Urschel, Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet ; PMLR 70:3511-3520

[abs] [Download PDF]

Automatic Discovery of the Statistical Types of Variables in a Dataset

Isabel Valera, Zoubin Ghahramani ; PMLR 70:3521-3529

[abs] [Download PDF]

Model-Independent Online Learning for Influence Maximization

Sharan Vaswani, Branislav Kveton, Zheng Wen, Mohammad Ghavamzadeh, Laks V. S. Lakshmanan, Mark Schmidt ; PMLR 70:3530-3539

[abs] [Download PDF][Supplementary PDF]

FeUdal Networks for Hierarchical Reinforcement Learning

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, Koray Kavukcuoglu ; PMLR 70:3540-3549

[abs] [Download PDF][Supplementary ZIP]

Scalable Multi-Class Gaussian Process Classification using Expectation Propagation

Carlos Villacampa-Calvo, Daniel Hernández-Lobato ; PMLR 70:3550-3559

[abs] [Download PDF][Supplementary ZIP]

Learning to Generate Long-term Future via Hierarchical Prediction

Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, Honglak Lee ; PMLR 70:3560-3569

[abs] [Download PDF][Supplementary PDF]

On orthogonality and learning recurrent networks with long term dependencies

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, Chris Pal ; PMLR 70:3570-3578

[abs] [Download PDF]

Fast Bayesian Intensity Estimation for the Permanental Process

Christian J. Walder, Adrian N. Bishop ; PMLR 70:3579-3588

[abs] [Download PDF][Supplementary PDF]

Optimal and Adaptive Off-policy Evaluation in Contextual Bandits

Yu-Xiang Wang, Alekh Agarwal, Miroslav Dudik ; PMLR 70:3589-3597

[abs] [Download PDF][Supplementary PDF]

Capacity Releasing Diffusion for Speed and Locality

Di Wang, Kimon Fountoulakis, Monika Henzinger, Michael W. Mahoney, Satish Rao ; PMLR 70:3598-3607

[abs] [Download PDF][Supplementary PDF]

Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging

Shusen Wang, Alex Gittens, Michael W. Mahoney ; PMLR 70:3608-3616

[abs] [Download PDF]

Robust Gaussian Graphical Model Estimation with Arbitrary Corruption

Lingxiao Wang, Quanquan Gu ; PMLR 70:3617-3626

[abs] [Download PDF]

Max-value Entropy Search for Efficient Bayesian Optimization

Zi Wang, Stefanie Jegelka ; PMLR 70:3627-3635

[abs] [Download PDF][Supplementary PDF]

Efficient Distributed Learning with Sparsity

Jialei Wang, Mladen Kolar, Nathan Srebro, Tong Zhang ; PMLR 70:3636-3645

[abs] [Download PDF][Supplementary PDF]

Robust Probabilistic Modeling with Bayesian Data Reweighting

Yixin Wang, Alp Kucukelbir, David M. Blei ; PMLR 70:3646-3655

[abs] [Download PDF][Supplementary PDF]

Batched High-dimensional Bayesian Optimization via Structural Kernel Learning

Zi Wang, Chengtao Li, Stefanie Jegelka, Pushmeet Kohli ; PMLR 70:3656-3664

[abs] [Download PDF][Supplementary PDF]

Tensor Decomposition via Simultaneous Power Iteration

Po-An Wang, Chi-Jen Lu ; PMLR 70:3665-3673

[abs] [Download PDF][Supplementary PDF]

Sequence Modeling via Segmentations

Chong Wang, Yining Wang, Po-Sen Huang, Abdelrahman Mohamed, Dengyong Zhou, Li Deng ; PMLR 70:3674-3683

[abs] [Download PDF]

Variational Policy for Guiding Point Processes

Yichen Wang, Grady Williams, Evangelos Theodorou, Le Song ; PMLR 70:3684-3693

[abs] [Download PDF][Supplementary PDF]

Exploiting Strong Convexity from Data with Primal-Dual First-Order Algorithms

Jialei Wang, Lin Xiao ; PMLR 70:3694-3702

[abs] [Download PDF][Supplementary PDF]

Beyond Filters: Compact Feature Map for Portable Deep Model

Yunhe Wang, Chang Xu, Chao Xu, Dacheng Tao ; PMLR 70:3703-3711

[abs] [Download PDF][Supplementary ZIP]

A Unified Variance Reduction-Based Framework for Nonconvex Low-Rank Matrix Recovery

Lingxiao Wang, Xiao Zhang, Quanquan Gu ; PMLR 70:3712-3721

[abs] [Download PDF][Supplementary PDF]

Source-Target Similarity Modelings for Multi-Source Transfer Gaussian Process Regression

Pengfei Wei, Ramon Sagarna, Yiping Ke, Yew-Soon Ong, Chi-Keong Goh ; PMLR 70:3722-3731

[abs] [Download PDF]

Latent Intention Dialogue Models

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, Steve Young ; PMLR 70:3732-3741

[abs] [Download PDF]

Unifying Task Specification in Reinforcement Learning

Martha White ; PMLR 70:3742-3750

[abs] [Download PDF][Supplementary PDF]

Learned Optimizers that Scale and Generalize

Olga Wichrowska, Niru Maheswaranathan, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Nando Freitas, Jascha Sohl-Dickstein ; PMLR 70:3751-3760

[abs] [Download PDF][Supplementary ZIP]

Exact Inference for Integer Latent-Variable Models

Kevin Winner, Debora Sujono, Dan Sheldon ; PMLR 70:3761-3770

[abs] [Download PDF][Supplementary PDF]

Tensor Belief Propagation

Andrew Wrigley, Wee Sun Lee, Nan Ye ; PMLR 70:3771-3779

[abs] [Download PDF][Supplementary PDF]

A Unified View of Multi-Label Performance Measures

Xi-Zhu Wu, Zhi-Hua Zhou ; PMLR 70:3780-3788

[abs] [Download PDF][Supplementary PDF]

Dual Supervised Learning

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, Tie-Yan Liu ; PMLR 70:3789-3798

[abs] [Download PDF][Supplementary PDF]

Learning Latent Space Models with Angular Constraints

Pengtao Xie, Yuntian Deng, Yi Zhou, Abhimanu Kumar, Yaoliang Yu, James Zou, Eric P. Xing ; PMLR 70:3799-3810

[abs] [Download PDF]

Uncorrelation and Evenness: a New Diversity-Promoting Regularizer

Pengtao Xie, Aarti Singh, Eric P. Xing ; PMLR 70:3811-3820

[abs] [Download PDF][Supplementary PDF]

Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence

Yi Xu, Qihang Lin, Tianbao Yang ; PMLR 70:3821-3830

[abs] [Download PDF][Supplementary PDF]

Learning Hawkes Processes from Short Doubly-Censored Event Sequences

Hongteng Xu, Dixin Luo, Hongyuan Zha ; PMLR 70:3831-3840

[abs] [Download PDF][Supplementary PDF]

Adaptive Consensus ADMM for Distributed Optimization

Zheng Xu, Gavin Taylor, Hao Li, Mário A. T. Figueiredo, Xiaoming Yuan, Tom Goldstein ; PMLR 70:3841-3850

[abs] [Download PDF][Supplementary PDF]

High-dimensional Non-Gaussian Single Index Models via Thresholded Score Function Estimation

Zhuoran Yang, Krishnakumar Balasubramanian, Han Liu ; PMLR 70:3851-3860

[abs] [Download PDF][Supplementary ZIP]

Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, Mingyi Hong ; PMLR 70:3861-3870

[abs] [Download PDF][Supplementary PDF]

On The Projection Operator to A Three-view Cardinality Constrained Set

Haichuan Yang, Shupeng Gui, Chuyang Ke, Daniel Stefankovic, Ryohei Fujimaki, Ji Liu ; PMLR 70:3871-3880

[abs] [Download PDF][Supplementary PDF]

Improved Variational Autoencoders for Text Modeling using Dilated Convolutions

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, Taylor Berg-Kirkpatrick ; PMLR 70:3881-3890

[abs] [Download PDF][Supplementary PDF]

Tensor-Train Recurrent Neural Networks for Video Classification

Yinchong Yang, Denis Krompass, Volker Tresp ; PMLR 70:3891-3900

[abs] [Download PDF]

A Richer Theory of Convex Constrained Optimization with Reduced Projections and Improved Rates

Tianbao Yang, Qihang Lin, Lijun Zhang ; PMLR 70:3901-3910

[abs] [Download PDF]

Sparse + Group-Sparse Dirty Models: Statistical Guarantees without Unreasonable Conditions and a Case for Non-Convexity

Eunho Yang, Aurélie C. Lozano ; PMLR 70:3911-3920

[abs] [Download PDF][Supplementary PDF]

Scalable Bayesian Rule Lists

Hongyu Yang, Cynthia Rudin, Margo Seltzer ; PMLR 70:3921-3930

[abs] [Download PDF]

Approximate Newton Methods and Their Local Convergence

Haishan Ye, Luo Luo, Zhihua Zhang ; PMLR 70:3931-3939

[abs] [Download PDF][Supplementary PDF]

A Simulated Annealing Based Inexact Oracle for Wasserstein Loss Minimization

Jianbo Ye, James Z. Wang, Jia Li ; PMLR 70:3940-3948

[abs] [Download PDF][Supplementary PDF]

Latent Feature Lasso

Ian En-Hsu Yen, Wei-Cheng Lee, Sung-En Chang, Arun Sai Suggala, Shou-De Lin, Pradeep Ravikumar ; PMLR 70:3949-3957

[abs] [Download PDF][Supplementary PDF]

Combined Group and Exclusive Sparsity for Deep Neural Networks

Jaehong Yoon, Sung Ju Hwang ; PMLR 70:3958-3966

[abs] [Download PDF]

Latent LSTM Allocation: Joint Clustering and Non-Linear Dynamic Modeling of Sequence Data

Manzil Zaheer, Amr Ahmed, Alexander J. Smola ; PMLR 70:3967-3976

[abs] [Download PDF]

Canopy Fast Sampling with Cover Trees

Manzil Zaheer, Satwik Kottur, Amr Ahmed, José Moura, Alex Smola ; PMLR 70:3977-3986

[abs] [Download PDF][Supplementary PDF]

Continual Learning Through Synaptic Intelligence

Friedemann Zenke, Ben Poole, Surya Ganguli ; PMLR 70:3987-3995

[abs] [Download PDF][Supplementary PDF]

Stochastic Gradient Monomial Gamma Sampler

Yizhe Zhang, Changyou Chen, Zhe Gan, Ricardo Henao, Lawrence Carin ; PMLR 70:3996-4005

[abs] [Download PDF]

Adversarial Feature Matching for Text Generation

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, Lawrence Carin ; PMLR 70:4006-4015

[abs] [Download PDF]

Scaling Up Sparse Support Vector Machines by Simultaneous Feature and Sample Reduction

Weizhong Zhang, Bin Hong, Wei Liu, Jieping Ye, Deng Cai, Xiaofei He, Jie Wang ; PMLR 70:4016-4025

[abs] [Download PDF][Supplementary PDF]

Re-revisiting Learning on Hypergraphs: Confidence Interval and Subgradient Method

Chenzi Zhang, Shuguang Hu, Zhihao Gavin Tang, T-H. Hubert Chan ; PMLR 70:4026-4034

[abs] [Download PDF]

ZipML: Training Linear Models with End-to-End Low Precision, and a Little Bit of Deep Learning

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, Ce Zhang ; PMLR 70:4035-4043

[abs] [Download PDF]

Convexified Convolutional Neural Networks

Yuchen Zhang, Percy Liang, Martin J. Wainwright ; PMLR 70:4044-4053

[abs] [Download PDF][Supplementary PDF]

Projection-free Distributed Online Learning in Networks

Wenpeng Zhang, Peilin Zhao, Wenwu Zhu, Steven C. H. Hoi, Tong Zhang ; PMLR 70:4054-4062

[abs] [Download PDF][Supplementary PDF]

Multi-Class Optimal Margin Distribution Machine

Teng Zhang, Zhi-Hua Zhou ; PMLR 70:4063-4071

[abs] [Download PDF]

Leveraging Node Attributes for Incomplete Relational Data

He Zhao, Lan Du, Wray Buntine ; PMLR 70:4072-4081

[abs] [Download PDF]

Theoretical Properties for Neural Networks with Weight Matrices of Low Displacement Rank

Liang Zhao, Siyu Liao, Yanzhi Wang, Zhe Li, Jian Tang, Bo Yuan ; PMLR 70:4082-4090

[abs] [Download PDF]

Learning Hierarchical Features from Deep Generative Models

Shengjia Zhao, Jiaming Song, Stefano Ermon ; PMLR 70:4091-4099

[abs] [Download PDF][Supplementary PDF]

Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture

Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S. Jaakkola, Matt T. Bianchi ; PMLR 70:4100-4109

[abs] [Download PDF]

Follow the Moving Leader in Deep Learning

Shuai Zheng, James T. Kwok ; PMLR 70:4110-4119

[abs] [Download PDF][Supplementary PDF]

Asynchronous Stochastic Gradient Descent with Delay Compensation

Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, Tie-Yan Liu ; PMLR 70:4120-4129

[abs] [Download PDF][Supplementary PDF]

Collect at Once, Use Effectively: Making Non-interactive Locally Private Learning Possible

Kai Zheng, Wenlong Mou, Liwei Wang ; PMLR 70:4130-4139

[abs] [Download PDF][Supplementary PDF]

Recovery Guarantees for One-hidden-layer Neural Networks

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, Inderjit S. Dhillon ; PMLR 70:4140-4149

[abs] [Download PDF]

Stochastic Adaptive Quasi-Newton Methods for Minimizing Expected Values

Chaoxu Zhou, Wenbo Gao, Donald Goldfarb ; PMLR 70:4150-4159

[abs] [Download PDF][Supplementary PDF]

Identify the Nash Equilibrium in Static Games with Random Payoffs

Yichi Zhou, Jialian Li, Jun Zhu ; PMLR 70:4160-4169

[abs] [Download PDF][Supplementary PDF]

When can Multi-Site Datasets be Pooled for Regression? Hypothesis Tests, ℓ_2-consistency and Neuroscience Applications

Hao Henry Zhou, Yilin Zhang, Vamsi K. Ithapu, Sterling C. Johnson, Grace Wahba, Vikas Singh ; PMLR 70:4170-4179

[abs] [Download PDF][Supplementary PDF]

High-Dimensional Variance-Reduced Stochastic Gradient Expectation-Maximization Algorithm

Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, Quanquan Gu ; PMLR 70:4180-4188

[abs] [Download PDF]

Recurrent Highway Networks

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutnik, Jürgen Schmidhuber ; PMLR 70:4189-4198

[abs] [Download PDF][Supplementary PDF]

Online Learning to Rank in Stochastic Click Models

Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, Zheng Wen ; PMLR 70:4199-4208

[abs] [Download PDF][Supplementary PDF]

subscribe via RSS

This site last compiled Thu, 27 Jul 2017 21:10:49 +0000

Github Account

Copyright © PMLR 2017. All rights reserved.

Appendix CUTURI02

https://web.archive.org/web/20171101000000*/http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf December 19, 2025

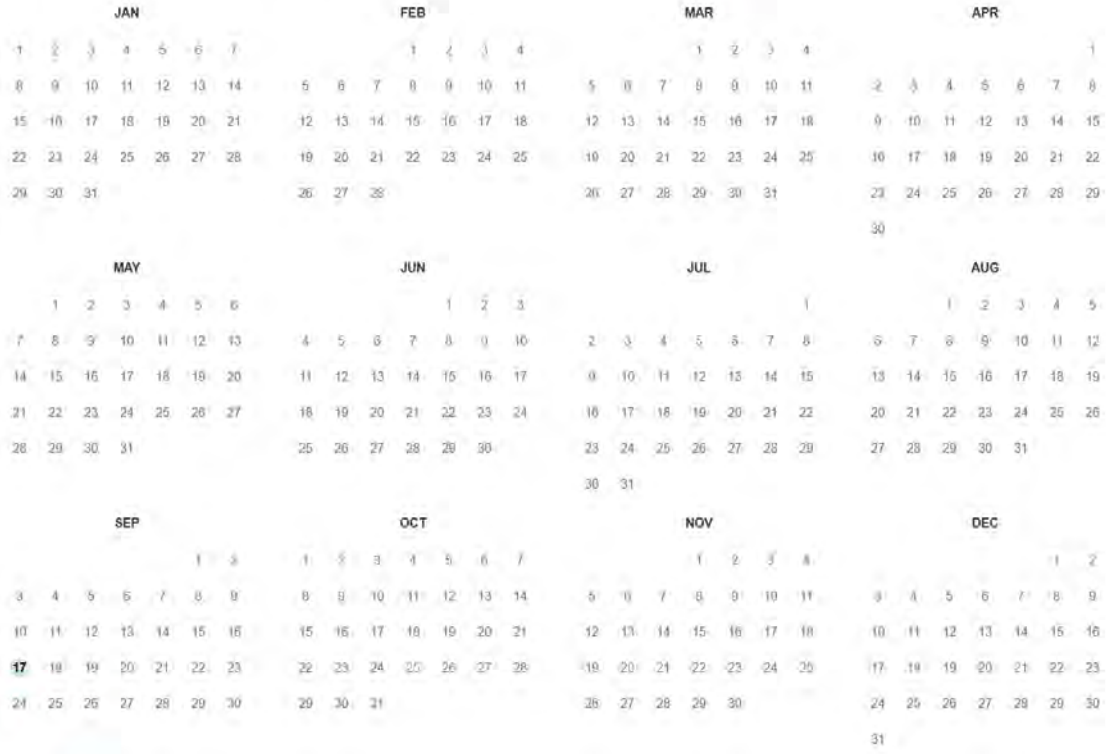
INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE INTERNET ARCHIVE WayBackMachine Explore more than 1 trillion web pages saved over time
http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **10 times** between September 17, 2017 and August 7, 2025.





Note

This calendar view maps the number of times **http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20170917200602/http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf December 19, 2025

INTERNET ARCHIVE WaybackMachine http://proceedings.mlr.press/v70/cuturi17a/cuturi17a.pdf Go AUG SEP 17 OCT 2016 2017 2018 About this capture
10 captures
17 Sep 2017 - 7 Aug 2025

cuturi17a.pdf 1 / 10 100%



1



2



Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi [1] Mathieu Blondel [2]

Abstract

We propose in this paper a differentiable learning loss between time series, building upon the celebrated dynamic time warping (DTW) discrepancy. Unlike the Euclidean distance, DTW can compare time series of variable size and is robust to shifts or dilatations across the time dimension. To compute DTW, one typically solves a minimal-cost alignment problem between two time series using dynamic programming. Our work takes advantage of a smoothed formulation of DTW, called soft-DTW, that computes the soft-minimum of all alignment costs. We show in this paper that soft-DTW is a *differentiable*

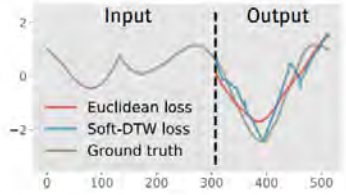


Figure 1. Given the first part of a time series, we trained two multi-layer perceptron (MLP) to predict the entire second part. Using the ShapesAll dataset, we used a Euclidean loss for the first MLP and the soft-DTW loss proposed in this paper for the second

Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi [1] **Mathieu Blondel** [2]

Abstract

We propose in this paper a differentiable learning loss between time series, building upon the celebrated dynamic time warping (DTW) discrepancy. Unlike the Euclidean distance, DTW can compare time series of variable size and is robust to shifts or dilatations across the time dimension. To compute DTW, one typically solves a minimal-cost alignment problem between two time series using dynamic programming. Our work takes advantage of a smoothed formulation of DTW, called soft-DTW, that computes the soft-minimum of all alignment costs. We show in this paper that soft-DTW is a *differentiable* loss function, and that both its value and gradient can be computed with quadratic time/space complexity (DTW has quadratic time but linear space complexity). We show that this regularization is particularly well suited to average and cluster time series under the DTW geometry, a task for which our proposal significantly outperforms existing baselines (Petitjean et al., 2011). Next, we propose to tune the parameters of a machine that outputs time series by minimizing its fit with ground-truth labels in a soft-DTW sense.

1. Introduction

The goal of supervised learning is to learn a mapping that links an input to an output objects, using examples of such pairs. This task is noticeably more difficult when the output objects have a structure, *i.e.* when they are not vectors (Bakir et al., 2007). We study here the case where each output object is a *time series*, namely a family of observations indexed by time. While it is tempting to treat time as yet another feature, and handle time series of vectors as the concatenation of all these vectors, several practical issues arise when taking this simplistic approach: Time-indexed phenomena can often be stretched in some areas along the time axis (a word uttered in a slightly slower pace than usual) with no impact on their characteristics; varying sampling conditions may mean they have different lengths; time series may not synchronized.

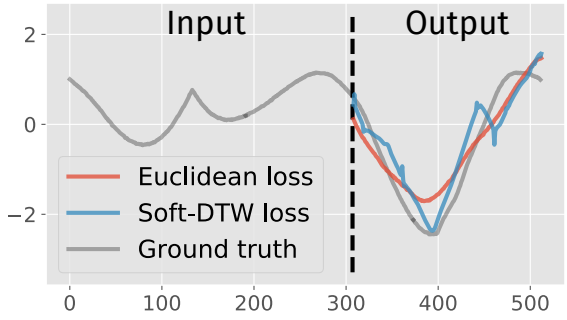


Figure 1. Given the first part of a time series, we trained two multi-layer perceptron (MLP) to predict the entire second part. Using the ShapesAll dataset, we used a Euclidean loss for the first MLP and the soft-DTW loss proposed in this paper for the second one. We display above the prediction obtained for a given test instance with either of these two MLPs in addition to the ground truth. Oftentimes, we observe that the soft-DTW loss enables us to better predict sharp changes. More time series predictions are given in Appendix F.

The DTW paradigm. Generative models for time series are usually built having the invariances above in mind: Such properties are typically handled through latent variables and/or Markovian assumptions (Lütkepohl, 2005, Part I,§18). A simpler approach, motivated by geometry, lies in the direct definition of a discrepancy between time series that encodes these invariances, such as the Dynamic Time Warping (DTW) score (Sakoe & Chiba, 1971; 1978). DTW computes the best possible alignment between two time series (the optimal alignment itself can also be of interest, see e.g. Garreau et al. 2014) of respective length n and m by computing first the $n \times m$ pairwise distance matrix between these points to solve then a dynamic program (DP) using Bellman's recursion with a quadratic (nm) cost.

The DTW geometry. Because it encodes efficiently a useful class of invariances, DTW has often been used in a discriminative framework (with a k-NN or SVM classifier) to predict a real or a class label output, and engineered to run

[1]CREST, ENSAE, Université Paris-Saclay, France [2]NTT Communication Science Laboratories, Seika-cho, Kyoto, Japan. Correspondence to: Marco Cuturi <marco.cuturi@ensae.fr>, Mathieu Blondel <mathieu@mblondel.org>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

IPR2026-00272
Krisp Technologies EX1010 Page 296

faster in that context (Yi et al., 1998). Recent works by Petitjean et al. (2011); Petitjean & Gançarski (2012) have, however, shown that DTW can be used for more innovative tasks, such as time series *averaging* using the DTW discrepancy (see Schultz & Jain 2017 for a gentle introduction to these ideas). More generally, the idea of synthetising time series centroids can be regarded as a first attempt to *output* entire time series using DTW as a fitting loss. From a computational perspective, these approaches are, however, hampered by the fact that DTW is not differentiable and unstable when used in an optimization pipeline.

Soft-DTW. In parallel to these developments, several authors have considered smoothed modifications of Bellman's recursion to define smoothed DP distances (Bahl & Jelinek, 1975; Ristad & Yianilos, 1998) or kernels (Saigo et al., 2004; Cuturi et al., 2007). When applied to the DTW discrepancy, that regularization results in a *soft-DTW* score, which considers the *soft-minimum* of the distribution of *all costs* spanned by *all* possible alignments between two time series. Despite considering all alignments and not just the optimal one, soft-DTW can be computed with a minor modification of Bellman's recursion, in which all $(\min, +)$ operations are replaced with $(+, \times)$. As a result, both DTW and soft-DTW have quadratic in time & linear in space complexity with respect to the sequences' lengths. Because soft-DTW can be used with kernel machines, one typically observes an increase in performance when using soft-DTW over DTW (Cuturi, 2011) for classification.

Our contributions. We explore in this paper another important benefit of smoothing DTW: unlike the original DTW discrepancy, soft-DTW is *differentiable* in all of its arguments. We show that the gradients of soft-DTW w.r.t to all of its variables can be computed as a by-product of the computation of the discrepancy itself, with an added quadratic storage cost. We use this fact to propose an alternative approach to the DBA (DTW Barycenter Averaging) clustering algorithm of (Petitjean et al., 2011), and observe that our smoothed approach significantly outperforms known baselines for that task. More generally, we propose to use soft-DTW as a *fitting term* to compare the output of a machine synthesizing a time series segment with a ground truth observation, in the same way that, for instance, a regularized Wasserstein distance was used to compute barycenters (Cuturi & Doucet, 2014), and later to fit discriminators that output histograms (Zhang et al., 2015; Rolet et al., 2016). When paired with a flexible learning architecture such as a neural network, soft-DTW allows for a differentiable end-to-end approach to design predictive and generative models for time series, as illustrated in Figure 1. Source code is available at `https://github.com/mblondel/soft-dtw`.

Structure. After providing background material, we show in §2 how soft-DTW can be differentiated w.r.t the locations of two time series. We follow in §3 by illustrating how these results can be directly used for tasks that require to output time series: averaging, clustering and prediction of time series. We close this paper with experimental results in §4 that showcase each of these potential applications.

Notations. We consider in what follows multivariate discrete time series of varying length taking values in $\Omega \subset \mathbb{R}^p$. A time series can be thus represented as a matrix of p lines and varying number of columns. We consider a differentiable substitution-cost function $\delta : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ which will be, in most cases, the quadratic Euclidean distance between two vectors. For an integer n we write $[\![n]\!]$ for the set $\{1, \dots, n\}$ of integers. Given two series' lengths n and m, we write $\mathcal{A}_{n,m} \subset \{0,1\}^{n \times m}$ for the set of (binary) alignment matrices, that is paths on a $n \times m$ matrix that connect the upper-left $(1,1)$ matrix entry to the lower-right (n,m) one using only $\downarrow, \rightarrow, \searrow$ moves. The cardinal of $\mathcal{A}_{n,m}$ is known as the delannoy$(n-1, m-1)$ number; that number grows exponentially with m and n.

2. The DTW and soft-DTW loss functions

We propose in this section a unified formulation for the original DTW discrepancy (Sakoe & Chiba, 1978) and the Global Alignment kernel (GAK) (Cuturi et al., 2007), which can be both used to compare two time series $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^{p \times n}$ and $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^{p \times m}$.

2.1. Alignment costs: optimality and sum

Given the cost matrix $\Delta(\mathbf{x}, \mathbf{y}) := \big[\delta(x_i, y_j)\big]_{ij} \in \mathbb{R}^{n \times m}$, the inner product $\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle$ of that matrix with an alignment matrix A in $\mathcal{A}_{n,m}$ gives the score of A, as illustrated in Figure 2. Both DTW and GAK consider the costs of all possible alignment matrices, yet do so differently:

$$\begin{aligned} \mathrm{DTW}(\mathbf{x}, \mathbf{y}) &:= \min_{A \in \mathcal{A}_{n,m}} \langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle, \\ k_{\mathrm{GA}}^{\gamma}(\mathbf{x}, \mathbf{y}) &:= \sum_{A \in \mathcal{A}_{n,m}} e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma}. \end{aligned} \tag{1}$$

DP Recursion. Sakoe & Chiba (1978) showed that the Bellman equation (1952) can be used to compute DTW. That recursion, which appears in line 5 of Algorithm 1 (disregarding for now the exponent γ), only involves $(\min, +)$ operations. When considering kernel k_{GA}^{γ} and, instead, its integration over all alignments (see e.g. Lasserre 2009), Cuturi et al. (2007, Theorem 2) and the highly related formulation of Saigo et al. (2004, p.1685) use an old algorithmic appraoch (Bahl & Jelinek, 1975) which consists in *(i)* replacing all costs by their neg-exponential; *(ii)* replace $(\min, +)$ operations with $(+, \times)$ operations. These two recursions can be in fact unified with the use of a soft-

IPR2026-00272
Krisp Technologies EX1010 Page 297

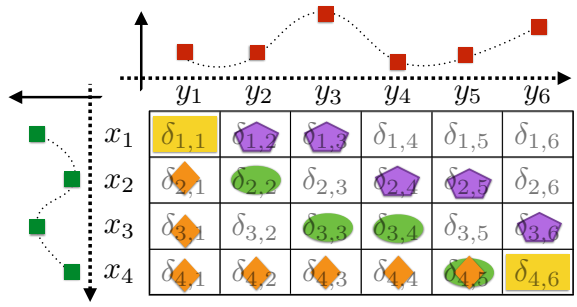


Figure 2. Three alignment matrices (orange, green, purple, in addition to the top-left and bottom-right entries) between two time series of length 4 and 6. The cost of an alignment is equal to the sum of entries visited along the path. DTW only considers the optimal alignment (here depicted in purple pentagons), whereas soft-DTW considers all delannoy$(n-1, m-1)$ possible alignment matrices.

minimum operator, which we present below.

Unified algorithm Both formulas in Eq. (1) can be computed with a single algorithm. That formulation is new to our knowledge. Consider the following generalized $\min$ operator, with a smoothing parameter $\gamma \geq 0$:

$$\min{}^{\gamma}\{a_1, \dots, a_n\} \coloneqq \begin{cases} \min_{i \leq n} a_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma}, & \gamma > 0. \end{cases}$$

With that operator, we can define γ-soft-DTW:

$$\boxed{\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y}) \coloneqq \min{}^{\gamma}\{\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle, A \in \mathcal{A}_{n,m}\}.}$$

The original DTW score is recovered by setting γ to 0. When $\gamma > 0$, we recover $\mathbf{dtw}_\gamma = -\gamma \log k^{\gamma}_{\mathrm{GA}}$. Most importantly, and in either case, $\mathbf{dtw}_\gamma$ can be computed using Algorithm 1, which requires (nm) operations and (nm) storage cost as well . That cost can be reduced to $2n$ with a more careful implementation if one only seeks to compute $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$, but the backward pass we consider next requires the entire matrix R of intermediary alignment costs. Note that, to ensure numerical stability, the operator $\min^{\gamma}$ must be computed using the usual log-sum-exp stabilization trick, namely that $\log \sum_i e^{z_i} = (\max_j z_j) + \log \sum_i e^{z_i - \max_j z_j}$.

2.2. Differentiation of soft-DTW

A small variation in the input $\mathbf{x}$ causes a small change in $\mathbf{dtw}_0(\mathbf{x}, \mathbf{y})$ or $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$. When considering $\mathbf{dtw}_0$, that change can be efficiently monitored only when the optimal alignment matrix $A^\star$ that arises when computing $\mathbf{dtw}_0(\mathbf{x}, \mathbf{y})$ in Eq. (1) is unique. As the minimum over a finite set of linear functions of Δ, $\mathbf{dtw}_0$ is therefore locally differentiable w.r.t. the cost matrix Δ, with gradient $A^\star$, a fact that has been exploited in all algorithms designed to

Algorithm 1 Forward recursion to compute $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$ and intermediate alignment costs

Inputs: $\mathbf{x}, \mathbf{y}$, smoothing $\gamma \geq 0$, distance function δ
$r_{0,0} = 0; r_{i,0} = r_{0,j} = \infty; i \in [\![n]\!], j \in [\![m]\!]$
for $j = 1, \dots, m$ **do**
  for $i = 1, \dots, n$ **do**
    $r_{i,j} = \delta(x_i, y_j) + \min^{\gamma}\{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\}$
  end for
end for
Output: $(r_{n,m}, R)$

average time series under the DTW metric (Petitjean et al., 2011; Schultz & Jain, 2017). To recover the gradient of $\mathbf{dtw}_0(\mathbf{x}, \mathbf{y})$ w.r.t. $\mathbf{x}$, we only need to apply the chain rule, thanks to the differentiability of the cost function:

$$\nabla_{\mathbf{x}} \mathbf{dtw}_0(\mathbf{x}, \mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^T A^\star, \qquad (2)$$

where $\partial \Delta(\mathbf{x}, \mathbf{y}) / \partial \mathbf{x}$ is the Jacobian of Δ w.r.t. $\mathbf{x}$, a linear map from $\mathbb{R}^{p \times n}$ to $\mathbb{R}^{n \times m}$. When δ is the squared Euclidean distance, the transpose of that Jacobian applied to a matrix $B \in \mathbb{R}^{n \times m}$ is ($\circ$ being the elementwise product):

$$(\partial \Delta(\mathbf{x}, \mathbf{y}) / \partial \mathbf{x})^T B = 2\left((\mathbf{1}_p \mathbf{1}_m^T B^T) \circ \mathbf{x} - \mathbf{y} B^T\right).$$

With continuous data, $A^\star$ is almost always likely to be unique, and therefore the gradient in Eq. (2) will be defined almost everywhere. However, that gradient, when it exists, will be discontinuous around those values $\mathbf{x}$ where a small change in $\mathbf{x}$ causes a change in $A^\star$, which is likely to hamper the performance of gradient descent methods.

The case $\gamma > 0$. An immediate advantage of soft-DTW is that it can be explicitly differentiated, a fact that was also noticed by Saigo et al. (2006) in the related case of edit distances. When $\gamma > 0$, the gradient of Eq. (1) is obtained via the chain rule,

$$\nabla_{\mathbf{x}} \mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^T \mathbb{E}_\gamma[A], \qquad (3)$$

$$\text{where} \quad \mathbb{E}_\gamma[A] \coloneqq \frac{1}{k^{\gamma}_{\mathrm{GA}}(\mathbf{x}, \mathbf{y})} \sum_{A \in \mathcal{A}_{n,m}} e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma} A,$$

is the average alignment matrix A under the Gibbs distribution $p_\gamma \propto e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma}$ defined on all alignments in $\mathcal{A}_{n,m}$. The kernel $k^{\gamma}_{\mathrm{GA}}(\mathbf{x}, \mathbf{y})$ can thus be interpreted as the normalization constant of p_γ. Of course, since $\mathcal{A}_{n,m}$ has exponential size in n and m, a naive summation is not tractable. Although a Bellman recursion to compute that average alignment matrix $\mathbb{E}_\gamma[A]$ exists (see Appendix A) that computation has *quartic* $(n^2 m^2)$ complexity. Note that

IPR2026-00272
Krisp Technologies EX1010 Page 298

this stands in stark contrast to the quadratic complexity obtained by Saigo et al. (2006) for edit-distances, which is due to the fact the sequences they consider can only take values in a *finite* alphabet. To compute the gradient of soft-DTW, we propose instead an algorithm that manages to remain *quadratic* (nm) in terms of complexity. The key to achieve this reduction is to apply the chain rule in *reverse* order of Bellman's recursion given in Algorithm 1, namely back-propagate. A similar idea was recently used to compute the gradient of ANOVA kernels in (Blondel et al., 2016).

2.3. Algorithmic differentiation

Differentiating algorithmically $\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$ requires doing first a forward pass of Bellman's equation to store all intermediary computations and recover $R = [r_{i,j}]$ when running Algorithm 1. The value of $\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$—stored in $r_{n,m}$ at the end of the forward recursion—is then impacted by a change in $r_{i,j}$ exclusively through the terms in which $r_{i,j}$ plays a role, namely the triplet of terms $r_{i+1,j}, r_{i,j+1}, r_{i+1,j+1}$. A straightforward application of the chain rule then gives

$$\underbrace{\frac{\partial r_{n,m}}{\partial r_{i,j}}}_{e_{i,j}} = \underbrace{\frac{\partial r_{n,m}}{\partial r_{i+1,j}}}_{e_{i+1,j}} \frac{\partial r_{i+1,j}}{\partial r_{i,j}} + \underbrace{\frac{\partial r_{n,m}}{\partial r_{i,j+1}}}_{e_{i,j+1}} \frac{\partial r_{i,j+1}}{\partial r_{i,j}} + \underbrace{\frac{\partial r_{n,m}}{\partial r_{i+1,j+1}}}_{e_{i+1,j+1}} \frac{\partial r_{i+1,j+1}}{\partial r_{i,j}},$$

in which we have defined the notation of the main object of interest of the backward recursion: $e_{i,j} := \frac{\partial r_{n,m}}{\partial r_{i,j}}$. The Bellman recursion evaluated at $(i+1,j)$ as shown in line 5 of Algorithm 1 (here $\delta_{i+1,j}$ is $\delta(x_{i+1}, y_j)$) yields :

$$r_{i+1,j} = \delta_{i+1,j} + \min{}^\gamma\{r_{i,j-1}, r_{i,j}, r_{i+1,j-1}\},$$

which, when differentiated w.r.t $r_{i,j}$ yields the ratio:

$$\frac{\partial r_{i+1,j}}{\partial r_{i,j}} = e^{-r_{i,j}/\gamma} \Big/ \left(e^{-r_{i,j-1}/\gamma} + e^{-r_{i,j}/\gamma} + e^{-r_{i+1,j-1}/\gamma}\right).$$

The logarithm of that derivative can be conveniently cast using evaluations of $\min^\gamma$ computed in the forward loop:

$$\begin{aligned}\gamma \log \frac{\partial r_{i+1,j}}{\partial r_{i,j}} &= \min{}^\gamma\{r_{i,j-1}, r_{i,j}, r_{i+1,j-1}\} - r_{i,j} \\ &= r_{i+1,j} - \delta_{i+1,j} - r_{i,j}.\end{aligned}$$

Similarly, the following relationships can also be obtained:

$$\gamma \log \frac{\partial r_{i,j+1}}{\partial r_{i,j}} = r_{i,j+1} - r_{i,j} - \delta_{i,j+1},$$

$$\gamma \log \frac{\partial r_{i+1,j+1}}{\partial r_{i,j}} = r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1}.$$

We have therefore obtained a *backward* recursion to compute the entire matrix $E = [e_{i,j}]$, starting from $e_{n,m} = \frac{\partial r_{n,m}}{\partial r_{n,m}} = 1$ down to $e_{1,1}$. To obtain $\nabla_{\mathbf{x}}\,\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$, notice that the derivatives w.r.t. the entries of the cost matrix Δ can be computed by $\frac{\partial r_{n,m}}{\partial \delta_{i,j}} = \frac{\partial r_{n,m}}{\partial r_{i,j}} \frac{\partial r_{i,j}}{\partial \delta_{i,j}} = e_{i,j} \cdot 1 = e_{i,j}$, and therefore we have that

$$\nabla_{\mathbf{x}}\,\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}\right)^T E,$$

where E is exactly the average alignment $\mathbb{E}_\gamma[A]$ in Eq. (3). These computations are summarized in Algorithm 2, which, once Δ has been computed, has complexity nm in time and space. Because $\min^\gamma$ has a $1/\gamma$-Lipschitz continuous gradient, the gradient of $\mathbf{dtw}_\gamma$ is $2/\gamma$-Lipschitz continuous when δ is the squared Euclidean distance.

Algorithm 2 Backward recursion to compute $\nabla_{\mathbf{x}}\,\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$

Inputs: $\mathbf{x},\mathbf{y}$, smoothing $\gamma \geq 0$, distance function δ
$(\cdot, R) = \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$, $\Delta = [\delta(x_i, y_j)]_{i,j}$
$\delta_{i,m+1} = \delta_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$e_{i,m+1} = e_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$r_{i,m+1} = r_{n+1,j} = -\infty, i \in [\![n]\!], j \in [\![m]\!]$
$\delta_{n+1,m+1} = 0, e_{n+1,m+1} = 1, r_{n+1,m+1} = r_{n,m}$
for $j = m, \ldots, 1$ **do**
  for $i = n, \ldots, 1$ **do**
    $a = \exp \frac{1}{\gamma}(r_{i+1,j} - r_{i,j} - \delta_{i+1,j})$
    $b = \exp \frac{1}{\gamma}(r_{i,j+1} - r_{i,j} - \delta_{i,j+1})$
    $c = \exp \frac{1}{\gamma}(r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1})$
    $e_{i,j} = e_{i+1,j} \cdot a + e_{i,j+1} \cdot b + e_{i+1,j+1} \cdot c$
  end for
end for
Output: $\nabla_{\mathbf{x}}\,\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}\right)^T E$

3. Learning with the soft-DTW loss

3.1. Averaging with the soft-DTW geometry

We study in this section a direct application of Algorithm 2 to the problem of computing Fréchet means (1948) of time series with respect to the $\mathbf{dtw}_\gamma$ discrepancy. Given a family of N times series $\mathbf{y}_1, \ldots, \mathbf{y}_N$, namely N matrices of p lines and varying number of columns, $m_1, \ldots, m_N$, we are interested in defining a single barycenter time series $\mathbf{x}$ for that family under a set of normalized weights $\lambda_1, \ldots, \lambda_N \in \mathbb{R}_+$ such that $\sum_{i=1}^N \lambda_i = 1$. Our goal is thus to solve approximately the following problem, in which we have assumed that $\mathbf{x}$ has fixed length n:

$$\min_{\mathbf{x} \in \mathbb{R}^{p \times n}} \sum_{i=1}^N \frac{\lambda_i}{m_i}\,\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}_i). \tag{4}$$

Note that each $\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}_i)$ term is divided by m_i, the length of $\mathbf{y}_i$. Indeed, since $\mathbf{dtw}_0$ is an increasing (roughly linearly) function of each of the input lengths n and m_i, we follow the convention of normalizing in practice each discrepancy by $n \times m_i$. Since the length n of $\mathbf{x}$ is here fixed across all evaluations, we do not need to divide the objective of Eq. (4) by n. Averaging under the soft-DTW geometry results in substantially different results than those that can be obtained with the Euclidean geometry (which can only be used in the case where all lengths $n = m_1 = \cdots =$

IPR2026-00272
Krisp Technologies EX1010 Page 299

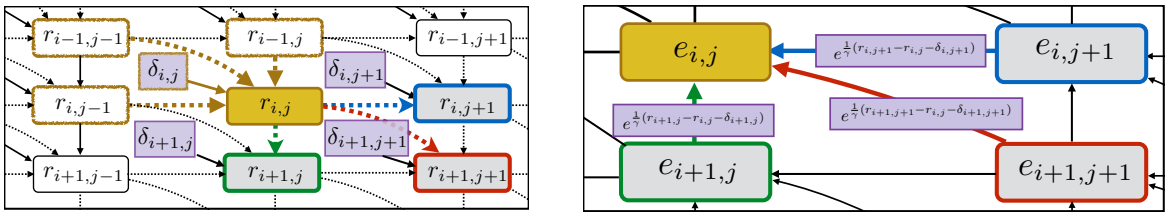
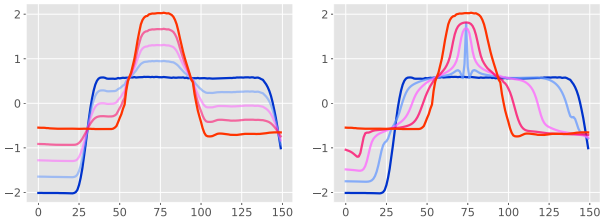


Figure 3. Sketch of the computational graph for soft-DTW, in the forward pass used to compute $\mathbf{dtw}_\gamma$ (left) and backward pass used to compute its gradient $\nabla_{\mathbf{x}} \mathbf{dtw}_\gamma$ (right). In both diagrams, purple shaded cells stand for data values available before the recursion starts, namely cost values (left) and multipliers computed using forward pass results (right). In the left diagram, the forward computation of $r_{i,j}$ as a function of its predecessors and $\delta_{i,j}$ is summarized with arrows. Dotted lines indicate a $\min^\gamma$ operation, solid lines an addition. From the perspective of the final term $r_{n,m}$, which stores $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$ at the lower right corner (not shown) of the computational graph, a change in $r_{i,j}$ only impacts $r_{n,m}$ through changes that $r_{i,j}$ causes to $r_{i+1,j}$, $r_{i,j+1}$ and $r_{i+1,j+1}$. These changes can be tracked using Eq. (2.3,2.3) and appear in lines 9-11 in Algorithm 2 as variables a, b, c, as well as in the purple shaded boxes in the backward pass (right) which represents the recursion of line 12 in Algorithm 2.

m_N are equal), as can be seen in the intuitive interpolations we obtain between two time series shown in Figure 4.

Non-convexity of $\mathbf{dtw}_\gamma$. A natural question that arises from Eq. (4) is whether that objective is convex or not. The answer is negative, in a way that echoes the non-convexity of the k-means objective as a function of cluster centroids locations. Indeed, for any alignment matrix A of suitable size, each map $\mathbf{x} \mapsto \langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle$ shares the same convexity/concavity property that δ may have. However, both $\min$ and $\min^\gamma$ can only preserve the *concavity* of elementary functions (Boyd & Vandenberghe, 2004, pp.72-74). Therefore $\mathbf{dtw}_\gamma$ will only be concave if δ is concave, or become instead a (non-convex) (soft) minimum of convex functions if δ is convex. When δ is a squared-Euclidean distance, $\mathbf{dtw}_0$ is a piecewise quadratic function of $\mathbf{x}$, as is also the case with the k-means energy (see for instance Figure 2 in Schultz & Jain 2017). Since this is the setting we consider here, all of the computations involving barycenters should be taken with a grain of salt, since we have no way of ensuring optimality when approximating Eq. (4).

Smoothing helps optimizing $\mathbf{dtw}_\gamma$. Smoothing can be regarded, however, as a way to "convexify" $\mathbf{dtw}_\gamma$. Indeed, notice that $\mathbf{dtw}_\gamma$ converges to the sum of all costs as $\gamma \to \infty$. Therefore, if δ is convex, $\mathbf{dtw}_\gamma$ will gradually become convex as γ grows. For smaller values of γ, one can intuitively foresee that using $\min^\gamma$ instead of a minimum will smooth out local minima and therefore provide a better (although slightly different from $\mathbf{dtw}_0$) optimization landscape. We believe this is why our approach recovers better results, **even when measured in the original $\mathbf{dtw}_0$ discrepancy**, than subgradient or alternating minimization approaches such as DBA (Petitjean et al., 2011), which can, on the contrary, get more easily stuck in local minima. Evidence for this statement is presented in the experimental section.



(a) Euclidean loss (b) Soft-DTW loss ($\gamma = 1$)

Figure 4. Interpolation between two time series (red and blue) on the Gun Point dataset. We computed the barycenter by solving Eq. (4) with (λ_1, λ_2) set to (0.25, 0.75), (0.5, 0.5) and (0.75, 0.25). The soft-DTW geometry leads to visibly different interpolations.

3.2. Clustering with the soft-DTW geometry

The (approximate) computation of $\mathbf{dtw}_\gamma$ barycenters can be seen as a first step towards the task of clustering time series under the $\mathbf{dtw}_\gamma$ discrepancy. Indeed, one can naturally formulate that problem as that of finding centroids $\mathbf{x}_1, \ldots, \mathbf{x}_k$ that minimize the following energy:

$$\min_{\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^{p \times n}} \sum_{i=1}^{N} \frac{1}{m_i} \min_{j \in [[k]]} \mathbf{dtw}_\gamma(\mathbf{x}_j, \mathbf{y}_i). \qquad (5)$$

To solve that problem one can resort to a direct generalization of Lloyd's algorithm (1982) in which each centering step and each clustering allocation step is done according to the $\mathbf{dtw}_\gamma$ discrepancy.

3.3. Learning prototypes for time series classification

One of the de-facto baselines for learning to classify time series is the k nearest neighbors (k-NN) algorithm, combined with DTW as discrepancy measure between time series. However, k-NN has two main drawbacks. First, the time series used for training must be stored, leading to potentially high storage cost. Second, in order to com-

IPR2026-00272
Krisp Technologies EX1010 Page 300

pute predictions on new time series, the DTW discrepancy must be computed with all training time series, leading to high computational cost. Both of these drawbacks can be addressed by the nearest centroid classifier (Hastie et al., 2001, p.670), (Tibshirani et al., 2002). This method chooses the class whose barycenter (centroid) is closest to the time series to classify. Although very simple, this method was shown to be competitive with k-NN, while requiring much lower computational cost at prediction time (Petitjean et al., 2014). Soft-DTW can naturally be used in a nearest centroid classifier, in order to compute the barycenter of each class at train time, and to compute the discrepancy between barycenters and time series, at prediction time.

3.4. Multistep-ahead prediction

Soft-DTW is ideally suited as a loss function for any task that requires time series outputs. As an example of such a task, we consider the problem of, given the first $1, \dots, t$ observations of a time series, predicting the remaining $(t+1), \dots, n$ observations. Let $\mathbf{x}^{t,t'} \in \mathbb{R}^{p \times (t'-t+1)}$ be the submatrix of $\mathbf{x} \in \mathbb{R}^{p \times n}$ of all columns with indices between t and t', where $1 \leq t < t' < n$. Learning to predict the segment of a time series can be cast as the problem

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \mathbf{dtw}_\gamma \left(f_\theta(\mathbf{x}_i^{1,t}), \mathbf{x}_i^{t+1,n} \right),$$

where $\{f_\theta\}$ is a set of parameterized function that take as input a time series and outputs a time series. Natural choices would be multi-layer perceptrons or recurrent neural networks (RNN), which have been historically trained with a Euclidean loss (Parlos et al., 2000, Eq.5).

4. Experimental results

Throughout this section, we use the UCR (University of California, Riverside) time series classification archive (Chen et al., 2015). We use a subset containing 79 datasets encompassing a wide variety of fields (astronomy, geology, medical imaging) and lengths. Datasets include class information (up to 60 classes) for each time series and are split into train and test sets. Due to the large number of datasets in the UCR archive, we choose to report only a summary of our results in the main manuscript. Detailed results are included in the appendices for interested readers.

4.1. Averaging experiments

In this section, we compare the soft-DTW barycenter approach presented in §3.1 to DBA (Petitjean et al., 2011) and a simple batch subgradient method.

Experimental setup. For each dataset, we choose a class at random, pick 10 time series in that class and compute their barycenter. For quantitative results below, we repeat this procedure 10 times and report the averaged results. For each method, we set the maximum number of iterations to 100. To minimize the proposed soft-DTW barycenter objective, Eq. (4), we use L-BFGS.

Table 1. Percentage of the datasets on which the proposed soft-DTW barycenter is achieving lower DTW loss (Equation (4) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	40.51%	3.80%
$\gamma = 0.1$	93.67%	46.83%
$\gamma = 0.01$	100%	79.75%
$\gamma = 0.001$	97.47%	89.87%
Comparison with subgradient method		
$\gamma = 1$	96.20%	35.44%
$\gamma = 0.1$	97.47%	72.15%
$\gamma = 0.01$	97.47%	92.41%
$\gamma = 0.001$	97.47%	97.47%

Qualitative results. We first visualize the barycenters obtained by soft-DTW when $\gamma = 1$ and $\gamma = 0.01$, by DBA and by the subgradient method. Figure 5 shows barycenters obtained using random initialization on the ECG200 dataset. More results with both random and Euclidean mean initialization are given in Appendix B and C.

We observe that both DBA or soft-DTW with low smoothing parameter γ yield barycenters that are spurious. On the other hand, a descent on the soft-DTW loss with sufficiently high γ converges to a reasonable solution. For example, as indicated in Figure 5 with DTW or soft-DTW ($\gamma = 0.01$), the small kink around $x = 15$ is not representative of any of the time series in the dataset. However, with soft-DTW ($\gamma = 1$), the barycenter closely matches the time series. This suggests that DTW or soft-DTW with too low γ can get stuck in bad local minima.

When using Euclidean mean initialization (only possible if time series have the same length), DTW or soft-DTW with low γ often yield barycenters that better match the shape of the time series. However, they tend to overfit: they absorb the idiosyncrasies of the data. In contrast, soft-DTW is able to learn barycenters that are much smoother.

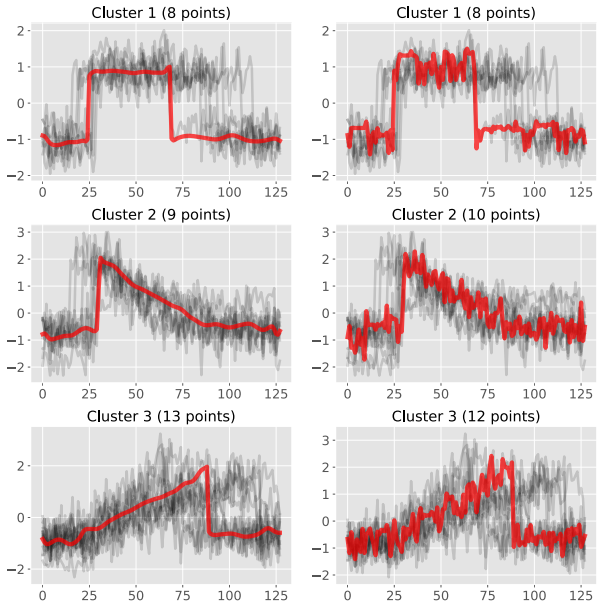
Quantitative results. Table 1 summarizes the percentage of datasets on which the proposed soft-DTW barycenter achieves lower DTW loss when varying the smoothing parameter γ. The actual loss values achieved by different methods are indicated in Appendix G and Appendix H.

As γ decreases, soft-DTW achieves a lower DTW loss than other methods on almost all datasets. This confirms our

IPR2026-00272
Krisp Technologies EX1010 Page 301



Figure 5. Comparison between our proposed soft barycenter and the barycenter obtained by DBA and the subgradient method, on the ECG200 dataset. When DTW is insufficiently smoothed, barycenters often get stuck in a bad local minimum that does not correctly match the time series.



(a) Soft-DTW ($\gamma = 1$) (b) DBA

Figure 6. Clusters obtained on the CBF dataset when plugging our proposed soft barycenter and that of DBA in Lloyd's algorithm. DBA absorbs the idiosyncrasies of the data, while soft-DTW can learn much smoother barycenters.

claim that the smoothness of soft-DTW leads to an objective that is better behaved and more amenable to optimization by gradient-descent methods.

4.2. k-means clustering experiments

We consider in this section the same computational tools used in §4.1 above, but use them to cluster time series.

Experimental setup. For all datasets, the number of clusters k is equal to the number of classes available in the dataset. Lloyd's algorithm alternates between a centering step (barycenter computation) and an assignment step. We set the maximum number of outer iterations to 30 and the maximum number of inner (barycenter) iterations to 100, as before. Again, for soft-DTW, we use L-BFGS.

Qualitative results. Figure 6 shows the clusters obtained when runing Lloyd's algorithm on the CBF dataset with soft-DTW ($\gamma = 1$) and DBA, in the case of random initialization. More results are included in Appendix E. Clearly, DTW absorbs the tiny details in the data, while soft-DTW is able to learn much smoother barycenters.

Quantitative results. Table 2 summarizes the percentage of datasets on which soft-DTW barycenter achieves lower k-means loss under DTW, i.e. Eq. (5) with $\gamma = 0$. The actual loss values achieved by all methods are indicated in Appendix I and Appendix J. The results confirm the same trend as for the barycenter experiments. Namely, as γ decreases, soft-DTW is able to achieve lower loss than other methods on a large proportion of the datasets. Note that we have not run experiments with smaller values of γ than 0.001, since $\mathbf{dtw}_{0.001}$ is very close to $\mathbf{dtw}_0$ in practice.

4.3. Time-series classification experiments

In this section, we investigate whether the smoothing in soft-DTW can act as a useful regularization and improve classification accuracy in the nearest centroid classifier.

Experimental setup. We use 50% of the data for training, 25% for validation and 25% for testing. We choose γ from 15 log-spaced values between 10^{-3} and 10.

Quantitative results. Each point in Figure 7 above the diagonal line represents a dataset for which using soft-DTW for barycenter computation rather than DBA improves the accuracy of the nearest centroid classifier. To summarize, we found that soft-DTW is working better or at least as well as DBA in 75% of the datasets.

4.4. Multistep-ahead prediction experiments

In this section, we present preliminary experiments for the task of multistep-ahead prediction, described in §3.4.

Experimental setup. We use the training and test sets predefined in the UCR archive. In both the training and test sets, we use the first 60% of the time series as input and the remaining 40% as output, ignoring class information. We then use the training set to learn a model that predicts the outputs from inputs and the test set to evaluate results with both Euclidean and DTW losses. In this experiment, we focus on a simple multi-layer perceptron (MLP) with one

IPR2026-00272
Krisp Technologies EX1010 Page 302

Table 2. Percentage of the datasets on which the proposed soft-DTW based k-**means** is achieving lower DTW loss (Equation (5) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	15.78%	29.31%
$\gamma = 0.1$	24.56%	24.13%
$\gamma = 0.01$	59.64%	55.17%
$\gamma = 0.001$	77.19%	68.97%
Comparison with subgradient method		
$\gamma = 1$	42.10%	46.44%
$\gamma = 0.1$	57.89%	50%
$\gamma = 0.01$	76.43%	65.52%
$\gamma = 0.001$	96.49%	84.48%

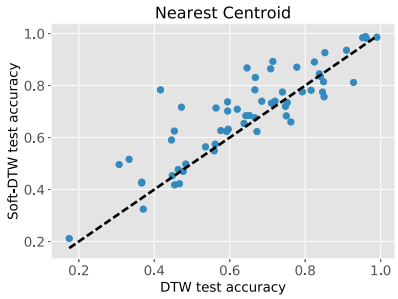


Figure 7. Each point above the diagonal represents a dataset where using our soft-DTW barycenter rather than that of DBA improves the accuracy of the nearest nearest centroid classifier. This is the case for **75%** of the datasets in the UCR archive.

hidden layer and sigmoid activation. We also experimented with linear models and recurrent neural networks (RNNs) but they did not improve over a simple MLP.

Implementation details. Deep learning frameworks such as Theano, TensorFlow and Chainer allow the user to specify a custom backward pass for their function. Implementing such a backward pass, rather than resorting to automatic differentiation (autodiff), is particularly important in the case of soft-DTW: First, the autodiff in these frameworks is designed for vectorized operations, whereas the dynamic program used by the forward pass of Algorithm 2 is inherently element-wise; Second, as we explained in §2.2, our backward pass is able to re-use log-sum-exp computations from the forward pass, leading to both lower computational cost and better numerical stability. We implemented a custom backward pass in Chainer, which can then be used to plug soft-DTW as a loss function in any network architecture. To estimate the MLP's parameters, we used Chainer's implementation of Adam (Kingma & Ba, 2014).

Qualitative results. Visualizations of the predictions obtained under Euclidean and soft-DTW losses are given in Figure 1, as well as in Appendix F. We find that for simple one-dimensional time series, an MLP works very well, showing its ability to capture patterns in the training set. Although the predictions under Euclidean and soft-DTW losses often agree with each other, they can sometimes be visibly different. Predictions under soft-DTW loss can confidently predict abrupt and sharp changes since those have a low DTW cost as long as such a sharp change is present, under a small time shift, in the ground truth.

Table 3. Averaged rank obtained by a multi-layer perceptron (MLP) under Euclidean and soft-DTW losses. Euclidean initialization means that we initialize the MLP trained with soft-DTW loss by the solution of the MLP trained with Euclidean loss.

Training loss	Random initialization	Euclidean initialization
When evaluating with DTW loss		
Euclidean	3.46	4.21
soft-DTW ($\gamma = 1$)	3.55	3.96
soft-DTW ($\gamma = 0.1$)	3.33	3.42
soft-DTW ($\gamma = 0.01$)	2.79	2.12
soft-DTW ($\gamma = 0.001$)	**1.87**	**1.29**
When evaluating with Euclidean loss		
Euclidean	**1.05**	**1.70**
soft-DTW ($\gamma = 1$)	2.41	2.99
soft-DTW ($\gamma = 0.1$)	3.42	3.38
soft-DTW ($\gamma = 0.01$)	4.13	3.64
soft-DTW ($\gamma = 0.001$)	3.99	3.29

Quantitative results. A comparison summary of our MLP under Euclidean and soft-DTW losses over the UCR archive is given in Table 3. Detailed results are given in the appendix. Unsurprisingly, we achieve lower DTW loss when training with the soft-DTW loss, and lower Euclidean loss when training with the Euclidean loss. Because DTW is robust to several useful invariances, a small error in the soft-DTW sense could be a more judicious choice than an error in an Euclidean sense for many applications.

5. Conclusion

We propose in this paper to turn the popular DTW discrepancy between time series into a full-fledged loss function between ground truth time series and outputs from a learning machine. We have shown experimentally that, on the existing problem of computing barycenters and clusters for time series data, our computational approach is superior to existing baselines. We have shown promising results on the problem of multistep-ahead time series prediction, which could prove extremely useful in settings where a user's actual loss function for time series is closer to the robust perspective given by DTW, than to the local parsing of the Euclidean distance.

Acknowledgements. MC gratefully acknowledges the support of a *chaire de l'IDEX Paris Saclay*.

IPR2026-00272
Krisp Technologies EX1010 Page 303

References

Bahl, L and Jelinek, Frederick. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, 21(4):404–411, 1975.

Bakir, GH, Hofmann, T, Schölkopf, B, Smola, AJ, Taskar, B, and Vishwanathan, SVN. *Predicting Structured Data*. Advances in neural information processing systems. MIT Press, Cambridge, MA, USA, 2007.

Bellman, Richard. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8): 716–719, 1952.

Blondel, Mathieu, Fujino, Akinori, Ueda, Naonori, and Ishihata, Masakazu. Higher-order factorization machines. In *Advances in Neural Information Processing Systems 29*, pp. 3351–3359. 2016.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Chen, Yanping, Keogh, Eamonn, Hu, Bing, Begum, Nurjahan, Bagnall, Anthony, Mueen, Abdullah, and Batista, Gustavo. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

Cuturi, Marco. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 929–936, 2011.

Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 685–693, 2014.

Cuturi, Marco, Vert, Jean-Philippe, Birkenes, Oystein, and Matsui, Tomoko. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pp. II–413, 2007.

Fréchet, Maurice. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pp. 215–310. Presses universitaires de France, 1948.

Garreau, Damien, Lajugie, Rémi, Arlot, Sylvain, and Bach, Francis. Metric learning for temporal sequence alignment. In *Advances in Neural Information Processing Systems*, pp. 1817–1825, 2014.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lasserre, Jean B. *Linear and integer programming vs linear integration and counting: a duality viewpoint*. Springer Science & Business Media, 2009.

Lloyd, Stuart. Least squares quantization in pcm. *IEEE Trans. on Information Theory*, 28(2):129–137, 1982.

Lütkepohl, Helmut. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Parlos, Alexander G, Rais, Omar T, and Atiya, Amir F. Multi-step-ahead prediction using dynamic recurrent neural networks. *Neural networks*, 13(7):765–786, 2000.

Petitjean, François and Gançarski, Pierre. Summarizing a set of time series by averaging: From steiner sequence to compact multiple alignment. *Theoretical Computer Science*, 414(1):76–91, 2012.

Petitjean, François, Ketterlin, Alain, and Gançarski, Pierre. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44 (3):678–693, 2011.

Petitjean, François, Forestier, Germain, Webb, Geoffrey I, Nicholson, Ann E, Chen, Yanping, and Keogh, Eamonn. Dynamic time warping averaging of time series allows faster and more accurate classification. In *ICDM*, pp. 470–479. IEEE, 2014.

Ristad, Eric Sven and Yianilos, Peter N. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed Wasserstein loss. *Proceedings of AISTATS'16*, 2016.

Saigo, Hiroto, Vert, Jean-Philippe, Ueda, Nobuhisa, and Akutsu, Tatsuya. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

Saigo, Hiroto, Vert, Jean-Philippe, and Akutsu, Tatsuya. Optimizing amino acid substitution matrices with a local alignment kernel. *BMC bioinformatics*, 7(1):246, 2006.

Sakoe, Hiroaki and Chiba, Seibi. A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics, Budapest*, volume 3, pp. 65–69, 1971.

Sakoe, Hiroaki and Chiba, Seibi. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, 26:43–49, 1978.

Schultz, David and Jain, Brijnesh. Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. *arXiv preprint arXiv:1701.06393*, 2017.

Tibshirani, Robert, Hastie, Trevor, Narasimhan, Balasubramanian, and Chu, Gilbert. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10): 6567–6572, 2002.

Yi, Byoung-Kee, Jagadish, HV, and Faloutsos, Christos. Efficient retrieval of similar time sequences under time warping. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pp. 201–208. IEEE, 1998.

Zhang, C., Frogner, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems 29*, 2015.

Appendix CUTURI03

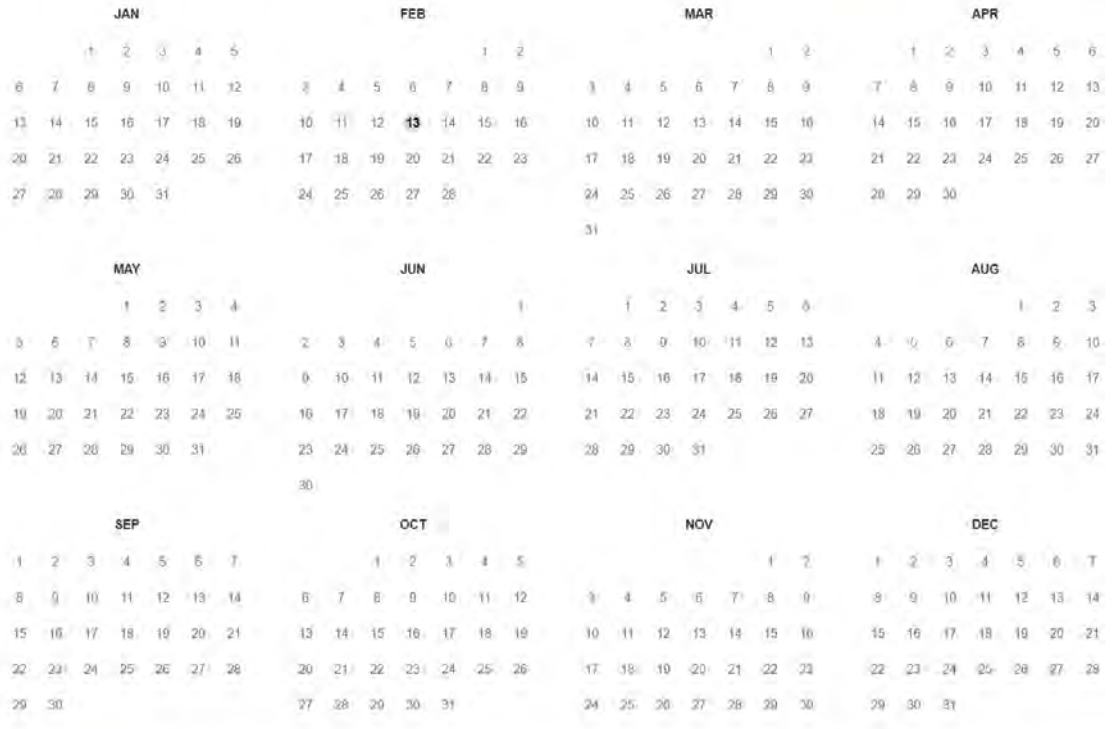
INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE INTERNET ARCHIVE WayBackMachine Explore more than 1 trillion web pages saved over time
https://arxiv.org/abs/1703.01541

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **30 times** between February 13, 2019 and September 14, 2025.





Note

This calendar view maps the number of times **https://arxiv.org/abs/1703.01541** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20190213125708/https://arxiv.org/abs/1703.01541 December 19, 2025

INTERNET ARCHIVE Wayback Machine https://arxiv.org/abs/1703.01541 Go JAN FEB 13 JUL 2018 2019 2020 About this capture
30 captures
13 Feb 2019 - 14 Sep 2025

Statistics > Machine Learning

Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi, Mathieu Blondel

(Submitted on 5 Mar 2017 (v1), last revised 20 Feb 2018 (this version, v2))

We propose in this paper a differentiable learning loss between time series, building upon the celebrated dynamic time warping (DTW) discrepancy. Unlike the Euclidean distance, DTW can compare time series of variable size and is robust to shifts or dilatations across the time dimension. To compute DTW, one typically solves a minimal-cost alignment problem between two time series using dynamic programming. Our work takes advantage of a smoothed formulation of DTW, called soft-DTW, that computes the soft-minimum of all alignment costs. We show in this paper that soft-DTW is a differentiable loss function, and that both its value and gradient can be computed with quadratic time/space complexity (DTW has quadratic time but linear space complexity). We show that this regularization is particularly well suited to average and cluster time series under the DTW geometry, a task for which our proposal significantly outperforms existing baselines. Next, we propose to tune the parameters of a machine that outputs time series by minimizing its fit with ground-truth labels in a soft-DTW sense.

Comments: Published in ICML 2017
Subjects: **Machine Learning (stat.ML)**
Cite as: **arXiv:1703.01541 [stat.ML]**
(or **arXiv:1703.01541v2 [stat.ML]** for this version)

Submission history

From: Mathieu Blondel [view email]
[v1] Sun, 5 Mar 2017 01:30:28 UTC (4,490 KB)
[v2] Tue, 20 Feb 2018 11:10:29 UTC (3,372 KB)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Download:
- PDF
- Other formats

(license)

Current browse context:
stat.ML
< prev | next >
new | recent | 1703

Change to browse by:
stat

References & Citations
- NASA ADS

Bookmark (what is this?)

Link back to: arXiv, form interface, contact.

Browse v0.1 released 2018-10-22 Feedback?

If you have a disability and are having trouble accessing information on this website or need materials in an alternate format, contact web-accessibility@cornell.edu for assistance.

Appendix CUTURI04

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

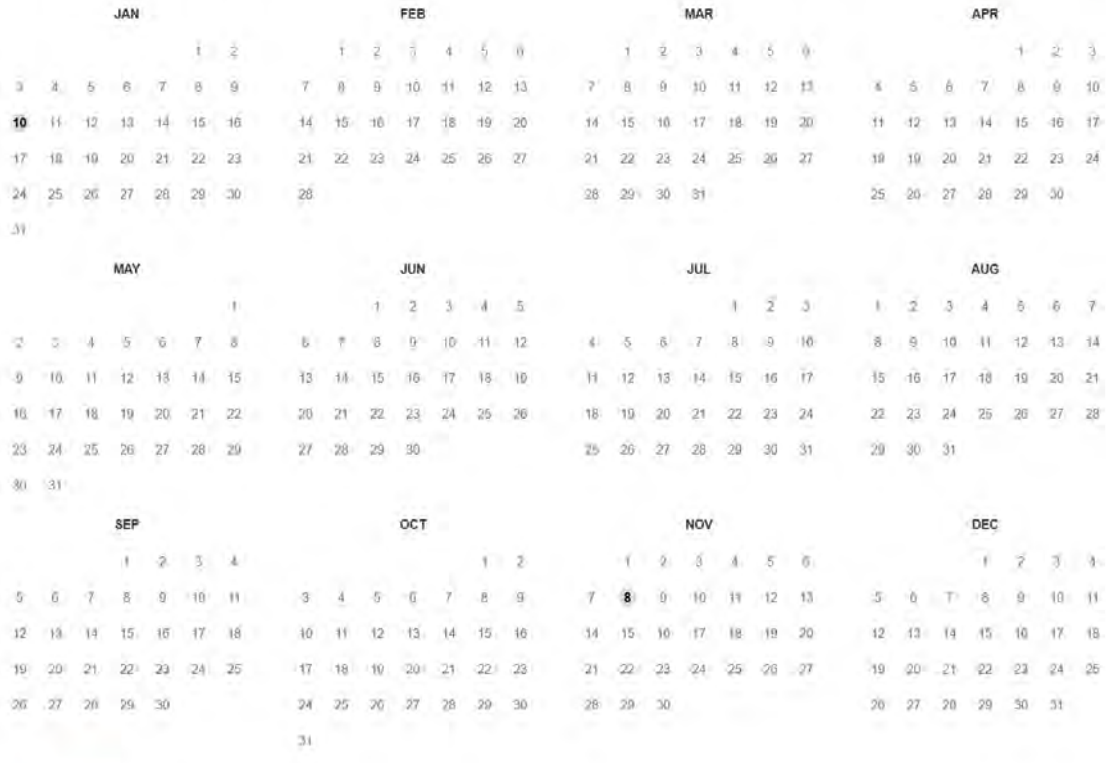
DONATE INTERNET ARCHIVE WayBackMachine Explore more than 1 trillion web pages saved over time

https://arxiv.org/pdf/1703.01541.pdf

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **9 times** between January 10, 2021 and September 18, 2025.





Note

This calendar view maps the number of times **https://arxiv.org/pdf/1703.01541.pdf** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20211108222341/https://arxiv.org/pdf/1703.01541.pdf December 19, 2025

INTERNET ARCHIVE WaybackMachine https://arxiv.org/pdf/1703.01541.pdf Go DCT NOV 08 2021 DEC
9 captures
10 Jan 2021 - 18 Sep 2025
2020 2022 About this capture



1



2



ML] 20 Feb 2018

Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi [1] Mathieu Blondel [2]

Abstract

We propose in this paper a differentiable learning loss between time series, building upon the celebrated dynamic time warping (DTW) discrepancy. Unlike the Euclidean distance, DTW can compare time series of variable size and is robust to shifts or dilatations across the time dimension. To compute DTW, one typically solves a minimal-cost alignment problem between two time series using dynamic programming. Our work takes advantage of a smoothed formulation of DTW, called soft-DTW, that computes the soft-minimum of all alignment costs. We show in this paper that soft-DTW is a *differentiable*

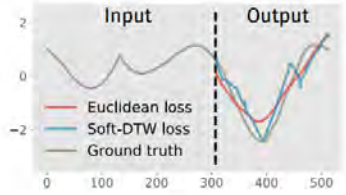


Figure 1. Given the first part of a time series, we trained two multi-layer perceptron (MLP) to predict the entire second part. Using the ShapesAll dataset, we used a Euclidean loss for the first MLP and the soft-DTW loss proposed in this paper for the second

arXiv:1703.01541v2 [stat.ML] 20 Feb 2018

Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi [1] **Mathieu Blondel** [2]

Abstract

We propose in this paper a differentiable learning loss between time series, building upon the celebrated dynamic time warping (DTW) discrepancy. Unlike the Euclidean distance, DTW can compare time series of variable size and is robust to shifts or dilatations across the time dimension. To compute DTW, one typically solves a minimal-cost alignment problem between two time series using dynamic programming. Our work takes advantage of a smoothed formulation of DTW, called soft-DTW, that computes the soft-minimum of all alignment costs. We show in this paper that soft-DTW is a *differentiable* loss function, and that both its value and gradient can be computed with quadratic time/space complexity (DTW has quadratic time but linear space complexity). We show that this regularization is particularly well suited to average and cluster time series under the DTW geometry, a task for which our proposal significantly outperforms existing baselines (Petitjean et al., 2011). Next, we propose to tune the parameters of a machine that outputs time series by minimizing its fit with ground-truth labels in a soft-DTW sense.

1. Introduction

The goal of supervised learning is to learn a mapping that links an input to an output objects, using examples of such pairs. This task is noticeably more difficult when the output objects have a structure, *i.e.* when they are not vectors (Bakir et al., 2007). We study here the case where each output object is a *time series*, namely a family of observations indexed by time. While it is tempting to treat time as yet another feature, and handle time series of vectors as the concatenation of all these vectors, several practical issues arise when taking this simplistic approach: Time-indexed phenomena can often be stretched in some areas along the time axis (a word uttered in a slightly slower pace than usual) with no impact on their characteristics; varying sampling conditions may mean they have different lengths; time series may not synchronized.

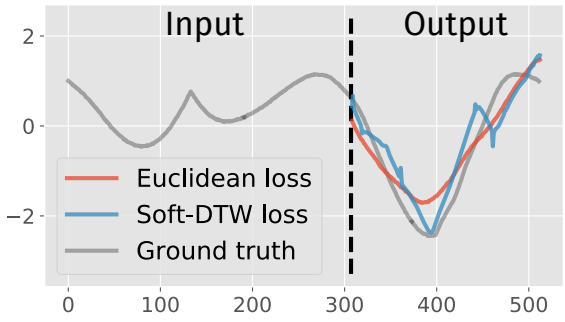


Figure 1. Given the first part of a time series, we trained two multi-layer perceptron (MLP) to predict the entire second part. Using the ShapesAll dataset, we used a Euclidean loss for the first MLP and the soft-DTW loss proposed in this paper for the second one. We display above the prediction obtained for a given test instance with either of these two MLPs in addition to the ground truth. Oftentimes, we observe that the soft-DTW loss enables us to better predict sharp changes. More time series predictions are given in Appendix F.

The DTW paradigm. Generative models for time series are usually built having the invariances above in mind: Such properties are typically handled through latent variables and/or Markovian assumptions (Lütkepohl, 2005, Part I,§18). A simpler approach, motivated by geometry, lies in the direct definition of a discrepancy between time series that encodes these invariances, such as the Dynamic Time Warping (DTW) score (Sakoe & Chiba, 1971; 1978). DTW computes the best possible alignment between two time series (the optimal alignment itself can also be of interest, see e.g. Garreau et al. 2014) of respective length n and m by computing first the $n \times m$ pairwise distance matrix between these points to solve then a dynamic program (DP) using Bellman's recursion with a quadratic (nm) cost.

The DTW geometry. Because it encodes efficiently a useful class of invariances, DTW has often been used in a discriminative framework (with a k-NN or SVM classifier) to predict a real or a class label output, and engineered to run

[1]CREST, ENSAE, Université Paris-Saclay, France [2]NTT Communication Science Laboratories, Seika-cho, Kyoto, Japan. Correspondence to: Marco Cuturi <marco.cuturi@ensae.fr>, Mathieu Blondel <mathieu@mblondel.org>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

faster in that context (Yi et al., 1998). Recent works by Petitjean et al. (2011); Petitjean & Gançarski (2012) have, however, shown that DTW can be used for more innovative tasks, such as time series *averaging* using the DTW discrepancy (see Schultz & Jain 2017 for a gentle introduction to these ideas). More generally, the idea of synthetising time series centroids can be regarded as a first attempt to *output* entire time series using DTW as a fitting loss. From a computational perspective, these approaches are, however, hampered by the fact that DTW is not differentiable and unstable when used in an optimization pipeline.

Soft-DTW. In parallel to these developments, several authors have considered smoothed modifications of Bellman's recursion to define smoothed DP distances (Bahl & Jelinek, 1975; Ristad & Yianilos, 1998) or kernels (Saigo et al., 2004; Cuturi et al., 2007). When applied to the DTW discrepancy, that regularization results in a *soft-DTW* score, which considers the *soft-minimum* of the distribution of *all costs* spanned by *all* possible alignments between two time series. Despite considering all alignments and not just the optimal one, soft-DTW can be computed with a minor modification of Bellman's recursion, in which all $(\min, +)$ operations are replaced with $(+, \times)$. As a result, both DTW and soft-DTW have quadratic in time & linear in space complexity with respect to the sequences' lengths. Because soft-DTW can be used with kernel machines, one typically observes an increase in performance when using soft-DTW over DTW (Cuturi, 2011) for classification.

Our contributions. We explore in this paper another important benefit of smoothing DTW: unlike the original DTW discrepancy, soft-DTW is *differentiable* in all of its arguments. We show that the gradients of soft-DTW w.r.t to all of its variables can be computed as a by-product of the computation of the discrepancy itself, with an added quadratic storage cost. We use this fact to propose an alternative approach to the DBA (DTW Barycenter Averaging) clustering algorithm of (Petitjean et al., 2011), and observe that our smoothed approach significantly outperforms known baselines for that task. More generally, we propose to use soft-DTW as a *fitting term* to compare the output of a machine synthesizing a time series segment with a ground truth observation, in the same way that, for instance, a regularized Wasserstein distance was used to compute barycenters (Cuturi & Doucet, 2014), and later to fit discriminators that output histograms (Zhang et al., 2015; Rolet et al., 2016). When paired with a flexible learning architecture such as a neural network, soft-DTW allows for a differentiable end-to-end approach to design predictive and generative models for time series, as illustrated in Figure 1. Source code is available at `https://github.com/mblondel/soft-dtw`.

Structure. After providing background material, we show in §2 how soft-DTW can be differentiated w.r.t the locations of two time series. We follow in §3 by illustrating how these results can be directly used for tasks that require to output time series: averaging, clustering and prediction of time series. We close this paper with experimental results in §4 that showcase each of these potential applications.

Notations. We consider in what follows multivariate discrete time series of varying length taking values in $\Omega \subset \mathbb{R}^p$. A time series can be thus represented as a matrix of p lines and varying number of columns. We consider a differentiable substitution-cost function $\delta : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ which will be, in most cases, the quadratic Euclidean distance between two vectors. For an integer n we write $[\![n]\!]$ for the set $\{1, \dots, n\}$ of integers. Given two series' lengths n and m, we write $\mathcal{A}_{n,m} \subset \{0,1\}^{n \times m}$ for the set of (binary) alignment matrices, that is paths on a $n \times m$ matrix that connect the upper-left $(1,1)$ matrix entry to the lower-right (n,m) one using only $\downarrow, \rightarrow, \searrow$ moves. The cardinal of $\mathcal{A}_{n,m}$ is known as the delannoy$(n-1, m-1)$ number; that number grows exponentially with m and n.

2. The DTW and soft-DTW loss functions

We propose in this section a unified formulation for the original DTW discrepancy (Sakoe & Chiba, 1978) and the Global Alignment kernel (GAK) (Cuturi et al., 2007), which can be both used to compare two time series $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^{p \times n}$ and $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^{p \times m}$.

2.1. Alignment costs: optimality and sum

Given the cost matrix $\Delta(\mathbf{x}, \mathbf{y}) \coloneqq \left[\delta(x_i, y_j)\right]_{ij} \in \mathbb{R}^{n \times m}$, the inner product $\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle$ of that matrix with an alignment matrix A in $\mathcal{A}_{n,m}$ gives the score of A, as illustrated in Figure 2. Both DTW and GAK consider the costs of all possible alignment matrices, yet do so differently:

$$\begin{aligned} \mathrm{DTW}(\mathbf{x}, \mathbf{y}) &\coloneqq \min_{A \in \mathcal{A}_{n,m}} \langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle, \\ k_{\mathrm{GA}}^{\gamma}(\mathbf{x}, \mathbf{y}) &\coloneqq \sum_{A \in \mathcal{A}_{n,m}} e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma}. \end{aligned} \tag{1}$$

DP Recursion. Sakoe & Chiba (1978) showed that the Bellman equation (1952) can be used to compute DTW. That recursion, which appears in line 5 of Algorithm 1 (disregarding for now the exponent γ), only involves $(\min, +)$ operations. When considering kernel k_{GA}^{γ} and, instead, its integration over all alignments (see e.g. Lasserre 2009), Cuturi et al. (2007, Theorem 2) and the highly related formulation of Saigo et al. (2004, p.1685) use an old algorithmic appraoch (Bahl & Jelinek, 1975) which consists in *(i)* replacing all costs by their neg-exponential; *(ii)* replace $(\min, +)$ operations with $(+, \times)$ operations. These two recursions can be in fact unified with the use of a soft-

IPR2026-00272
Krisp Technologies EX1010 Page 313

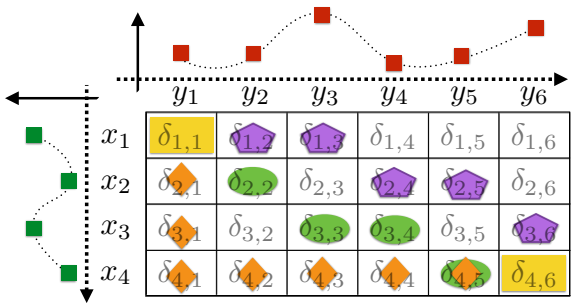


Figure 2. Three alignment matrices (orange, green, purple, in addition to the top-left and bottom-right entries) between two time series of length 4 and 6. The cost of an alignment is equal to the sum of entries visited along the path. DTW only considers the optimal alignment (here depicted in purple pentagons), whereas soft-DTW considers all delannoy$(n-1, m-1)$ possible alignment matrices.

minimum operator, which we present below.

Unified algorithm Both formulas in Eq. (1) can be computed with a single algorithm. That formulation is new to our knowledge. Consider the following generalized $\min$ operator, with a smoothing parameter $\gamma \geq 0$:

$$\min{}^{\gamma}\{a_1, \dots, a_n\} \coloneqq \begin{cases} \min_{i \leq n} a_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma}, & \gamma > 0. \end{cases}$$

With that operator, we can define γ-soft-DTW:

$$\boxed{\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y}) \coloneqq \min{}^{\gamma}\{\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle, A \in \mathcal{A}_{n,m}\}.}$$

The original DTW score is recovered by setting γ to 0. When $\gamma > 0$, we recover $\mathbf{dtw}_\gamma = -\gamma \log k_{\text{GA}}^{\gamma}$. Most importantly, and in either case, $\mathbf{dtw}_\gamma$ can be computed using Algorithm 1, which requires (nm) operations and (nm) storage cost as well . That cost can be reduced to $2n$ with a more careful implementation if one only seeks to compute $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$, but the backward pass we consider next requires the entire matrix R of intermediary alignment costs. Note that, to ensure numerical stability, the operator $\min^{\gamma}$ must be computed using the usual log-sum-exp stabilization trick, namely that $\log \sum_i e^{z_i} = (\max_j z_j) + \log \sum_i e^{z_i - \max_j z_j}$.

2.2. Differentiation of soft-DTW

A small variation in the input $\mathbf{x}$ causes a small change in $\mathbf{dtw}_0(\mathbf{x}, \mathbf{y})$ or $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$. When considering $\mathbf{dtw}_0$, that change can be efficiently monitored only when the optimal alignment matrix $A^\star$ that arises when computing $\mathbf{dtw}_0(\mathbf{x}, \mathbf{y})$ in Eq. (1) is unique. As the minimum over a finite set of linear functions of Δ, $\mathbf{dtw}_0$ is therefore locally differentiable w.r.t. the cost matrix Δ, with gradient $A^\star$, a fact that has been exploited in all algorithms designed to

Algorithm 1 Forward recursion to compute $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$ and intermediate alignment costs

Inputs: $\mathbf{x}, \mathbf{y}$, smoothing $\gamma \geq 0$, distance function δ
$r_{0,0} = 0; r_{i,0} = r_{0,j} = \infty; i \in [\![n]\!], j \in [\![m]\!]$
for $j = 1, \dots, m$ **do**
  for $i = 1, \dots, n$ **do**
    $r_{i,j} = \delta(x_i, y_j) + \min^{\gamma}\{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\}$
  end for
end for
Output: $(r_{n,m}, R)$

average time series under the DTW metric (Petitjean et al., 2011; Schultz & Jain, 2017). To recover the gradient of $\mathbf{dtw}_0(\mathbf{x}, \mathbf{y})$ w.r.t. $\mathbf{x}$, we only need to apply the chain rule, thanks to the differentiability of the cost function:

$$\nabla_{\mathbf{x}} \mathbf{dtw}_0(\mathbf{x}, \mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^T A^\star, \qquad (2)$$

where $\partial \Delta(\mathbf{x}, \mathbf{y}) / \partial \mathbf{x}$ is the Jacobian of Δ w.r.t. $\mathbf{x}$, a linear map from $\mathbb{R}^{p \times n}$ to $\mathbb{R}^{n \times m}$. When δ is the squared Euclidean distance, the transpose of that Jacobian applied to a matrix $B \in \mathbb{R}^{n \times m}$ is ($\circ$ being the elementwise product):

$$(\partial \Delta(\mathbf{x}, \mathbf{y}) / \partial \mathbf{x})^T B = 2 \left((\mathbf{1}_p \mathbf{1}_m^T B^T) \circ \mathbf{x} - \mathbf{y} B^T \right).$$

With continuous data, $A^\star$ is almost always likely to be unique, and therefore the gradient in Eq. (2) will be defined almost everywhere. However, that gradient, when it exists, will be discontinuous around those values $\mathbf{x}$ where a small change in $\mathbf{x}$ causes a change in $A^\star$, which is likely to hamper the performance of gradient descent methods.

The case $\gamma > 0$. An immediate advantage of soft-DTW is that it can be explicitly differentiated, a fact that was also noticed by Saigo et al. (2006) in the related case of edit distances. When $\gamma > 0$, the gradient of Eq. (1) is obtained via the chain rule,

$$\nabla_{\mathbf{x}} \mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^T \mathbb{E}_\gamma[A], \qquad (3)$$

$$\text{where} \quad \mathbb{E}_\gamma[A] \coloneqq \frac{1}{k_{\text{GA}}^{\gamma}(\mathbf{x}, \mathbf{y})} \sum_{A \in \mathcal{A}_{n,m}} e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma} A,$$

is the average alignment matrix A under the Gibbs distribution $p_\gamma \propto e^{-\langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \gamma}$ defined on all alignments in $\mathcal{A}_{n,m}$. The kernel $k_{\text{GA}}^{\gamma}(\mathbf{x}, \mathbf{y})$ can thus be interpreted as the normalization constant of p_γ. Of course, since $\mathcal{A}_{n,m}$ has exponential size in n and m, a naive summation is not tractable. Although a Bellman recursion to compute that average alignment matrix $\mathbb{E}_\gamma[A]$ exists (see Appendix A) that computation has *quartic* $(n^2 m^2)$ complexity. Note that

IPR2026-00272
Krisp Technologies EX1010 Page 314

this stands in stark contrast to the quadratic complexity obtained by Saigo et al. (2006) for edit-distances, which is due to the fact the sequences they consider can only take values in a *finite* alphabet. To compute the gradient of soft-DTW, we propose instead an algorithm that manages to remain *quadratic* (nm) in terms of complexity. The key to achieve this reduction is to apply the chain rule in *reverse* order of Bellman's recursion given in Algorithm 1, namely back-propagate. A similar idea was recently used to compute the gradient of ANOVA kernels in (Blondel et al., 2016).

2.3. Algorithmic differentiation

Differentiating algorithmically $\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$ requires doing first a forward pass of Bellman's equation to store all intermediary computations and recover $R = [r_{i,j}]$ when running Algorithm 1. The value of $\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$—stored in $r_{n,m}$ at the end of the forward recursion—is then impacted by a change in $r_{i,j}$ exclusively through the terms in which $r_{i,j}$ plays a role, namely the triplet of terms $r_{i+1,j}, r_{i,j+1}, r_{i+1,j+1}$. A straightforward application of the chain rule then gives

$$\underbrace{\frac{\partial r_{n,m}}{\partial r_{i,j}}}_{e_{i,j}} = \underbrace{\frac{\partial r_{n,m}}{\partial r_{i+1,j}}}_{e_{i+1,j}} \frac{\partial r_{i+1,j}}{\partial r_{i,j}} + \underbrace{\frac{\partial r_{n,m}}{\partial r_{i,j+1}}}_{e_{i,j+1}} \frac{\partial r_{i,j+1}}{\partial r_{i,j}} + \underbrace{\frac{\partial r_{n,m}}{\partial r_{i+1,j+1}}}_{e_{i+1,j+1}} \frac{\partial r_{i+1,j+1}}{\partial r_{i,j}},$$

in which we have defined the notation of the main object of interest of the backward recursion: $e_{i,j} := \frac{\partial r_{n,m}}{\partial r_{i,j}}$. The Bellman recursion evaluated at $(i+1,j)$ as shown in line 5 of Algorithm 1 (here $\delta_{i+1,j}$ is $\delta(x_{i+1}, y_j)$) yields :

$$r_{i+1,j} = \delta_{i+1,j} + \min\nolimits^\gamma\{r_{i,j-1}, r_{i,j}, r_{i+1,j-1}\},$$

which, when differentiated w.r.t $r_{i,j}$ yields the ratio:

$$\frac{\partial r_{i+1,j}}{\partial r_{i,j}} = e^{-r_{i,j}/\gamma} \Big/ \left(e^{-r_{i,j-1}/\gamma} + e^{-r_{i,j}/\gamma} + e^{-r_{i+1,j-1}/\gamma}\right).$$

The logarithm of that derivative can be conveniently cast using evaluations of $\min^\gamma$ computed in the forward loop:

$$\begin{aligned}\gamma \log \frac{\partial r_{i+1,j}}{\partial r_{i,j}} &= \min\nolimits^\gamma\{r_{i,j-1}, r_{i,j}, r_{i+1,j-1}\} - r_{i,j} \\ &= r_{i+1,j} - \delta_{i+1,j} - r_{i,j}.\end{aligned}$$

Similarly, the following relationships can also be obtained:

$$\gamma \log \frac{\partial r_{i,j+1}}{\partial r_{i,j}} = r_{i,j+1} - r_{i,j} - \delta_{i,j+1},$$

$$\gamma \log \frac{\partial r_{i+1,j+1}}{\partial r_{i,j}} = r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1}.$$

We have therefore obtained a *backward* recursion to compute the entire matrix $E = [e_{i,j}]$, starting from $e_{n,m} = \frac{\partial r_{n,m}}{\partial r_{n,m}} = 1$ down to $e_{1,1}$. To obtain $\nabla_{\mathbf{x}}\, \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$, notice that the derivatives w.r.t. the entries of the cost matrix Δ can be computed by $\frac{\partial r_{n,m}}{\partial \delta_{i,j}} = \frac{\partial r_{n,m}}{\partial r_{i,j}} \frac{\partial r_{i,j}}{\partial \delta_{i,j}} = e_{i,j} \cdot 1 = e_{i,j}$, and therefore we have that

$$\nabla_{\mathbf{x}}\, \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}\right)^T E,$$

where E is exactly the average alignment $\mathbb{E}_\gamma[A]$ in Eq. (3). These computations are summarized in Algorithm 2, which, once Δ has been computed, has complexity nm in time and space. Because $\min^\gamma$ has a $1/\gamma$-Lipschitz continuous gradient, the gradient of $\mathbf{dtw}_\gamma$ is $2/\gamma$-Lipschitz continuous when δ is the squared Euclidean distance.

Algorithm 2 Backward recursion to compute $\nabla_{\mathbf{x}}\, \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$

Inputs: $\mathbf{x}, \mathbf{y}$, smoothing $\gamma \geq 0$, distance function δ
$(\cdot, R) = \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y})$, $\Delta = [\delta(x_i, y_j)]_{i,j}$
$\delta_{i,m+1} = \delta_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$e_{i,m+1} = e_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$r_{i,m+1} = r_{n+1,j} = -\infty, i \in [\![n]\!], j \in [\![m]\!]$
$\delta_{n+1,m+1} = 0, e_{n+1,m+1} = 1, r_{n+1,m+1} = r_{n,m}$
for $j = m, \ldots, 1$ **do**
  for $i = n, \ldots, 1$ **do**
    $a = \exp \frac{1}{\gamma}(r_{i+1,j} - r_{i,j} - \delta_{i+1,j})$
    $b = \exp \frac{1}{\gamma}(r_{i,j+1} - r_{i,j} - \delta_{i,j+1})$
    $c = \exp \frac{1}{\gamma}(r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1})$
    $e_{i,j} = e_{i+1,j} \cdot a + e_{i,j+1} \cdot b + e_{i+1,j+1} \cdot c$
  end for
end for
Output: $\nabla_{\mathbf{x}}\, \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}}\right)^T E$

3. Learning with the soft-DTW loss

3.1. Averaging with the soft-DTW geometry

We study in this section a direct application of Algorithm 2 to the problem of computing Fréchet means (1948) of time series with respect to the $\mathbf{dtw}_\gamma$ discrepancy. Given a family of N times series $\mathbf{y}_1, \ldots, \mathbf{y}_N$, namely N matrices of p lines and varying number of columns, $m_1, \ldots, m_N$, we are interested in defining a single barycenter time series $\mathbf{x}$ for that family under a set of normalized weights $\lambda_1, \ldots, \lambda_N \in \mathbb{R}_+$ such that $\sum_{i=1}^N \lambda_i = 1$. Our goal is thus to solve approximately the following problem, in which we have assumed that $\mathbf{x}$ has fixed length n:

$$\min_{\mathbf{x} \in \mathbb{R}^{p \times n}} \sum_{i=1}^N \frac{\lambda_i}{m_i}\, \mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}_i). \tag{4}$$

Note that each $\mathbf{dtw}_\gamma(\mathbf{x},\mathbf{y}_i)$ term is divided by m_i, the length of $\mathbf{y}_i$. Indeed, since $\mathbf{dtw}_0$ is an increasing (roughly linearly) function of each of the input lengths n and m_i, we follow the convention of normalizing in practice each discrepancy by $n \times m_i$. Since the length n of $\mathbf{x}$ is here fixed across all evaluations, we do not need to divide the objective of Eq. (4) by n. Averaging under the soft-DTW geometry results in substantially different results than those that can be obtained with the Euclidean geometry (which can only be used in the case where all lengths $n = m_1 = \cdots =$

IPR2026-00272
Krisp Technologies EX1010 Page 315

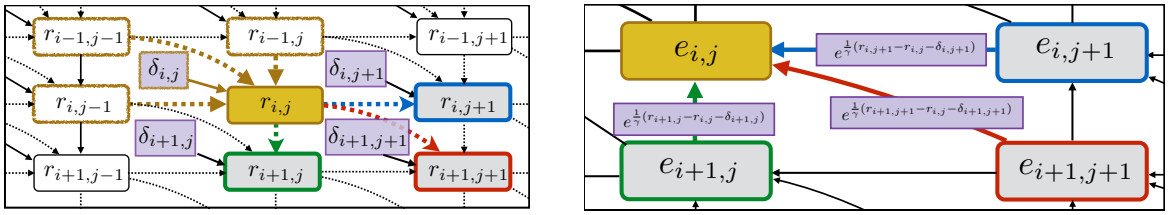
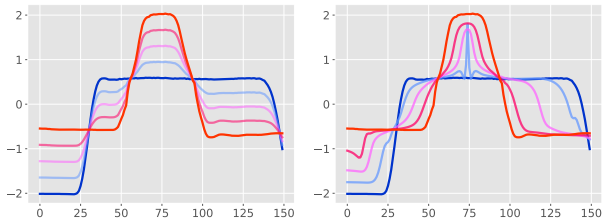


Figure 3. Sketch of the computational graph for soft-DTW, in the forward pass used to compute $\mathbf{dtw}_\gamma$ (left) and backward pass used to compute its gradient $\nabla_{\mathbf{x}} \mathbf{dtw}_\gamma$ (right). In both diagrams, purple shaded cells stand for data values available before the recursion starts, namely cost values (left) and multipliers computed using forward pass results (right). In the left diagram, the forward computation of $r_{i,j}$ as a function of its predecessors and $\delta_{i,j}$ is summarized with arrows. Dotted lines indicate a $\min^\gamma$ operation, solid lines an addition. From the perspective of the final term $r_{n,m}$, which stores $\mathbf{dtw}_\gamma(\mathbf{x}, \mathbf{y})$ at the lower right corner (not shown) of the computational graph, a change in $r_{i,j}$ only impacts $r_{n,m}$ through changes that $r_{i,j}$ causes to $r_{i+1,j}$, $r_{i,j+1}$ and $r_{i+1,j+1}$. These changes can be tracked using Eq. (2.3,2.3) and appear in lines 9-11 in Algorithm 2 as variables a, b, c, as well as in the purple shaded boxes in the backward pass (right) which represents the recursion of line 12 in Algorithm 2.

m_N are equal), as can be seen in the intuitive interpolations we obtain between two time series shown in Figure 4.

Non-convexity of $\mathbf{dtw}_\gamma$. A natural question that arises from Eq. (4) is whether that objective is convex or not. The answer is negative, in a way that echoes the non-convexity of the k-means objective as a function of cluster centroids locations. Indeed, for any alignment matrix A of suitable size, each map $\mathbf{x} \mapsto \langle A, \Delta(\mathbf{x}, \mathbf{y}) \rangle$ shares the same convexity/concavity property that δ may have. However, both $\min$ and $\min^\gamma$ can only preserve the *concavity* of elementary functions (Boyd & Vandenberghe, 2004, pp.72-74). Therefore $\mathbf{dtw}_\gamma$ will only be concave if δ is concave, or become instead a (non-convex) (soft) minimum of convex functions if δ is convex. When δ is a squared-Euclidean distance, $\mathbf{dtw}_0$ is a piecewise quadratic function of $\mathbf{x}$, as is also the case with the k-means energy (see for instance Figure 2 in Schultz & Jain 2017). Since this is the setting we consider here, all of the computations involving barycenters should be taken with a grain of salt, since we have no way of ensuring optimality when approximating Eq. (4).

Smoothing helps optimizing $\mathbf{dtw}_\gamma$. Smoothing can be regarded, however, as a way to "convexify" $\mathbf{dtw}_\gamma$. Indeed, notice that $\mathbf{dtw}_\gamma$ converges to the sum of all costs as $\gamma \to \infty$. Therefore, if δ is convex, $\mathbf{dtw}_\gamma$ will gradually become convex as γ grows. For smaller values of γ, one can intuitively foresee that using $\min^\gamma$ instead of a minimum will smooth out local minima and therefore provide a better (although slightly different from $\mathbf{dtw}_0$) optimization landscape. We believe this is why our approach recovers better results, **even when measured in the original $\mathbf{dtw}_0$ discrepancy**, than subgradient or alternating minimization approaches such as DBA (Petitjean et al., 2011), which can, on the contrary, get more easily stuck in local minima. Evidence for this statement is presented in the experimental section.



(a) Euclidean loss (b) Soft-DTW loss ($\gamma = 1$)

Figure 4. Interpolation between two time series (red and blue) on the Gun Point dataset. We computed the barycenter by solving Eq. (4) with (λ_1, λ_2) set to (0.25, 0.75), (0.5, 0.5) and (0.75, 0.25). The soft-DTW geometry leads to visibly different interpolations.

3.2. Clustering with the soft-DTW geometry

The (approximate) computation of $\mathbf{dtw}_\gamma$ barycenters can be seen as a first step towards the task of clustering time series under the $\mathbf{dtw}_\gamma$ discrepancy. Indeed, one can naturally formulate that problem as that of finding centroids $\mathbf{x}_1, \ldots, \mathbf{x}_k$ that minimize the following energy:

$$\min_{\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^{p \times n}} \sum_{i=1}^{N} \frac{1}{m_i} \min_{j \in [[k]]} \mathbf{dtw}_\gamma(\mathbf{x}_j, \mathbf{y}_i). \qquad (5)$$

To solve that problem one can resort to a direct generalization of Lloyd's algorithm (1982) in which each centering step and each clustering allocation step is done according to the $\mathbf{dtw}_\gamma$ discrepancy.

3.3. Learning prototypes for time series classification

One of the de-facto baselines for learning to classify time series is the k nearest neighbors (k-NN) algorithm, combined with DTW as discrepancy measure between time series. However, k-NN has two main drawbacks. First, the time series used for training must be stored, leading to potentially high storage cost. Second, in order to com-

pute predictions on new time series, the DTW discrepancy must be computed with all training time series, leading to high computational cost. Both of these drawbacks can be addressed by the nearest centroid classifier (Hastie et al., 2001, p.670), (Tibshirani et al., 2002). This method chooses the class whose barycenter (centroid) is closest to the time series to classify. Although very simple, this method was shown to be competitive with k-NN, while requiring much lower computational cost at prediction time (Petitjean et al., 2014). Soft-DTW can naturally be used in a nearest centroid classifier, in order to compute the barycenter of each class at train time, and to compute the discrepancy between barycenters and time series, at prediction time.

3.4. Multistep-ahead prediction

Soft-DTW is ideally suited as a loss function for any task that requires time series outputs. As an example of such a task, we consider the problem of, given the first $1, \dots, t$ observations of a time series, predicting the remaining $(t+1), \dots, n$ observations. Let $\mathbf{x}^{t,t'} \in \mathbb{R}^{p \times (t'-t+1)}$ be the submatrix of $\mathbf{x} \in \mathbb{R}^{p \times n}$ of all columns with indices between t and t', where $1 \leq t < t' < n$. Learning to predict the segment of a time series can be cast as the problem

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \mathbf{dtw}_\gamma \left(f_\theta(\mathbf{x}_i^{1,t}), \mathbf{x}_i^{t+1,n} \right),$$

where $\{f_\theta\}$ is a set of parameterized function that take as input a time series and outputs a time series. Natural choices would be multi-layer perceptrons or recurrent neural networks (RNN), which have been historically trained with a Euclidean loss (Parlos et al., 2000, Eq.5).

4. Experimental results

Throughout this section, we use the UCR (University of California, Riverside) time series classification archive (Chen et al., 2015). We use a subset containing 79 datasets encompassing a wide variety of fields (astronomy, geology, medical imaging) and lengths. Datasets include class information (up to 60 classes) for each time series and are split into train and test sets. Due to the large number of datasets in the UCR archive, we choose to report only a summary of our results in the main manuscript. Detailed results are included in the appendices for interested readers.

4.1. Averaging experiments

In this section, we compare the soft-DTW barycenter approach presented in §3.1 to DBA (Petitjean et al., 2011) and a simple batch subgradient method.

Experimental setup. For each dataset, we choose a class at random, pick 10 time series in that class and compute their barycenter. For quantitative results below, we repeat this procedure 10 times and report the averaged results. For each method, we set the maximum number of iterations to 100. To minimize the proposed soft-DTW barycenter objective, Eq. (4), we use L-BFGS.

Table 1. Percentage of the datasets on which the proposed soft-DTW barycenter is achieving lower DTW loss (Equation (4) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	40.51%	3.80%
$\gamma = 0.1$	93.67%	46.83%
$\gamma = 0.01$	100%	79.75%
$\gamma = 0.001$	97.47%	89.87%
Comparison with subgradient method		
$\gamma = 1$	96.20%	35.44%
$\gamma = 0.1$	97.47%	72.15%
$\gamma = 0.01$	97.47%	92.41%
$\gamma = 0.001$	97.47%	97.47%

Qualitative results. We first visualize the barycenters obtained by soft-DTW when $\gamma = 1$ and $\gamma = 0.01$, by DBA and by the subgradient method. Figure 5 shows barycenters obtained using random initialization on the ECG200 dataset. More results with both random and Euclidean mean initialization are given in Appendix B and C.

We observe that both DBA or soft-DTW with low smoothing parameter γ yield barycenters that are spurious. On the other hand, a descent on the soft-DTW loss with sufficiently high γ converges to a reasonable solution. For example, as indicated in Figure 5 with DTW or soft-DTW ($\gamma = 0.01$), the small kink around $x = 15$ is not representative of any of the time series in the dataset. However, with soft-DTW ($\gamma = 1$), the barycenter closely matches the time series. This suggests that DTW or soft-DTW with too low γ can get stuck in bad local minima.

When using Euclidean mean initialization (only possible if time series have the same length), DTW or soft-DTW with low γ often yield barycenters that better match the shape of the time series. However, they tend to overfit: they absorb the idiosyncrasies of the data. In contrast, soft-DTW is able to learn barycenters that are much smoother.

Quantitative results. Table 1 summarizes the percentage of datasets on which the proposed soft-DTW barycenter achieves lower DTW loss when varying the smoothing parameter γ. The actual loss values achieved by different methods are indicated in Appendix G and Appendix H.

As γ decreases, soft-DTW achieves a lower DTW loss than other methods on almost all datasets. This confirms our

IPR2026-00272
Krisp Technologies EX1010 Page 317

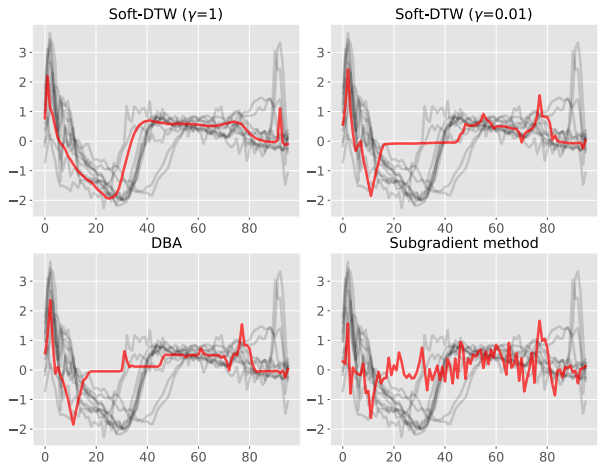
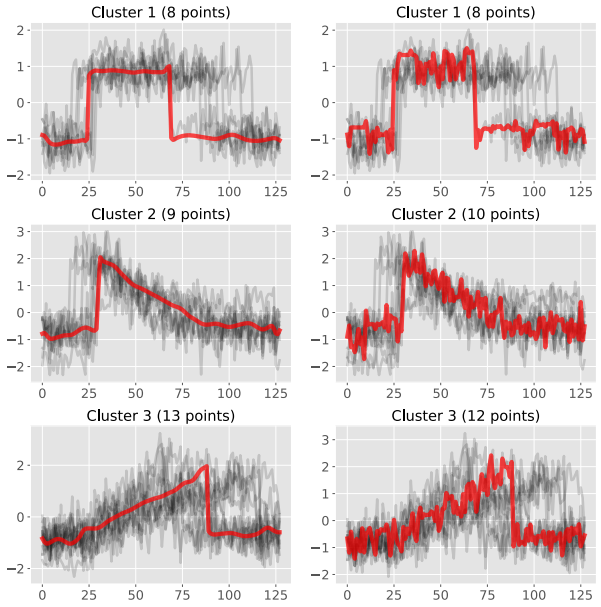


Figure 5. Comparison between our proposed soft barycenter and the barycenter obtained by DBA and the subgradient method, on the ECG200 dataset. When DTW is insufficiently smoothed, barycenters often get stuck in a bad local minimum that does not correctly match the time series.



(a) Soft-DTW ($\gamma = 1$) (b) DBA

Figure 6. Clusters obtained on the CBF dataset when plugging our proposed soft barycenter and that of DBA in Lloyd's algorithm. DBA absorbs the idiosyncrasies of the data, while soft-DTW can learn much smoother barycenters.

claim that the smoothness of soft-DTW leads to an objective that is better behaved and more amenable to optimization by gradient-descent methods.

4.2. k-means clustering experiments

We consider in this section the same computational tools used in §4.1 above, but use them to cluster time series.

Experimental setup. For all datasets, the number of clusters k is equal to the number of classes available in the dataset. Lloyd's algorithm alternates between a centering step (barycenter computation) and an assignment step. We set the maximum number of outer iterations to 30 and the maximum number of inner (barycenter) iterations to 100, as before. Again, for soft-DTW, we use L-BFGS.

Qualitative results. Figure 6 shows the clusters obtained when runing Lloyd's algorithm on the CBF dataset with soft-DTW ($\gamma = 1$) and DBA, in the case of random initialization. More results are included in Appendix E. Clearly, DTW absorbs the tiny details in the data, while soft-DTW is able to learn much smoother barycenters.

Quantitative results. Table 2 summarizes the percentage of datasets on which soft-DTW barycenter achieves lower k-means loss under DTW, i.e. Eq. (5) with $\gamma = 0$. The actual loss values achieved by all methods are indicated in Appendix I and Appendix J. The results confirm the same trend as for the barycenter experiments. Namely, as γ decreases, soft-DTW is able to achieve lower loss than other methods on a large proportion of the datasets. Note that we have not run experiments with smaller values of γ than 0.001, since $\mathbf{dtw}_{0.001}$ is very close to $\mathbf{dtw}_0$ in practice.

4.3. Time-series classification experiments

In this section, we investigate whether the smoothing in soft-DTW can act as a useful regularization and improve classification accuracy in the nearest centroid classifier.

Experimental setup. We use 50% of the data for training, 25% for validation and 25% for testing. We choose γ from 15 log-spaced values between 10^{-3} and 10.

Quantitative results. Each point in Figure 7 above the diagonal line represents a dataset for which using soft-DTW for barycenter computation rather than DBA improves the accuracy of the nearest centroid classifier. To summarize, we found that soft-DTW is working better or at least as well as DBA in 75% of the datasets.

4.4. Multistep-ahead prediction experiments

In this section, we present preliminary experiments for the task of multistep-ahead prediction, described in §3.4.

Experimental setup. We use the training and test sets predefined in the UCR archive. In both the training and test sets, we use the first 60% of the time series as input and the remaining 40% as output, ignoring class information. We then use the training set to learn a model that predicts the outputs from inputs and the test set to evaluate results with both Euclidean and DTW losses. In this experiment, we focus on a simple multi-layer perceptron (MLP) with one

IPR2026-00272
Krisp Technologies EX1010 Page 318

Table 2. Percentage of the datasets on which the proposed soft-DTW based k-**means** is achieving lower DTW loss (Equation (5) with $\gamma = 0$) than competing methods.

	Random initialization	Euclidean mean initialization
Comparison with DBA		
$\gamma = 1$	15.78%	29.31%
$\gamma = 0.1$	24.56%	24.13%
$\gamma = 0.01$	59.64%	55.17%
$\gamma = 0.001$	77.19%	68.97%
Comparison with subgradient method		
$\gamma = 1$	42.10%	46.44%
$\gamma = 0.1$	57.89%	50%
$\gamma = 0.01$	76.43%	65.52%
$\gamma = 0.001$	96.49%	84.48%

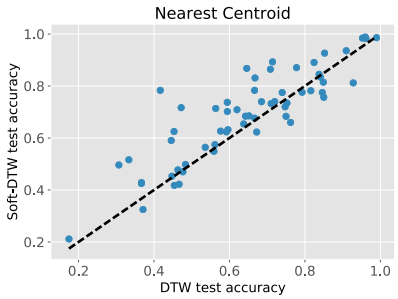


Figure 7. Each point above the diagonal represents a dataset where using our soft-DTW barycenter rather than that of DBA improves the accuracy of the nearest nearest centroid classifier. This is the case for **75%** of the datasets in the UCR archive.

hidden layer and sigmoid activation. We also experimented with linear models and recurrent neural networks (RNNs) but they did not improve over a simple MLP.

Implementation details. Deep learning frameworks such as Theano, TensorFlow and Chainer allow the user to specify a custom backward pass for their function. Implementing such a backward pass, rather than resorting to automatic differentiation (autodiff), is particularly important in the case of soft-DTW: First, the autodiff in these frameworks is designed for vectorized operations, whereas the dynamic program used by the forward pass of Algorithm 1 is inherently element-wise; Second, as we explained in §2.2, our backward pass is able to re-use log-sum-exp computations from the forward pass, leading to both lower computational cost and better numerical stability. We implemented a custom backward pass in Chainer, which can then be used to plug soft-DTW as a loss function in any network architecture. To estimate the MLP's parameters, we used Chainer's implementation of Adam (Kingma & Ba, 2014).

Qualitative results. Visualizations of the predictions obtained under Euclidean and soft-DTW losses are given in Figure 1, as well as in Appendix F. We find that for simple one-dimensional time series, an MLP works very well, showing its ability to capture patterns in the training set. Although the predictions under Euclidean and soft-DTW losses often agree with each other, they can sometimes be visibly different. Predictions under soft-DTW loss can confidently predict abrupt and sharp changes since those have a low DTW cost as long as such a sharp change is present, under a small time shift, in the ground truth.

Table 3. Averaged rank obtained by a multi-layer perceptron (MLP) under Euclidean and soft-DTW losses. Euclidean initialization means that we initialize the MLP trained with soft-DTW loss by the solution of the MLP trained with Euclidean loss.

Training loss	Random initialization	Euclidean initialization
When evaluating with DTW loss		
Euclidean	3.46	4.21
soft-DTW ($\gamma = 1$)	3.55	3.96
soft-DTW ($\gamma = 0.1$)	3.33	3.42
soft-DTW ($\gamma = 0.01$)	2.79	2.12
soft-DTW ($\gamma = 0.001$)	**1.87**	**1.29**
When evaluating with Euclidean loss		
Euclidean	**1.05**	**1.70**
soft-DTW ($\gamma = 1$)	2.41	2.99
soft-DTW ($\gamma = 0.1$)	3.42	3.38
soft-DTW ($\gamma = 0.01$)	4.13	3.64
soft-DTW ($\gamma = 0.001$)	3.99	3.29

Quantitative results. A comparison summary of our MLP under Euclidean and soft-DTW losses over the UCR archive is given in Table 3. Detailed results are given in the appendix. Unsurprisingly, we achieve lower DTW loss when training with the soft-DTW loss, and lower Euclidean loss when training with the Euclidean loss. Because DTW is robust to several useful invariances, a small error in the soft-DTW sense could be a more judicious choice than an error in an Euclidean sense for many applications.

5. Conclusion

We propose in this paper to turn the popular DTW discrepancy between time series into a full-fledged loss function between ground truth time series and outputs from a learning machine. We have shown experimentally that, on the existing problem of computing barycenters and clusters for time series data, our computational approach is superior to existing baselines. We have shown promising results on the problem of multistep-ahead time series prediction, which could prove extremely useful in settings where a user's actual loss function for time series is closer to the robust perspective given by DTW, than to the local parsing of the Euclidean distance.

Acknowledgements. MC gratefully acknowledges the support of a *chaire de l'IDEX Paris Saclay*.

IPR2026-00272
Krisp Technologies EX1010 Page 319

References

Bahl, L and Jelinek, Frederick. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, 21(4):404–411, 1975.

Bakir, GH, Hofmann, T, Schölkopf, B, Smola, AJ, Taskar, B, and Vishwanathan, SVN. *Predicting Structured Data*. Advances in neural information processing systems. MIT Press, Cambridge, MA, USA, 2007.

Bellman, Richard. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8): 716–719, 1952.

Blondel, Mathieu, Fujino, Akinori, Ueda, Naonori, and Ishihata, Masakazu. Higher-order factorization machines. In *Advances in Neural Information Processing Systems 29*, pp. 3351–3359. 2016.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Chen, Yanping, Keogh, Eamonn, Hu, Bing, Begum, Nurjahan, Bagnall, Anthony, Mueen, Abdullah, and Batista, Gustavo. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.

Cuturi, Marco. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 929–936, 2011.

Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 685–693, 2014.

Cuturi, Marco, Vert, Jean-Philippe, Birkenes, Oystein, and Matsui, Tomoko. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pp. II–413, 2007.

Fréchet, Maurice. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pp. 215–310. Presses universitaires de France, 1948.

Garreau, Damien, Lajugie, Rémi, Arlot, Sylvain, and Bach, Francis. Metric learning for temporal sequence alignment. In *Advances in Neural Information Processing Systems*, pp. 1817–1825, 2014.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lasserre, Jean B. *Linear and integer programming vs linear integration and counting: a duality viewpoint*. Springer Science & Business Media, 2009.

Lloyd, Stuart. Least squares quantization in pcm. *IEEE Trans. on Information Theory*, 28(2):129–137, 1982.

Lütkepohl, Helmut. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Parlos, Alexander G, Rais, Omar T, and Atiya, Amir F. Multi-step-ahead prediction using dynamic recurrent neural networks. *Neural networks*, 13(7):765–786, 2000.

Petitjean, François and Gançarski, Pierre. Summarizing a set of time series by averaging: From steiner sequence to compact multiple alignment. *Theoretical Computer Science*, 414(1):76–91, 2012.

Petitjean, François, Ketterlin, Alain, and Gançarski, Pierre. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44 (3):678–693, 2011.

Petitjean, François, Forestier, Germain, Webb, Geoffrey I, Nicholson, Ann E, Chen, Yanping, and Keogh, Eamonn. Dynamic time warping averaging of time series allows faster and more accurate classification. In *ICDM*, pp. 470–479. IEEE, 2014.

Ristad, Eric Sven and Yianilos, Peter N. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.

Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed Wasserstein loss. *Proceedings of AISTATS'16*, 2016.

Saigo, Hiroto, Vert, Jean-Philippe, Ueda, Nobuhisa, and Akutsu, Tatsuya. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

Saigo, Hiroto, Vert, Jean-Philippe, and Akutsu, Tatsuya. Optimizing amino acid substitution matrices with a local alignment kernel. *BMC bioinformatics*, 7(1):246, 2006.

Sakoe, Hiroaki and Chiba, Seibi. A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics, Budapest*, volume 3, pp. 65–69, 1971.

Sakoe, Hiroaki and Chiba, Seibi. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, 26:43–49, 1978.

IPR2026-00272
Krisp Technologies EX1010 Page 320

Schultz, David and Jain, Brijnesh. Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. *arXiv preprint arXiv:1701.06393*, 2017.

Tibshirani, Robert, Hastie, Trevor, Narasimhan, Balasubramanian, and Chu, Gilbert. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10): 6567–6572, 2002.

Yi, Byoung-Kee, Jagadish, HV, and Faloutsos, Christos. Efficient retrieval of similar time sequences under time warping. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pp. 201–208. IEEE, 1998.

Zhang, C., Frogner, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems 29*, 2015.

Appendix material

A. Recursive forward computation of the average path matrix

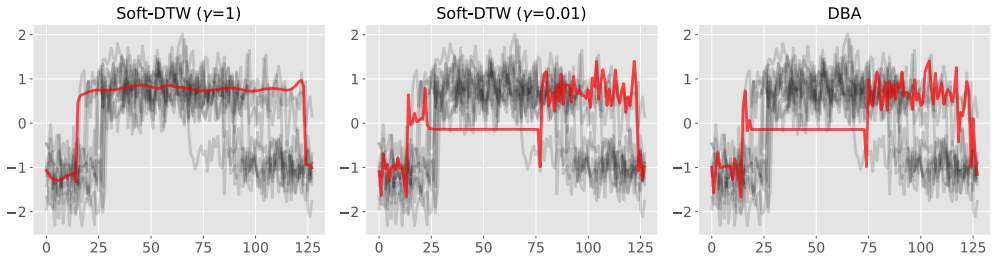
The average alignment under Gibbs distribution p_γ can be computed with the following forward recurrence, which mimics closely Bellman's original recursion. For each $i \in [\![n]\!], j \in [\![m]\!]$, define

$$E_{i+1,j+1} = \begin{bmatrix} e^{-\delta_{i+1,j+1}/\gamma} E_{i,j} & \mathbf{0}_i \\ \mathbf{0}_j^T & e^{-r_{i+1,j+1}/\gamma} \end{bmatrix} + \begin{bmatrix} e^{-\delta_{i+1,j+1}/\gamma} E_{i,j+1} \\ \mathbf{0}_j^T \quad e^{-r_{i+1,j+1}/\gamma} \end{bmatrix} + \begin{bmatrix} e^{-\delta_{i+1,j+1}/\gamma} E_{i+1,j} & \mathbf{0}_i \\ & e^{-r_{ij}/\gamma} \end{bmatrix}$$

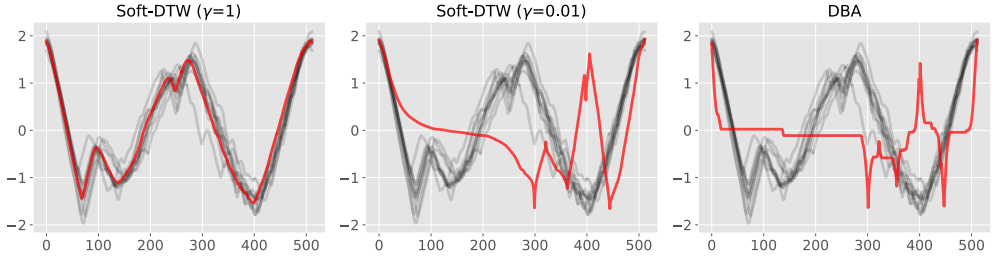
Here terms r_{ij} are computed following the recursion in Algorithm 2. Border matrices are initialized to 0, except for $E_{1,1}$ which is initialized to $[1]$. Upon completion, the average alignment matrix is stored in $E_{n,m}$.

The operation above consists in summing three matrices of size $(i+1, j+1)$. There are exactly (nm) such updates. A careful implementation of this algorithm, that would only store two arrays of matrices, as Algorithm 1 only store two arrays of values, can be carried out in $nm \min(n, m)$ space but it would still require $(nm)^2$ operations.

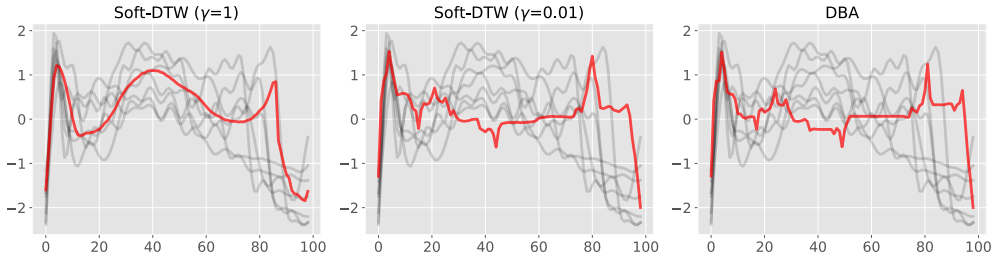
B. Barycenters obtained with random initialization



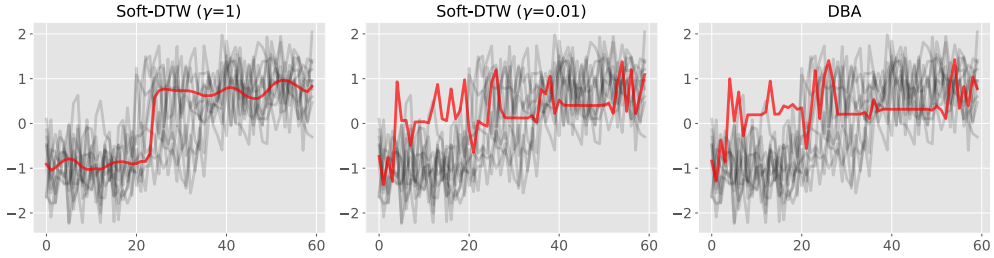
(a) CBF



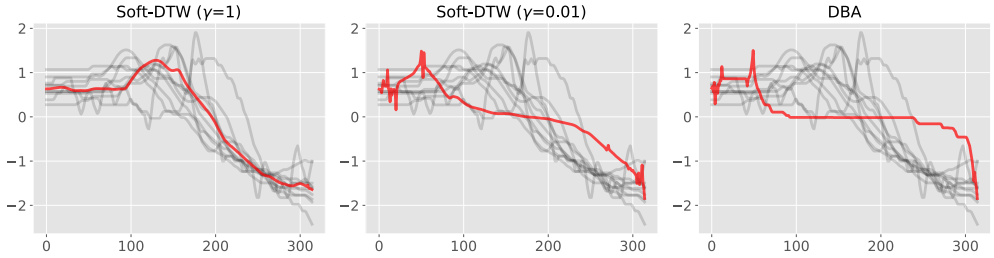
(b) Herring



(c) Medical Images



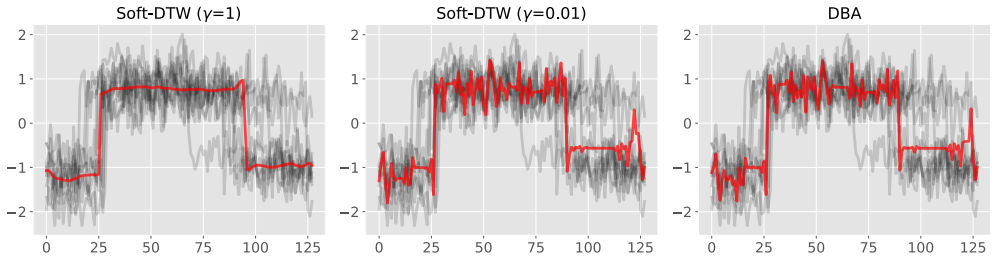
(d) Synthetic Control



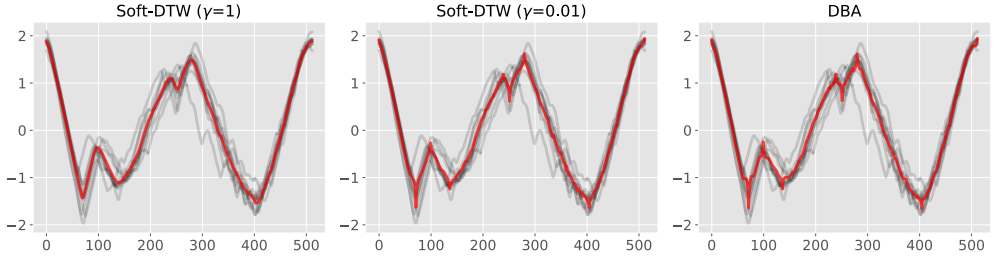
(e) Wave Gesture Library Y

IPR2026-00272
Krisp Technologies EX1010 Page 323

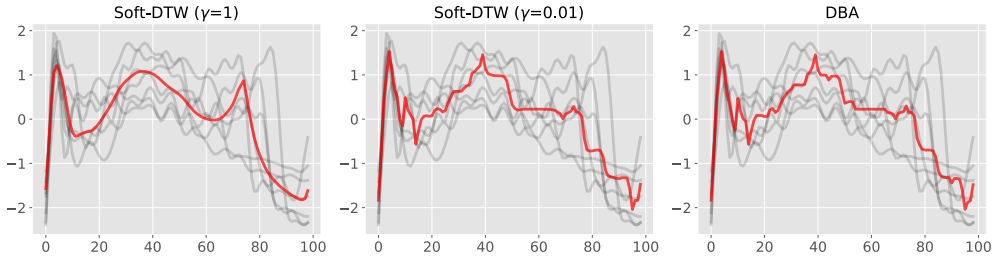
C. Barycenters obtained with Euclidean mean initialization



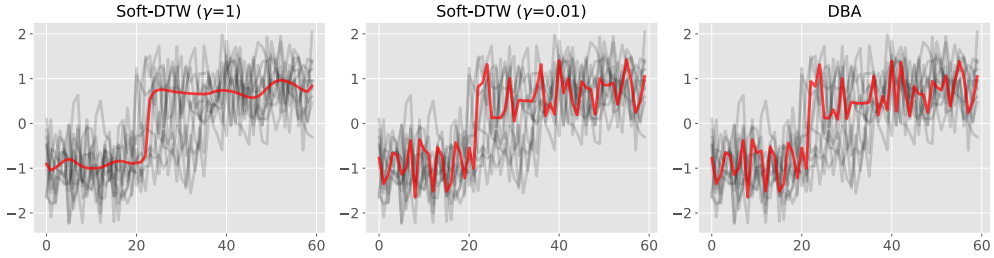
(a) CBF



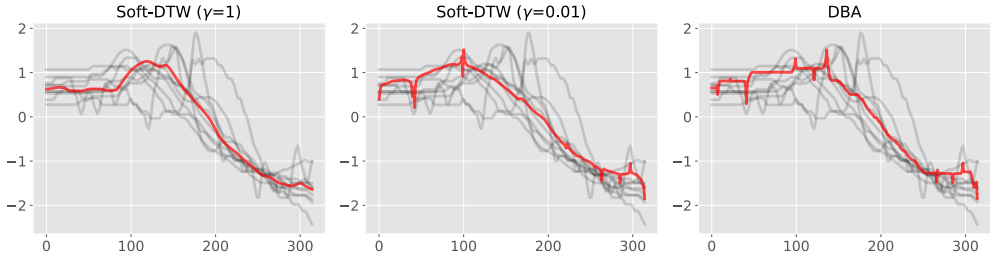
(b) Herring



(c) Medical Images



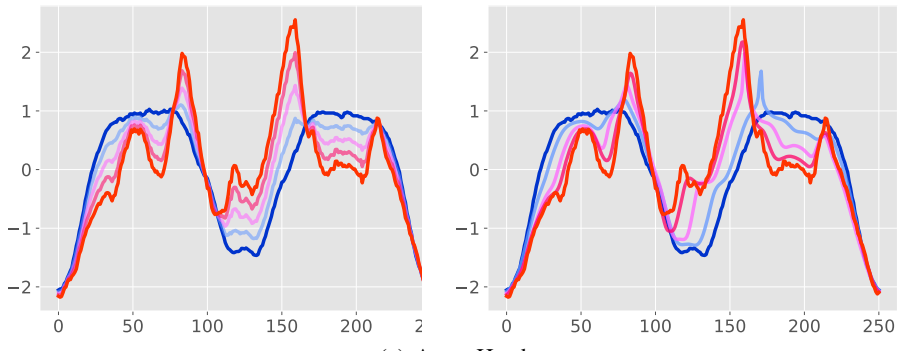
(d) Synthetic Control

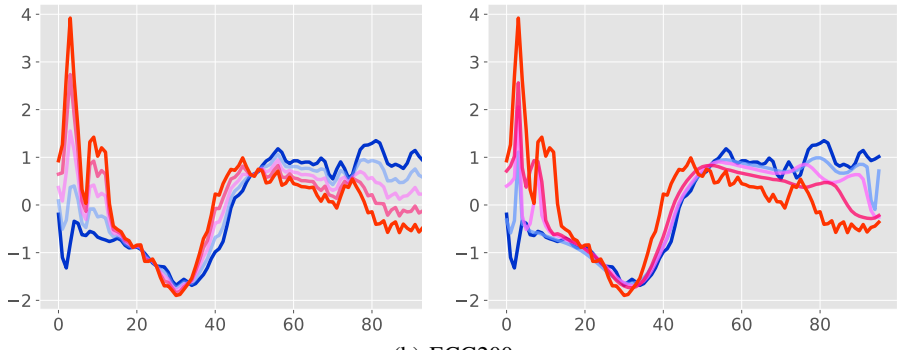


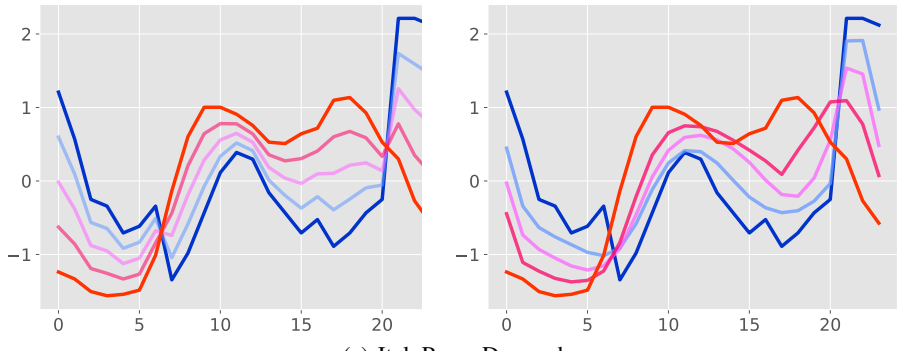
(e) Wave Gesture Library Y

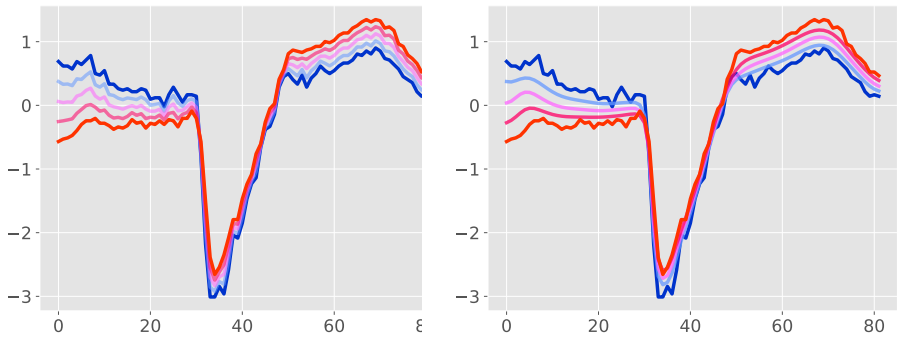
D. More interpolation results

Left: results obtained under Euclidean loss. Right: results obtained under soft-DTW ($\alpha = 1$) loss.







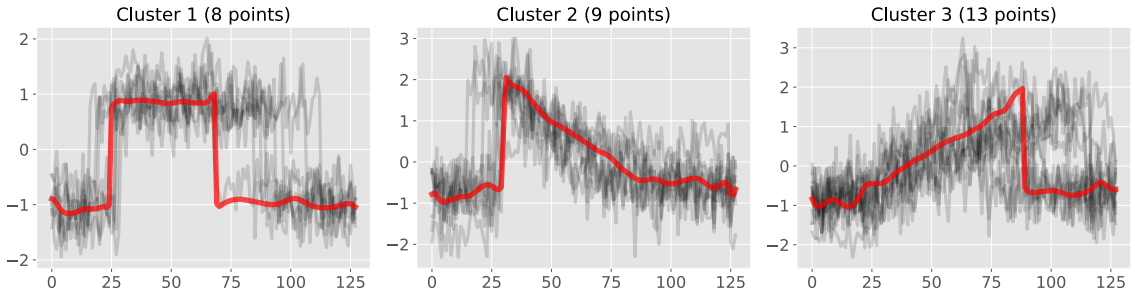


(d) TwoLeadECG

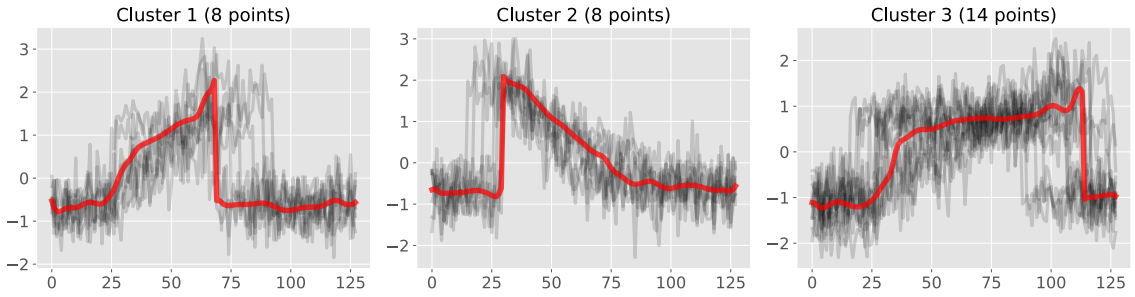
IPR2026-00272
Krisp Technologies EX1010 Page 325

E. Clusters obtained by k-means under DTW or soft-DTW geometry

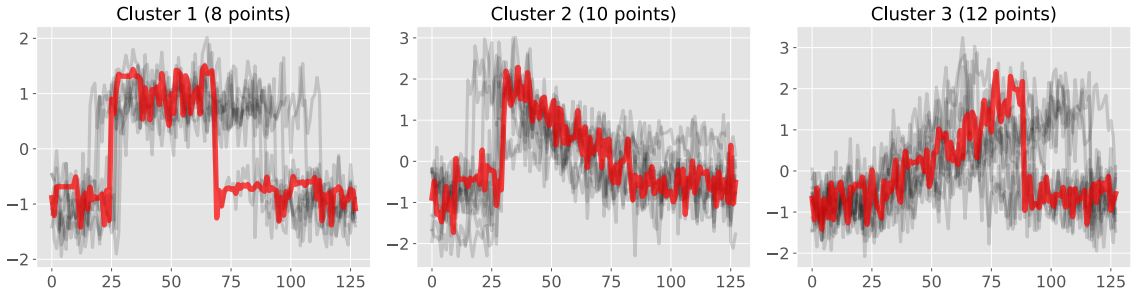
CBF dataset



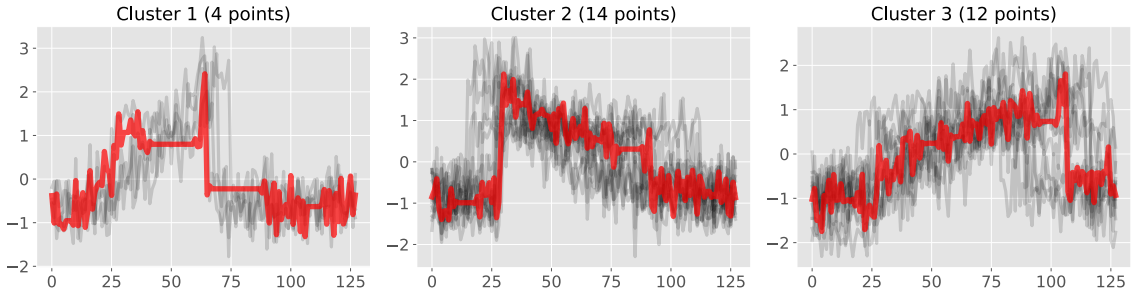
(a) Soft-DTW ($\gamma = 1$, random initialization)



(b) Soft-DTW ($\gamma = 1$, Euclidean mean initialization)

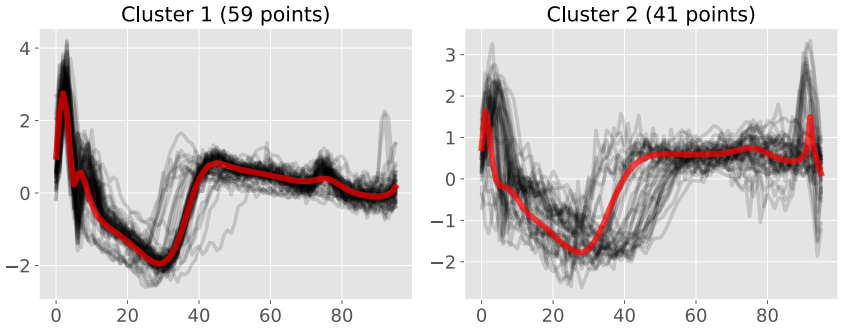


(c) DBA (random initialization)

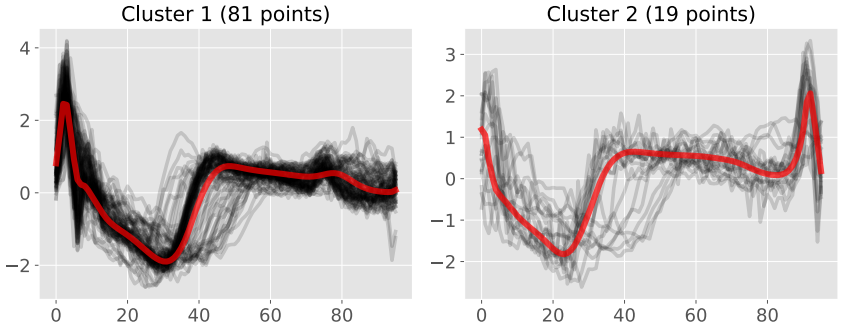


(d) DBA (Euclidean mean initialization)

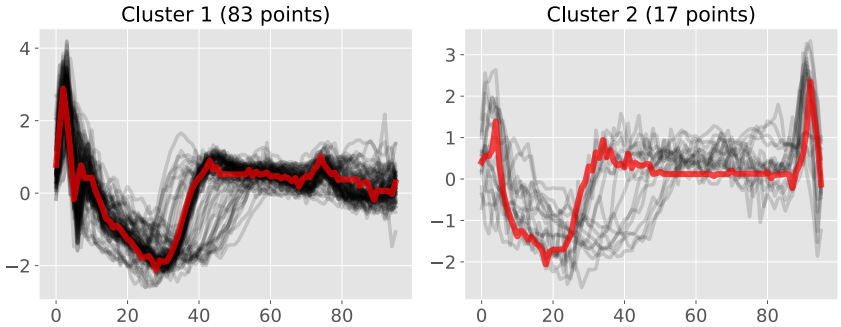
ECG200 dataset



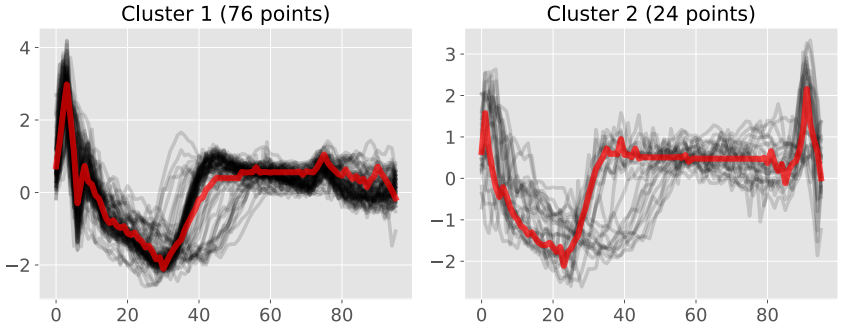
(a) Soft-DTW ($\gamma = 1$, random initialization)



(b) Soft-DTW ($\gamma = 1$, Euclidean mean initialization)

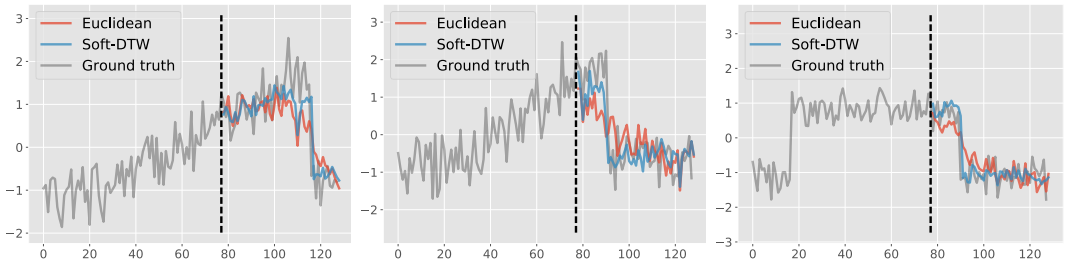


(c) DBA (random initialization)

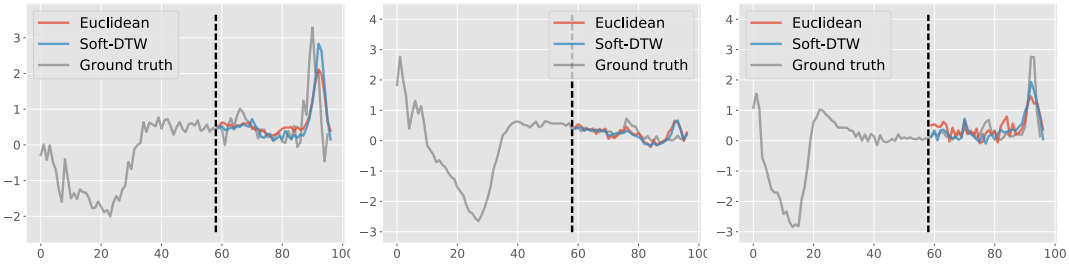


(d) DBA (Euclidean mean initialization)

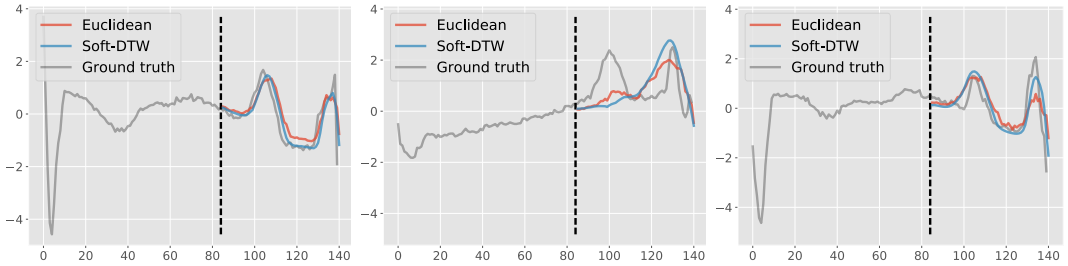
F. More visualizations of time-series prediction



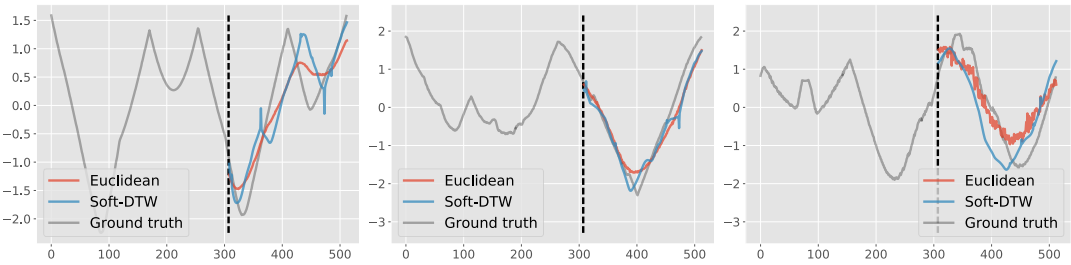
(a) CBF



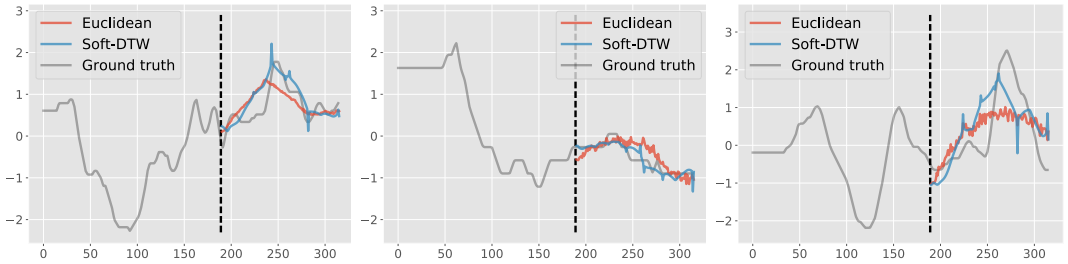
(b) ECG200



(c) ECG5000



(d) ShapesAll



(e) uWaveGestureLibrary_Y

IPR2026-00272
Krisp Technologies EX1010 Page 328

G. Barycenters: DTW loss (Eq. 4 with $\gamma = 0$) achieved with random init

Dataset	Soft-DTW $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Subgradient method	DBA	Euclidean mean
50words	5.000	2.785	**2.513**	2.721	44.399	4.554	25.388
Adiac	0.235	0.207	0.257	0.428	25.533	0.754	**0.177**
ArrowHead	2.390	1.598	**1.487**	1.664	36.125	2.512	2.743
Beef	10.471	6.541	**6.200**	6.238	88.100	7.780	25.347
BeetleFly	35.790	23.655	**22.559**	23.105	77.993	25.122	191.574
BirdChicken	23.300	12.542	**11.164**	11.954	45.777	12.820	92.061
CBF	21.098	**11.949**	12.564	12.667	30.281	14.836	28.236
Car	2.639	1.750	**1.611**	1.914	80.437	2.609	5.106
ChlorineConcentration	22.260	**13.932**	14.818	15.044	32.134	16.168	15.411
CinC_ECG_torso	118.872	80.248	**76.536**	76.812	262.221	90.663	761.238
Coffee	1.036	0.871	1.262	1.630	41.741	2.380	**0.591**
Computers	231.421	182.380	**178.184**	179.886	∞	183.388	391.830
Cricket_X	41.514	29.290	**28.424**	28.851	70.128	28.955	104.699
Cricket_Y	51.858	**30.321**	30.337	31.041	70.989	33.098	107.712
Cricket_Z	43.458	30.264	**28.668**	29.373	68.382	32.182	128.129
DiatomSizeReduction	0.055	0.054	0.064	0.132	49.308	0.418	**0.033**
DistalPhalanxOutlineAgeGroup	1.380	1.074	1.407	1.509	11.539	1.761	**0.981**
DistalPhalanxOutlineCorrect	2.501	**1.968**	2.267	2.634	13.169	2.991	2.374
DistalPhalanxTW	1.148	0.906	1.015	1.159	10.957	1.228	**0.867**
ECG200	7.374	**6.400**	6.871	7.047	19.514	8.257	10.107
ECG5000	**8.951**	8.961	10.601	10.265	34.558	12.098	14.517
ECGFiveDays	9.816	**9.019**	9.364	9.407	17.898	9.837	23.975
Earthquakes	148.959	**85.219**	85.470	85.515	∞	85.788	330.776
ElectricDevices	27.852	**23.769**	23.783	24.009	∞	24.869	56.470
FISH	0.978	**0.641**	0.662	0.957	63.566	1.680	1.543
FaceAll	15.068	**12.373**	13.248	13.494	24.582	15.980	29.982
FaceFour	15.500	**14.519**	15.002	14.849	∞	16.339	25.410
FacesUCR	15.033	**13.077**	13.405	14.174	30.621	14.796	27.026
FordA	56.936	**45.492**	46.170	46.038	96.087	49.723	218.482
FordB	59.117	47.812	**47.058**	47.642	102.279	50.262	250.595
Gun_Point	7.204	2.507	**2.037**	2.211	22.590	2.374	7.286
Ham	24.833	**19.101**	20.397	20.713	55.769	22.807	30.685
HandOutlines	3.400	**2.690**	2.814	28.759	353.235	3.422	7.838
Haptics	16.424	14.351	**14.129**	14.320	172.988	16.464	39.559
Herring	1.212	**0.946**	1.022	1.349	71.388	2.097	1.884
InlineSkate	83.107	29.672	**22.819**	N/A	N/A	N/A	N/A
ItalyPowerDemand	2.442	**2.124**	2.316	2.372	5.434	2.355	2.329
MedicalImages	6.934	**5.809**	5.980	6.089	22.777	6.252	10.911
MiddlePhalanxOutlineAgeGroup	0.858	0.753	1.305	1.375	11.474	1.605	**0.624**
MiddlePhalanxOutlineCorrect	0.832	0.714	0.985	1.030	11.643	1.678	**0.611**
MiddlePhalanxTW	0.755	0.581	0.963	1.206	10.684	1.274	**0.447**
MoteStrain	24.177	21.639	21.616	**21.554**	32.007	22.437	26.646
NonInvasiveFatalECG_Thorax1	6.671	**3.324**	4.738	5.031	59.162	6.378	3.568
NonInvasiveFatalECG_Thorax2	**2.559**	3.159	3.587	4.097	40339.200	6.494	2.864
OSULeaf	30.041	20.692	20.034	**19.950**	76.057	23.915	136.512
OliveOil	0.657	0.959	1.494	1.804	95.499	3.420	**0.008**
PhalangesOutlinesCorrect	1.383	**1.114**	1.405	1.516	11.070	1.743	1.210
Phoneme	99.205	**72.412**	73.666	73.767	157.124	78.664	138.157
Plane	1.079	**0.849**	1.220	1.629	20.328	2.111	1.209
ProximalPhalanxOutlineAgeGroup	0.618	0.511	0.691	0.878	10.437	1.177	**0.322**
ProximalPhalanxOutlineCorrect	0.749	0.654	0.833	0.882	10.767	1.111	**0.615**
ProximalPhalanxTW	0.653	0.536	0.672	0.778	10.377	1.133	**0.462**
RefrigerationDevices	159.745	146.601	**140.634**	141.200	∞	148.931	363.732
ScreenType	156.442	123.746	**121.432**	122.247	∞	132.748	334.379
ShapeletSim	236.039	**123.605**	125.657	126.062	154.480	130.428	283.458
ShapesAll	15.267	7.108	**6.466**	6.818	58.734	8.236	67.790
SmallKitchenAppliances	176.073	164.419	**162.009**	162.585	∞	168.170	524.542
SonyAIBORobotSurface	5.735	**4.916**	5.425	5.337	∞	5.813	5.430
SonyAIBORobotSurfaceII	11.994	**11.048**	11.278	11.405	6637415.793	11.481	13.783
StarLightCurves	22.342	11.757	7.934	**7.654**	102.255	11.600	47.353
Strawberry	1.392	**1.191**	1.477	1.627	28.696	2.347	1.300
SwedishLeaf	2.409	**1.968**	2.476	2.904	19.638	3.283	6.274
Symbols	0.845	0.488	**0.439**	0.460	∞	0.828	4.953
ToeSegmentation1	35.904	27.461	**26.554**	26.988	63.040	29.838	129.858
ToeSegmentation2	34.177	24.476	**23.003**	23.194	∞	24.944	170.222
Trace	2.686	1.453	**0.870**	1.031	43.017	2.233	26.037
TwoLeadECG	1.811	**1.514**	1.641	1.701	7.961	1.802	2.216
Two_Patterns	12.048	9.294	**7.764**	8.143	22.489	8.937	60.963
UWaveGestureLibraryAll	68.276	42.692	**38.327**	40.320	∞	49.486	181.901
Wine	0.728	0.500	0.746	1.147	32.463	1.812	**0.094**
WordsSynonyms	9.305	4.917	**4.491**	4.740	48.605	7.209	29.713
Worms	100.683	64.029	61.527	61.296	**35.906**	68.282	421.381
WormsTwoClass	110.292	68.932	66.258	65.964	**37.047**	72.387	430.774
synthetic_control	14.366	**7.115**	7.506	7.516	15.931	8.123	12.187
uWaveGestureLibrary_X	27.610	16.618	14.902	**14.442**	∞	18.269	75.119
uWaveGestureLibrary_Y	29.964	16.106	14.556	**14.450**	∞	15.961	74.405
uWaveGestureLibrary_Z	40.154	24.001	**22.462**	22.656	∞	25.040	107.540
wafer	25.831	**23.595**	25.828	25.195	∞	27.323	65.100
yoga	27.418	13.524	**11.828**	12.051	40.171	15.319	111.236

IPR2026-00272
Krisp Technologies EX1010 Page 329

H. Barycenters: DTW loss (Eq. (4) with $\gamma = 0$) achieved with Euclidean init

Dataset	Soft-DTW $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Subgradient method	DBA	Euclidean mean
50words	5.400	2.895	**2.355**	2.439	4.064	2.595	22.294
Adiac	0.124	0.103	0.089	**0.069**	0.081	0.071	0.103
ArrowHead	2.677	1.759	**1.282**	1.327	1.587	1.411	2.965
Beef	14.814	6.412	**5.252**	5.694	11.112	5.528	31.486
BeetleFly	33.082	20.819	**20.781**	22.127	25.554	21.960	191.285
BirdChicken	21.646	9.445	**7.807**	8.026	473.653	8.243	70.614
CBF	22.498	11.844	**11.433**	11.597	15.321	12.291	28.228
Car	1.556	0.932	**0.693**	0.901	1.171	1.079	2.439
ChlorineConcentration	19.239	10.663	**10.434**	10.468	11.370	10.638	13.549
CinC_ECG_torso	112.562	78.292	69.415	70.383	76.693	**68.641**	751.445
Coffee	1.078	0.657	0.460	**0.393**	0.435	0.399	0.571
Computers	172.590	**138.605**	144.576	146.409	∞	154.956	381.271
Cricket_X	48.334	35.136	**33.103**	33.312	42.018	34.430	125.879
Cricket_Y	41.804	31.395	**31.044**	31.158	35.957	31.749	97.393
Cricket_Z	46.957	**33.453**	34.005	33.708	45.125	36.025	140.474
DiatomSizeReduction	0.039	0.033	0.028	0.021	0.024	**0.019**	0.032
DistalPhalanxOutlineAgeGroup	1.578	0.988	0.784	**0.779**	0.847	0.794	1.075
DistalPhalanxOutlineCorrect	2.878	2.002	**1.751**	1.754	2475.922	1.790	2.780
DistalPhalanxTW	1.377	0.837	0.655	**0.651**	0.773	0.667	0.997
ECG200	7.266	5.608	**5.395**	5.424	5.955	5.494	9.638
ECG5000	12.430	10.377	**10.332**	10.343	12.340	10.595	18.886
ECGFiveDays	8.416	7.452	**7.046**	7.101	145.106	7.145	23.477
Earthquakes	172.035	91.568	**90.684**	91.071	∞	92.126	335.240
ElectricDevices	30.832	**26.480**	27.131	27.076	∞	27.615	57.938
FISH	1.183	0.806	0.541	**0.508**	0.645	0.551	1.933
FaceAll	18.102	13.305	13.104	**13.074**	16.491	13.915	40.404
FaceFour	17.070	13.069	**12.984**	13.091	∞	13.568	28.203
FacesUCR	17.172	**13.081**	13.293	13.394	15.780	13.498	35.942
FordA	53.903	42.199	**41.835**	41.966	53.545	44.259	235.362
FordB	61.168	48.150	**47.327**	47.743	60.120	50.121	246.802
Gun_Point	5.924	2.132	1.695	**1.666**	2.543	1.682	5.906
Ham	25.353	18.841	17.457	**17.294**	∞	17.917	32.456
HandOutlines	2.238	1.718	1.004	0.527	∞	**0.515**	6.452
Haptics	12.554	8.874	**7.785**	8.197	12.193	8.219	35.613
Herring	1.655	1.117	0.809	**0.760**	0.956	0.817	2.564
InlineSkate	100.849	46.460	**27.248**	35.578	N/A	N/A	N/A
ItalyPowerDemand	2.597	1.990	**1.956**	1.985	2.132	1.997	2.449
MedicalImages	5.719	4.319	4.145	**4.070**	4.791	4.371	8.047
MiddlePhalanxOutlineAgeGroup	0.870	0.578	0.427	**0.415**	0.464	0.426	0.552
MiddlePhalanxOutlineCorrect	0.799	0.609	0.460	**0.443**	0.501	0.461	0.577
MiddlePhalanxTW	0.658	0.466	0.335	**0.321**	0.358	0.332	0.434
MoteStrain	24.451	**20.720**	20.829	21.057	∞	21.273	26.694
NonInvasiveFatalECG_Thorax1	1.619	1.384	0.907	0.785	**0.691**	0.814	1.400
NonInvasiveFatalECG_Thorax2	1.624	1.370	0.932	**0.827**	2.163	0.853	1.409
OSULeaf	27.428	18.666	**18.544**	18.595	24.692	20.244	135.980
OliveOil	0.367	0.074	0.022	0.013	0.011	**0.009**	0.011
PhalangesOutlinesCorrect	1.172	0.895	0.699	**0.695**	0.766	0.704	1.002
Phoneme	135.535	**104.971**	105.478	108.031	126.055	108.513	254.392
Plane	0.928	0.600	0.404	**0.399**	0.499	0.430	1.203
ProximalPhalanxOutlineAgeGroup	0.820	0.502	0.361	**0.346**	0.390	0.356	0.512
ProximalPhalanxOutlineCorrect	0.816	0.630	0.463	**0.452**	0.517	0.461	0.669
ProximalPhalanxTW	0.637	0.431	0.313	**0.304**	0.341	0.308	0.471
RefrigerationDevices	154.420	**133.321**	135.721	135.300	∞	142.697	358.823
ScreenType	189.188	143.582	143.894	**141.776**	∞	148.464	325.840
ShapeletSim	231.937	124.443	**122.000**	122.506	154.089	127.977	284.079
ShapesAll	13.416	7.519	**6.420**	6.509	7.317	7.478	80.306
SmallKitchenAppliances	188.030	173.670	169.755	**167.097**	∞	173.004	505.356
SonyAIBORobotSurface	5.715	4.002	3.870	3.896	∞	**3.828**	5.444
SonyAIBORobotSurfaceII	11.300	8.947	**8.853**	8.871	12.651	8.977	14.225
StarLightCurves	13.581	6.619	4.054	**3.765**	7.247	4.517	30.354
Strawberry	2.218	1.413	1.128	**1.070**	1.374	1.156	2.128
SwedishLeaf	2.957	2.068	**2.049**	2.081	2.520	2.163	6.236
Symbols	0.762	0.451	0.412	**0.401**	∞	0.474	4.822
ToeSegmentation1	35.832	26.067	26.337	**25.735**	31.157	27.493	131.683
ToeSegmentation2	34.264	22.238	**20.800**	21.563	∞	23.080	164.101
Trace	1.737	1.744	1.508	**1.378**	4.170	1.969	26.814
TwoLeadECG	1.533	1.172	**1.030**	1.043	1.323	1.093	2.046
Two_Patterns	10.891	7.505	**6.045**	6.079	18.987	6.584	66.027
UWaveGestureLibraryAll	67.549	38.179	**32.894**	33.426	∞	39.241	167.486
Wine	0.707	0.188	0.127	0.111	0.114	**0.110**	0.118
WordsSynonyms	9.804	7.282	**6.711**	6.785	8.884	6.868	39.843
Worms	101.850	61.067	58.725	**56.793**	244.738	63.234	415.674
WormsTwoClass	122.901	68.771	**64.655**	64.898	1297.616	72.011	395.088
synthetic_control	18.147	**9.189**	9.307	9.350	11.520	9.614	19.237
uWaveGestureLibrary_X	34.423	19.787	18.746	**17.807**	∞	24.269	93.839
uWaveGestureLibrary_Y	27.744	14.309	**13.010**	13.607	∞	15.283	51.854
uWaveGestureLibrary_Z	21.927	10.081	8.456	**8.453**	∞	11.040	47.947
wafer	32.561	29.197	28.908	**28.820**	∞	33.379	67.413
yoga	23.698	11.632	9.433	**9.204**	16.239	10.058	93.688

IPR2026-00272
Krisp Technologies EX1010 Page 330

I. k-means clustering: DTW loss achieved (Eq. (5) with $\gamma = 0$, log-scaled) when using random initialization

Dataset	Soft-DTW $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Subgradient method	DBA	Euclidean mean
50words	16.294	16.193	**16.125**	16.135	16.163	16.156	16.205
Adiac	**11.933**	**11.933**	**11.933**	**11.933**	**11.933**	**11.933**	**11.933**
ArrowHead	9.020	8.757	8.699	**8.687**	8.732	8.692	8.958
Beef	11.215	11.095	11.069	**11.061**	**11.061**	11.117	11.215
BeetleFly	9.946	9.618	**9.531**	9.592	9.619	9.591	10.368
BirdChicken	9.996	9.652	**9.374**	9.515	9.870	9.585	10.335
CBF	10.150	10.065	**10.005**	10.006	10.009	10.009	10.150
Car	9.392	9.290	9.067	**9.039**	9.059	9.046	9.276
ChlorineConcentration	15.512	15.214	15.182	**15.175**	15.176	15.176	15.331
CinC_ECG_torso	13.134	**12.837**	12.848	12.868	13.621	12.877	13.621
Coffee	7.150	6.893	6.692	**6.628**	6.693	6.651	6.825
Computers	**16.420**	16.498	16.489	16.502	16.960	16.475	16.960
Cricket_X	16.922	16.696	16.629	**16.628**	16.649	16.653	16.955
Cricket_Y	16.783	16.588	16.545	**16.533**	16.570	16.570	16.803
Cricket_Z	16.874	16.669	16.597	**16.593**	16.620	16.620	16.981
DiatomSizeReduction	5.959	5.907	5.889	**5.739**	5.798	5.758	5.932
DistalPhalanxOutlineAgeGroup	11.193	11.220	11.202	11.198	11.194	11.196	**11.158**
DistalPhalanxOutlineCorrect	12.467	12.373	**12.340**	12.342	12.494	12.350	12.483
DistalPhalanxTW	11.244	11.260	11.263	11.251	11.264	11.261	**11.222**
ECG200	11.395	11.317	11.323	**11.274**	11.300	11.289	11.501
ECG5000	16.169	**16.084**	16.142	16.136	16.137	16.136	16.211
ECGFiveDays	8.734	8.579	8.522	**8.513**	8.713	8.533	8.818
Earthquakes	14.757	14.727	**14.726**	14.728	14.757	**14.726**	14.757
ElectricDevices	22.404	22.428	22.401	22.398	22.630	22.399	**22.332**
FISH	10.841	10.740	10.594	**10.514**	10.560	10.566	10.841
FaceAll	16.272	16.187	16.185	16.183	16.197	**16.182**	16.291
FaceFour	10.422	10.318	**10.302**	10.316	10.575	10.321	10.533
FacesUCR	14.479	14.432	14.426	**14.423**	14.430	14.429	14.431
FordA	18.604	18.390	18.388	18.387	18.977	**18.385**	18.977
FordB	17.620	17.429	17.425	17.426	17.466	**17.416**	17.998
Gun_Point	10.242	10.019	9.843	9.743	9.883	**9.738**	10.130
Ham	12.772	12.545	12.488	**12.473**	13.240	12.506	12.957
MedicalImages	15.081	14.982	14.985	**14.979**	14.986	14.986	15.032
MiddlePhalanxOutlineAgeGroup	9.909	9.919	9.856	**9.818**	9.824	9.822	9.856
MiddlePhalanxOutlineCorrect	11.121	11.088	10.984	10.951	10.962	10.961	**10.923**
MiddlePhalanxTW	**10.514**	**10.514**	**10.514**	**10.514**	**10.514**	**10.514**	**10.514**
MoteStrain	9.560	9.484	9.460	9.451	**9.201**	9.470	9.557
NonInvasiveFatalECG_Thorax1	**17.728**	N/A	N/A	N/A	N/A	N/A	N/A
ProximalPhalanxTW	11.055	10.993	10.978	**10.958**	10.968	10.965	10.968
RefrigerationDevices	17.391	17.351	**17.311**	17.322	17.758	17.324	17.758
ScreenType	17.467	17.388	17.306	17.297	18.126	**17.289**	17.838
ShapeletSim	11.176	**10.896**	10.905	10.906	10.916	10.915	11.176
ShapesAll	17.539	17.405	**17.331**	17.333	17.605	17.357	17.509
SmallKitchenAppliances	17.551	17.611	**17.537**	17.606	18.007	17.561	18.007
SonyAIBORobotSurface	8.181	7.959	**7.934**	7.943	8.247	7.958	8.084
SonyAIBORobotSurfaceII	9.349	**9.265**	9.267	9.267	9.325	9.277	9.338
StarLightCurves	19.435	19.110	**19.012**	N/A	N/A	N/A	N/A
Trace	14.570	14.570	14.556	**14.550**	14.555	14.556	14.553
TwoLeadECG	6.939	6.939	6.892	6.879	6.936	6.892	**6.743**
Two_Patterns	17.416	17.379	17.317	17.325	17.524	**17.307**	17.524
UWaveGestureLibraryAll	18.911	18.641	**18.514**	18.531	19.282	18.537	19.244
Wine	7.527	7.297	6.482	6.358	6.390	6.353	**6.223**
WordsSynonyms	15.209	15.093	**15.024**	15.025	15.053	15.036	15.159
Worms	14.184	14.051	**13.889**	13.896	14.943	13.968	14.648
WormsTwoClass	13.727	13.471	**13.429**	13.462	14.944	13.494	14.944
synthetic_control	15.338	15.303	15.292	15.291	15.303	15.295	**15.278**
uWaveGestureLibrary_X	18.789	**18.568**	N/A	N/A	N/A	N/A	N/A

IPR2026-00272
Krisp Technologies EX1010 Page 331

J. k-means clustering: DTW loss achieved (Eq. (5) with $\gamma = 0$, log-scaled) when using Euclidean mean initialization

Dataset	Soft-DTW $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Subgradient method	DBA	Euclidean mean
50words	16.233	16.145	16.046	**16.035**	16.045	16.233	16.233
Adiac	12.311	12.311	12.264	12.241	12.234	**12.233**	12.311
ArrowHead	9.014	8.963	8.766	**8.746**	8.851	8.809	9.014
Beef	11.225	11.110	11.088	11.079	11.077	**11.070**	11.300
BeetleFly	9.895	9.290	9.268	**9.240**	9.512	10.926	10.926
BirdChicken	10.032	9.542	9.352	**9.338**	9.422	9.414	10.335
CBF	10.246	9.995	9.910	**9.908**	9.933	9.921	10.246
Car	9.276	9.229	8.989	**8.910**	8.936	8.935	9.276
ChlorineConcentration	15.331	15.291	15.254	**15.252**	15.270	**15.252**	15.331
CinC_ECG_torso	13.197	12.803	12.752	**12.728**	13.723	13.723	13.723
Coffee	6.825	6.825	6.668	6.599	6.605	**6.591**	6.825
Computers	16.417	16.346	16.301	**16.289**	17.167	16.342	17.167
Cricket_X	16.895	16.719	16.622	16.623	16.612	**16.600**	16.987
Cricket_Y	16.770	16.651	16.546	**16.514**	16.515	16.527	16.861
Cricket_Z	16.924	16.748	16.670	**16.633**	16.653	16.653	17.028
DiatomSizeReduction	5.963	5.963	5.963	**5.884**	5.897	5.896	5.963
DistalPhalanxOutlineAgeGroup	**11.164**	**11.164**	**11.164**	**11.164**	**11.164**	**11.164**	**11.164**
DistalPhalanxOutlineCorrect	12.544	12.533	12.494	12.475	**12.240**	12.479	12.577
DistalPhalanxTW	**11.242**	11.259	11.256	11.243	11.251	11.245	11.259
ECG200	11.462	11.291	11.239	**11.222**	11.231	11.234	11.503
ECG5000	16.253	16.180	16.171	16.183	**16.170**	16.172	16.262
ECGFiveDays	8.738	8.614	**8.543**	8.549	8.709	8.559	8.818
Earthquakes	15.113	14.625	14.601	14.599	15.952	**14.597**	15.952
ElectricDevices	22.295	22.325	22.291	22.290	22.379	**22.283**	22.379
FISH	10.904	10.843	10.589	10.543	**10.527**	10.555	10.904
FaceAll	16.278	16.162	16.145	16.145	16.152	**16.140**	16.347
FaceFour	10.376	10.273	10.239	**10.226**	10.566	10.241	10.566
FacesUCR	14.472	14.434	14.406	14.407	**14.391**	14.481	14.481
FordA	18.581	**18.354**	**18.354**	18.360	20.038	20.038	20.038
FordB	17.649	17.443	17.429	17.436	17.466	**17.427**	19.143
Gun_Point	10.334	10.027	9.806	**9.751**	9.902	9.833	10.334
Ham	12.805	12.603	12.559	**12.558**	12.974	12.561	12.974
HandOutlines	**13.712**	N/A	N/A	N/A	N/A	N/A	N/A
MedicalImages	15.082	14.963	**14.940**	14.942	14.950	14.941	15.091
MiddlePhalanxOutlineAgeGroup	9.856	9.856	9.855	**9.821**	9.823	**9.821**	9.856
MiddlePhalanxOutlineCorrect	10.962	10.962	10.962	10.959	**10.950**	**10.950**	10.962
MiddlePhalanxTW	**10.558**	10.587	10.587	10.569	10.572	10.570	10.587
MoteStrain	9.551	9.454	**9.413**	9.451	9.557	9.446	9.557
NonInvasiveFatalECG_Thorax1	**17.765**	N/A	N/A	N/A	N/A	N/A	N/A
ProximalPhalanxTW	10.978	**10.973**	10.978	10.978	10.978	10.978	10.978
RefrigerationDevices	17.260	17.202	17.093	**17.073**	18.140	17.095	18.140
ScreenType	17.430	17.359	17.294	**17.292**	17.838	17.323	17.838
ShapeletSim	11.497	10.864	**10.845**	10.853	10.865	11.608	11.608
ShapesAll	17.560	17.431	17.335	**17.328**	17.560	17.560	17.560
SmallKitchenAppliances	17.310	**17.273**	17.357	17.357	18.206	18.206	18.206
SonyAIBORobotSurface	8.084	7.980	**7.941**	7.947	8.084	7.948	8.084
SonyAIBORobotSurfaceII	9.338	9.207	9.196	**9.195**	9.338	**9.195**	9.338
StarLightCurves	19.457	19.178	**19.083**	N/A	N/A	N/A	N/A
Trace	14.553	14.553	14.553	**14.549**	14.553	14.553	14.553
TwoLeadECG	6.743	6.705	6.623	**6.606**	6.666	6.633	6.743
Two_Patterns	**17.084**	17.363	17.242	17.255	17.518	17.316	17.942
UWaveGestureLibraryAll	18.820	18.613	18.539	**18.488**	19.259	18.508	19.259
Wine	6.223	6.223	6.223	6.223	6.223	**6.205**	6.223
WordsSynonyms	15.184	15.036	**14.947**	14.951	14.965	14.959	15.196
Worms	14.043	13.860	13.791	13.777	14.696	**13.772**	14.696
WormsTwoClass	13.699	13.440	**13.322**	13.337	15.076	13.390	15.076
synthetic_control	15.472	15.367	15.337	15.338	**15.330**	15.336	15.472
uWaveGestureLibrary_X	18.844	**18.562**	N/A	N/A	N/A	N/A	N/A

IPR2026-00272
Krisp Technologies EX1010 Page 332

K. Time-series prediction: DTW loss achieved when using random init

Dataset	Soft-DTW loss $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Euclidean loss
50words	6.473	**4.921**	4.999	6.489	18.734
Adiac	0.094	**0.074**	0.078	0.109	0.103
ArrowHead	1.851	**1.708**	1.933	1.909	2.073
Beef	12.229	**8.688**	10.244	9.126	22.228
BeetleFly	35.037	25.439	27.588	**23.494**	50.610
BirdChicken	31.878	19.914	25.100	**14.981**	30.693
CBF	10.802	**9.263**	9.595	10.151	12.868
Car	1.724	2.307	2.202	**1.318**	1.588
ChlorineConcentration	7.876	2.108	2.331	1.735	**0.769**
CinC_ECG_torso	45.675	26.337	**23.567**	24.550	48.171
Coffee	0.914	0.727	1.662	1.883	**0.660**
Computers	92.584	84.723	78.953	**75.435**	235.208
Cricket_X	9.394	8.042	**7.123**	7.226	12.080
Cricket_Y	11.989	9.643	**9.534**	9.545	15.002
Cricket_Z	9.161	6.889	**6.585**	7.200	11.003
DiatomSizeReduction	1.182	0.922	**0.820**	0.897	1.203
DistalPhalanxOutlineAgeGroup	0.426	0.291	0.541	0.496	**0.231**
DistalPhalanxOutlineCorrect	0.494	0.476	0.564	0.591	**0.351**
DistalPhalanxTW	0.441	0.330	0.305	1.214	**0.231**
ECG200	1.874	**1.716**	1.884	1.734	1.905
ECG5000	4.895	4.705	4.543	**4.441**	5.463
ECGFiveDays	1.834	1.944	1.699	**1.642**	2.220
Earthquakes	74.738	59.973	60.877	**57.827**	147.980
ElectricDevices	20.186	**15.125**	15.218	15.287	37.121
FISH	0.464	0.429	**0.354**	0.459	0.462
FaceAll	9.317	7.451	7.902	**7.276**	10.716
FaceFour	**19.564**	20.881	28.150	28.839	46.841
FacesUCR	15.359	**14.643**	16.143	17.428	28.576
Gun_Point	0.896	**0.805**	0.923	0.834	0.858
Ham	20.154	17.931	17.786	**17.413**	24.340
Haptics	**16.174**	17.775	17.142	17.423	23.130
Herring	1.000	0.712	**0.666**	0.762	0.865
InsectWingbeatSound	3.460	2.823	2.458	**2.220**	5.437
ItalyPowerDemand	0.911	0.893	**0.711**	0.798	0.881
LargeKitchenAppliances	63.153	60.739	**60.157**	61.841	266.853
Lighting2	73.293	66.341	**65.335**	66.881	147.668
Lighting7	44.446	42.699	**40.608**	41.502	68.902
Meat	0.162	0.242	0.246	0.650	**0.099**
MedicalImages	1.023	0.853	**0.708**	0.778	1.211
MiddlePhalanxOutlineAgeGroup	0.343	0.347	0.570	0.400	**0.312**
MiddlePhalanxOutlineCorrect	0.278	0.204	0.227	0.202	**0.182**
MiddlePhalanxTW	0.251	0.153	0.445	0.314	**0.132**
MoteStrain	10.188	**9.986**	11.119	10.250	11.183
NonInvasiveFatalECG_Thorax1	1.002	0.920	0.675	**0.634**	1.219
OSULeaf	15.125	11.722	11.086	**10.775**	30.739
OliveOil	0.476	0.683	2.082	2.076	**0.020**
PhalangesOutlinesCorrect	0.352	0.216	0.352	0.338	**0.170**
Phoneme	160.536	150.017	148.175	**145.093**	219.704
Plane	0.619	**0.564**	0.834	0.788	0.630
ProximalPhalanxOutlineAgeGroup	0.134	0.062	0.105	0.118	**0.046**
ProximalPhalanxOutlineCorrect	0.129	0.047	0.089	0.128	**0.044**
ProximalPhalanxTW	0.154	0.077	0.102	0.150	**0.055**
RefrigerationDevices	108.421	93.519	**89.370**	89.873	160.361
ShapeletSim	102.413	**70.455**	71.156	72.094	108.936
ShapesAll	10.391	9.027	7.850	**7.207**	18.348
SonyAIBORobotSurface	4.453	4.494	**4.318**	4.910	4.388
SonyAIBORobotSurfaceII	8.072	8.302	**7.758**	8.669	8.628
Strawberry	0.123	0.088	0.137	0.100	**0.081**
SwedishLeaf	1.486	1.277	1.316	**1.169**	1.633
Symbols	17.963	14.039	15.172	**13.192**	38.268
ToeSegmentation1	23.866	22.987	26.056	**22.401**	35.806
ToeSegmentation2	41.450	33.100	**30.931**	31.106	61.899
Trace	0.563	0.379	0.352	**0.279**	0.582
TwoLeadECG	0.441	0.394	**0.318**	0.320	0.336
Two_Patterns	15.035	10.100	10.588	**8.584**	35.923
UWaveGestureLibraryAll	40.324	28.975	26.193	**25.897**	93.019
Wine	0.164	0.203	1.417	0.958	**0.028**
WordsSynonyms	12.466	10.437	**9.165**	9.219	32.003
Worms	81.236	63.938	60.950	**59.995**	114.528
WormsTwoClass	78.455	66.609	**60.207**	61.685	122.619
synthetic_control	7.709	**5.315**	5.390	5.506	7.690
uWaveGestureLibrary_X	13.096	9.995	10.143	**9.433**	19.995
uWaveGestureLibrary_Y	9.793	7.272	7.327	**7.225**	17.706
uWaveGestureLibrary_Z	11.883	8.909	8.494	**8.416**	20.092
wafer	1.049	0.473	**0.386**	0.496	2.636
yoga	2.932	2.431	**1.995**	3.309	4.305

IPR2026-00272
Krisp Technologies EX1010 Page 333

L. Time-series prediction: DTW loss achieved when using Euclidean init

Dataset	Soft-DTW loss $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Euclidean loss
50words	6.330	5.628	4.885	**4.553**	18.734
Adiac	0.082	0.076	**0.064**	0.079	0.103
ArrowHead	1.823	2.016	**1.762**	2.106	2.073
Beef	7.250	6.940	7.146	**3.757**	22.228
BeetleFly	32.430	**26.600**	27.199	29.003	50.610
BirdChicken	24.952	22.600	**19.914**	20.540	30.693
CBF	10.744	8.978	9.215	**8.398**	12.868
Car	0.906	0.812	**0.709**	0.740	1.588
ChlorineConcentration	6.018	0.979	**0.695**	0.698	0.769
CinC_ECG_torso	29.892	**18.638**	19.635	19.191	48.171
Coffee	0.870	0.582	0.511	**0.496**	0.660
Computers	86.619	**79.250**	82.215	81.417	235.208
Cricket_X	10.954	8.200	**7.932**	8.296	12.080
Cricket_Y	11.901	10.150	10.265	**9.574**	15.002
Cricket_Z	9.714	7.760	**7.544**	8.041	11.003
DiatomSizeReduction	0.964	**0.852**	0.874	0.869	1.203
DistalPhalanxOutlineAgeGroup	0.403	0.206	**0.175**	0.177	0.231
DistalPhalanxOutlineCorrect	0.515	0.310	0.300	**0.262**	0.351
DistalPhalanxTW	0.468	0.228	0.186	**0.178**	0.231
ECG200	1.907	1.541	1.565	**1.536**	1.905
ECG5000	4.737	4.190	4.398	**4.148**	5.463
ECGFiveDays	1.584	1.396	**1.322**	1.335	2.220
Earthquakes	71.461	**55.819**	56.504	57.153	147.980
ElectricDevices	19.499	15.045	**14.999**	15.228	37.121
FISH	0.439	0.353	0.319	**0.318**	0.462
FaceAll	9.309	8.687	**7.803**	7.853	10.716
FaceFour	20.483	**20.411**	21.259	21.444	46.841
FacesUCR	14.984	14.530	**14.403**	14.729	28.576
Gun_Point	0.447	0.368	0.300	**0.297**	0.858
Ham	16.152	14.717	**12.252**	13.424	24.340
Haptics	15.177	14.275	12.394	**11.931**	23.130
Herring	0.310	0.305	0.292	**0.249**	0.865
InsectWingbeatSound	3.104	2.346	2.186	**2.036**	5.437
ItalyPowerDemand	0.802	**0.595**	0.623	0.654	0.881
LargeKitchenAppliances	61.531	63.834	59.116	**57.219**	266.853
Lighting2	65.602	62.240	61.561	**60.826**	147.668
Lighting7	43.930	41.668	40.535	**39.422**	68.902
Meat	0.173	0.140	0.103	**0.077**	0.099
MedicalImages	0.932	0.671	**0.615**	0.651	1.211
MiddlePhalanxOutlineAgeGroup	0.283	0.200	**0.134**	0.154	0.312
MiddlePhalanxOutlineCorrect	0.269	0.169	**0.133**	0.148	0.182
MiddlePhalanxTW	0.225	0.126	0.111	**0.094**	0.132
MoteStrain	9.704	**9.321**	9.385	10.156	11.183
NonInvasiveFatalECG_Thorax1	0.921	0.736	0.540	**0.536**	1.219
OSULeaf	16.168	14.372	13.354	**12.350**	30.739
OliveOil	0.179	0.298	0.034	0.026	**0.020**
PhalangesOutlinesCorrect	0.328	0.231	0.161	**0.153**	0.170
Phoneme	173.501	**158.036**	159.293	158.776	219.704
Plane	0.454	0.343	**0.311**	0.370	0.630
ProximalPhalanxOutlineAgeGroup	0.137	0.056	0.039	**0.036**	0.046
ProximalPhalanxOutlineCorrect	0.118	0.046	0.033	**0.030**	0.044
ProximalPhalanxTW	0.164	0.067	0.054	**0.045**	0.055
RefrigerationDevices	107.693	95.458	**90.328**	91.687	160.361
ShapeletSim	106.805	73.626	**72.985**	73.987	108.936
ShapesAll	11.946	7.724	**7.127**	7.522	18.348
SonyAIBORobotSurface	4.068	3.619	**3.400**	3.737	4.388
SonyAIBORobotSurfaceII	6.954	6.585	**6.410**	6.593	8.628
Strawberry	0.110	0.082	0.071	**0.069**	0.081
SwedishLeaf	1.438	1.151	**1.095**	1.108	1.633
Symbols	14.717	**12.040**	15.229	16.199	38.268
ToeSegmentation1	24.293	**20.905**	22.914	22.566	35.806
ToeSegmentation2	44.439	36.333	**35.804**	41.121	61.899
Trace	0.578	0.331	0.272	**0.250**	0.582
TwoLeadECG	0.157	0.131	**0.129**	0.149	0.336
Two_Patterns	14.843	11.616	11.622	**11.059**	35.923
UWaveGestureLibraryAll	42.336	30.864	27.572	**26.573**	93.019
Wine	0.069	0.263	0.045	0.029	**0.028**
WordsSynonyms	12.654	10.089	9.887	**8.946**	32.003
Worms	74.589	71.946	70.245	**67.669**	114.528
WormsTwoClass	81.311	**61.360**	62.281	74.672	122.619
synthetic_control	7.455	5.509	5.374	**5.369**	7.690
uWaveGestureLibrary_X	14.151	10.065	9.231	**9.197**	19.995
uWaveGestureLibrary_Y	9.852	7.285	7.066	**7.036**	17.706
uWaveGestureLibrary_Z	12.019	8.893	**8.453**	8.513	20.092
wafer	1.125	0.838	**0.765**	0.818	2.636
yoga	2.510	2.209	**1.807**	1.851	4.305

IPR2026-00272
Krisp Technologies EX1010 Page 334

Appendix CUTURI05

```
LEADER 00000nam  2200337I  4500
001    on1029804631
003    OCoLC
005    20180720013834.0
008    180329s2017    paua     b    101 0 eng d
019    1029847828|a1029874219|a1029889911
020    1510855149
020    9781510855144
035    (OCoLC)1029804631|z(OCoLC)1029847828|z(OCoLC)1029874219
       |z(OCoLC)1029889911
040    YDX|beng|cYDX|dLTSCA|dUOH
049    UOHA
050  4 Q325.5|b.M34 2017
090    Q325.5 .I.58 2017
111 2  International Conference on Machine Learning|n(34th :
       |d2017 :|cSydney, N.S.W.)
245 10 34th International Conference on Machine Learning (ICML
       2017) :|bSydney, Australia 6-11 August 2017
264  1 Stroudsburg, Pennsylvania :|bInternational Machine
       Learning Society (IMLS),|c[2017]
264  4 |c©2017
300    8 volumes :|billustrations,|c28 cm
336    text|btxt|2rdacontent
337    unmediated|bn|2rdamedia
338    volume|bnc|2rdacarrier
504    Includes bibliographical references and index
650  0 Machine learning|vCongresses
710 2  International Machine Learning Society,|eissuing body
850    UOH
902    LTS
915    Holding set with OCLC SHR March 29, 2023
```

LOCATION	CALL #	STATUS
Science Library-Lower Level	Q325.5 .I.58 2017 v.1	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.2	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.3	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.4	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.5	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.6	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.7	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.8	IN LIBRARY

© University of Saskatchewan
Disclaimer | Privacy

Ask Us

Printed Material

Author International Conference on Machine Learning (34th : 2017 : Sydney, N.S.W.)

Title **34th International Conference on Machine Learning (ICML 2017) : Sydney, Australia 6-11 August 2017**

Published Stroudsburg, Pennsylvania : International Machine Learning Society (IMLS), [2017]

©2017

LOCATION	CALL #	STATUS
Science Library-Lower Level	Q325.5 .I.58 2017 v.1	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.2	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.3	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.4	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.5	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.6	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.7	IN LIBRARY
Science Library-Lower Level	Q325.5 .I.58 2017 v.8	IN LIBRARY

Send via Text Message / Email

Description 8 volumes : illustrations, 28 cm

Subject Machine learning -- Congresses

Bibliog. Includes bibliographical references and index

Other Author International Machine Learning Society, issuing body

ISBN 1510855149

9781510855144

Permanent Link to This Record

Back to Top

Record: Previous Record Next Record

© University of Saskatchewan
Disclaimer | Privacy

Ask Us

Appendix
KAMRUZZAMAN01

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

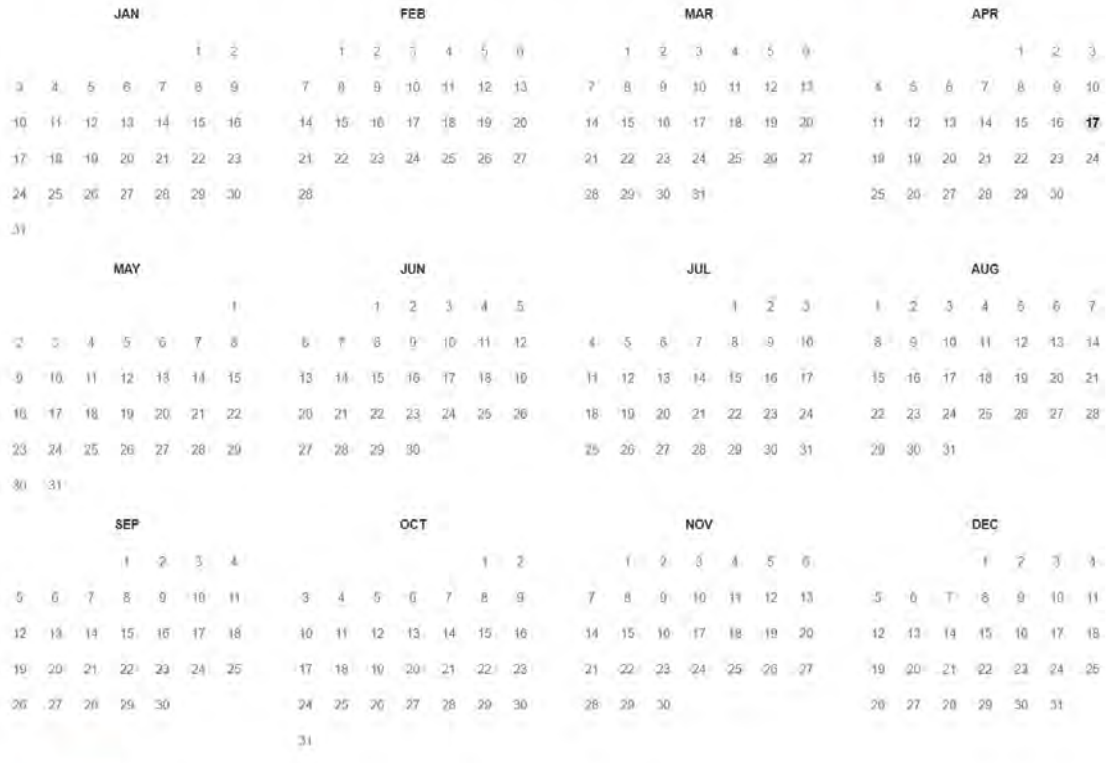
DONATE INTERNET ARCHIVE WayBackMachine Explore more than 1 trillion web pages saved over time

https://arxiv.org/abs/1009.4972

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **4 times** between April 17, 2021 and March 17, 2025.





Note

This calendar view maps the number of times **https://arxiv.org/abs/1009.4972** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20210417200127/https://arxiv.org/abs/1009.4972 January 21, 2026

INTERNET ARCHIVE WaybackMachine https://arxiv.org/abs/1009.4972 Go MAR APR 17 2021 APR 2024 About this capture
4 captures 17 Apr 2021 - 17 Mar 2025 2020

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > cs > arXiv:1009.4972

Search... All fields Search
Help | Advanced Search

Computer Science > Machine Learning

[Submitted on 25 Sep 2010]

Speaker Identification using MFCC-Domain Support Vector Machine

S. M. Kamruzzaman, A. N. M. Rezaul Karim, Md. Saiful Islam, Md. Emdadul Haque

Speech recognition and speaker identification are important for authentication and verification in security purpose, but they are difficult to achieve. Speaker identification methods can be divided into text-independent and text-dependent. This paper presents a technique of text-dependent speaker identification using MFCC-domain support vector machine (SVM). In this work, melfrequency cepstrum coefficients (MFCCs) and their statistical distribution properties are used as features, which will be inputs to the neural network. This work firstly used sequential minimum optimization (SMO) learning technique for SVM that improve performance over traditional techniques Chunking, Osuna. The cepstrum coefficients representing the speaker characteristics of a speech segment are computed by nonlinear filter bank analysis and discrete cosine transform. The speaker identification ability and convergence speed of the SVMs are investigated for different combinations of features. Extensive experimental results on several samples show the effectiveness of the proposed approach.

Comments: 5 Pages, International Journal
Subjects: **Machine Learning (cs.LG)**; Sound (cs.SD)
Journal reference: International Journal of Electrical and Power Engineering, Vol. 1, No. 3, pp. 274-278, 2007
DOI: 10.3923/ijepe.2007.274.278
Cite as: arXiv:1009.4972 **[cs.LG]**
(or arXiv:1009.4972v1 **[cs.LG]** for this version)

Submission history

From: S. M. Kamruzzaman [view email]
[v1] Sat, 25 Sep 2010 05:32:44 UTC (52 KB)

Download:
- PDF only
(license)

Current browse context:
cs.LG
< prev | next >
new | recent | 1009
Change to browse by:
cs
cs.SD

References & Citations
- NASA ADS
- Google Scholar
- Semantic Scholar

DBLP - CS Bibliography
listing | bibtex
S. M. Kamruzzaman
A. N. M. Rezaul Karim
Md. Saiful Islam
Md. Emdadul Haque

Export Bibtex Citation

Bookmark

Bibliographic Tools | Code | Related Papers | About arXivLabs

Bibliographic and Citation Tools

Bibliographic Explorer (What is the Explorer?)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

About | Help | Contact | Subscribe | Copyright | Privacy Policy | Web Accessibility Assistance | arXiv Operational Status
Get status notifications via email or slack

Appendix
KAMRUZZAMAN02

https://web.archive.org/web/20200801000000*/https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf January 21, 2026

INTERNET ARCHIVE WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES SIGN UP | LOG IN UPLOAD Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

DONATE

INTERNET ARCHIVE WayBackMachine

Explore more than 1 trillion web pages saved over time

https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf

Calendar · Collections · Changes · Summary · Site Map · URLs

Saved **6 times** between August 25, 2020 and April 21, 2024.

2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 **2020** 2021 2022 2023 2024 2025 2026

JAN FEB MAR APR

MAY JUN JUL AUG

SEP OCT NOV DEC

Note

This calendar view maps the number of times **https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf** was crawled by the Wayback Machine, *not* how many times the site was actually updated. More info in the FAQ.

FAQ | Contact Us | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

https://web.archive.org/web/20200825105542/https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf January 21, 2026
INTERNET ARCHIVE WaybackMachine https://arxiv.org/ftp/arxiv/papers/1009/1009.4972.pdf Go JUL AUG JUL
6 captures 25 Aug 2020 - 21 Apr 2024 2019 25 2020 2022 About this capture

Speaker Identification using MFCC-Domain Support Vector Machine

S. M. Kamruzzaman[1], A. N. M. Rezaul Karim[2], Md. Saiful Islam[3] and Md. Emdadul Haque[1]

[1]Department of Information and Communication Engineering
[3]Department of Computer Science and Engineering
University of Rajshahi, Rajshahi-6205, Bangladesh
[2]Department of Computer Science and Engineering
International Islamic University Chittagong, Bangladesh
Email: smzaman@gmail.com

Abstract: Speech recognition and speaker identification are important for authentication and verification in security purpose, but they are difficult to achieve. Speaker identification methods can be divided into text-independent and text-dependent. This paper presents a technique of text-dependent speaker identification using MFCC-domain support vector machine (SVM). In this work, melfrequency cepstrum coefficients (MFCCs) and their statistical distribution properties are used as features, which will be inputs to the neural network. This work firstly used sequential minimum optimization (SMO) learning technique for SVM that improve performance over traditional techniques Chunking, Osuna. The cepstrum coefficients representing the speaker characteristics of a speech segment are computed by nonlinear filter bank analysis and discrete cosine transform. The speaker identification ability and convergence speed of the SVMs are investigated for different combinations of features. Extensive experimental results on several samples show the effectiveness of the proposed approach.

Speaker Identification using MFCC-Domain Support Vector Machine

S. M. Kamruzzaman[1], A. N. M. Rezaul Karim[2], Md. Saiful Islam[3] and Md. Emdadul Haque[1]

[1]Department of Information and Communication Engineering
[3]Department of Computer Science and Engineering
University of Rajshahi, Rajshahi-6205, Bangladesh
[2]Department of Computer Science and Engineering
International Islamic University Chittagong, Bangladesh
Email: smzaman@gmail.com

Abstract: Speech recognition and speaker identification are important for authentication and verification in security purpose, but they are difficult to achieve. Speaker identification methods can be divided into text-independent and text-dependent. This paper presents a technique of text-dependent speaker identification using MFCC-domain support vector machine (SVM). In this work, melfrequency cepstrum coefficients (MFCCs) and their statistical distribution properties are used as features, which will be inputs to the neural network. This work firstly used sequential minimum optimization (SMO) learning technique for SVM that improve performance over traditional techniques Chunking, Osuna. The cepstrum coefficients representing the speaker characteristics of a speech segment are computed by nonlinear filter bank analysis and discrete cosine transform. The speaker identification ability and convergence speed of the SVMs are investigated for different combinations of features. Extensive experimental results on several samples show the effectiveness of the proposed approach.

Keywords: Speech recognition, speaker identification, MFCC, support vector machine, neural networks, chunking, osuna, discrete cosine transform.

1. Introduction

Speaker identification has been the subject of active research for many years, and has many potential applications where propriety of information is a concern. Speaker identification is the process of automatically recognizing a speaker by machine using the speaker's voice [11]. The most common application of speaker identification systems is in access control, for example, access to a room or privileged information over the telephone [14]. Also it has a very useful usage for speaker adaptation in automatic speech recognition system.

Speaker recognition can be classified into identification and verification [13]. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Speaker recognition methods can also be divided into text-independent and text-dependent methods [9]. In a text-independent system, speaker models capture characteristics of somebody's speech, which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc [2]. The choice of which technology to use is application-specific

The objective of this work is to design an efficient system for human speech recognition that is able to identify and verify human speech more accurately. This work presents a technique of text-dependent speaker identification using MFCC-domain support vector machine (SVM). Mel-frequency cepstrum coefficients (MFCCs) and their statistical distribution properties are used as features, which will be inputs to the neural network [8].

2. Voice processing

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate). The speech signal is a slowly time varying signal (it is called quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 0.2s or more) the signal characteristics change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as linear prediction coding (LPC), mel-frequency cepstrum coefficients

Corresponding Author: S. M. Kamruzzaman, Department of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh, Email: smzaman@gmail.com

IPR2026-00272
Krisp Technologies EX1010 Page 346

(MFCC), and others. MFCC is perhaps the best known and most popular, and this feature has been used in this paper. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [1] [4].

2.1. The MFCC processor

A block diagram of the structure of an MFCC processor is given in Figure 1. The speech input is recorded at a sampling rate of 22050Hz. This sampling frequency is chosen to minimize the effects of *aliasing* in the analog-to-digital conversion process.

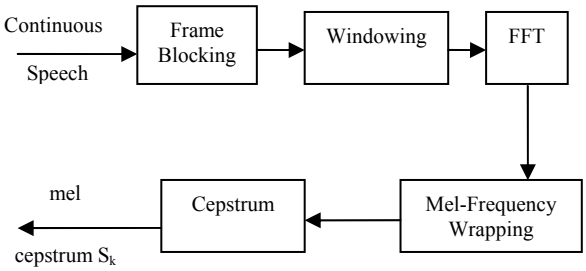


Figure 1: Block diagram of the MFCC processor.

2.2. Mel-frequency wrapping

The speech signal consists of tones with different frequencies. For each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following formula to compute the mels for a given frequency *f* in Hz:

mel(f)= 2595*log10(1+f/700) …….. (1)

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired melfrequency component. The filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

2.3. Cepstrum

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the discrete cosine transform (DCT). The MFCCs may be calculated using the following equation [3] [5]:

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k)\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \quad \text{.....(2)}$$

where n=1, 2, …, K. The number of mel cepstrum coefficients, K, is typically chosen as 20. The first component, $\tilde{c} \sim 0$, is excluded from the DCT since it represents the mean value of the input signal which carries little speaker specific information. By applying the procedure described above, for each speech frame of about 30 ms with overlap, a set of mel-frequency cepstrum coefficients is computed. This set of coefficients is called an *acoustic vecto*r. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker [7]. Therefore each input utterance is transformed into a sequence of acoustic vectors. The next section describes how these acoustic vectors can be used to represent and recognize the voice characteristic of a speaker.

3. Feature matching

The state-of-the-art feature matching techniques used in speaker recognition include, NETtalk, Time-Delay Neural Network (TDNN), Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). A new approach, Support Vector Machine is used here, due to high accuracy.

3.1. Support vector machine

The support vector machine (SVM) is a linear machine pioneered by Vapnik [12]. Like multiplayer perceptron and radial basis function networks SVM can be used for pattern classification and nonlinear regression. The SVM become popular due to its attractive features and promising empirical performance. The idea of support vector machine is to construct a hyperplane as the decision surface in such a way that that the margin of separation between positive and negative example maximized. The SVM is an approximate implementation of the method of structural risk minimization (SRM). SRM principles have been shown to be superior to empirical risk minimization (ERM) principle employed by conventional neural networks. SRM

minimize an upper bound on the expected risk as opposed to ERM that minimizes the error on the training data [6].

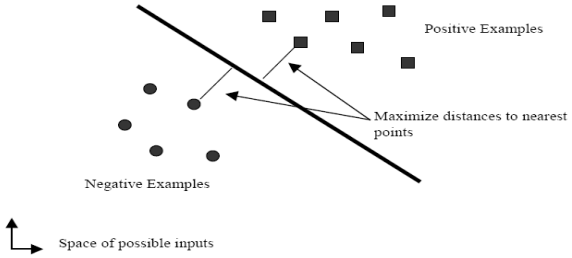


Figure 2: A linear support vector machine**.**

3.2. SVM training process

Though a solution for the quadratic optimization is guaranteed based on the KKT conditions, the number of computation required can be very high depending on the separability of the data and the number of training data points. Due to the size, the quadratic optimization problem that arises from SVMs cannot be easily solved via standard QP techniques. The quadratic form involves a matrix that has a number of elements equal to the square of the number of training examples. This matrix cannot be fit into 128 Megabytes if there are more than 4000 training examples. Vapnik [15] describes a method to solve the SVM quadratic problem, which has since been known as "chunking." The chunking algorithm uses the fact that the value of the quadratic form is the same if you remove the rows and columns of the matrix that corresponds to zero Lagrange multipliers. Therefore, the large quadratic optimization problem can be broken down into a series of smaller quadratic problems, whose ultimate goal is to identify all of the non-zero Lagrange multipliers and discard all of the zero Lagrange multipliers. At every step, chunking solves a quadratic optimization problem that consists of the following examples: every non-zero Lagrange multiplier from the last step, and the M worst examples that violate the KKT conditions [4], for some value of M. If there are fewer than M examples that violate the KKT conditions at a step, all of the violating examples are added in. Each quadratic sub-problem is initialized with the results of the previous sub-problem. At the last step, the entire set of non-zero Lagrange multipliers has been identified, hence the last step solves the large problem. Chunking seriously reduces the size of the matrix from the number of training examples squared to approximately the number of non-zero Lagrange multipliers squared. However, chunking still cannot handle large-scale training problems, since even this reduced matrix cannot fit into memory.

In 1997, Osuna, et al. [10] proved a theorem which suggests a whole new set of quadratic algorithms for SVMs. The theorem proves that the large quadratic optimization problem can be broken down into a series of smaller quadratic sub-problems. As long as at least one example that violates the KKT conditions is added to the examples for the previous sub-problem, each step will reduce the overall objective function and maintain a feasible point that obeys all of the constraints. Therefore, as sequence of quadratic sub-problems that always add at least one violator will be guaranteed to converge. Notice that the chunking algorithm obeys the conditions of the theorem, and hence will converge.

Osuna, et al. suggests keeping a constant size matrix for every quadratic sub-problem, which implies adding and deleting the same number of examples at every step [10]. Using a constant-size matrix will allow the training on arbitrarily sized data sets. The algorithm given in Osuna's paper [10] suggests adding one example and subtracting one example every step. Clearly this would be inefficient, because it would use an entire numerical quadratic optimization step to cause one training example to obey the KKT conditions. In practice, researchers add and subtract multiple examples according to unpublished heuristics [10]. In any event, a numerical QP solver is required for all of these methods. Numerical quadratic optimization is notoriously tricky to get right; there are many numerical precision issues that need to be addressed.

For each method, three steps are illustrated. The horizontal thin line at every step represents the training set, while the thick boxes represent the Lagrange multipliers being optimized at that step. For chunking, a fixed number of examples are added every step, while the zero Lagrange multipliers are discarded at every step.

Thus, the number of examples trained per step tends to grow. For Osuna's algorithm, a fixed number of examples are optimized every step: the same number of examples is added to and discarded from the problem at every step. For SMO, only two examples are analytically optimized at every step, so that each step is very fast.

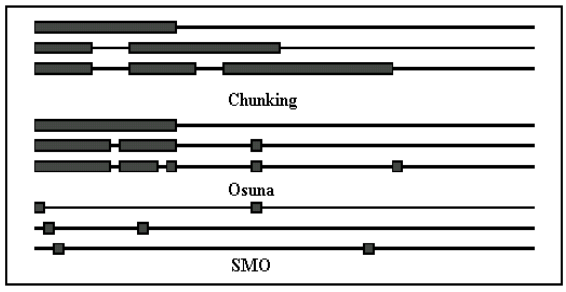


Figure 3: Three alternative methods for training SVMs: Chunking, Osuna's algorithm, and SMO.

3.3. Sequential minimal optimization (SMO)

Sequential minimal optimization (SMO) is a simple algorithm that can quickly solve the SVM problem without any extra matrix storage and without using numerical quadratic optimization steps at all. SMO decomposes the overall quadratic problem into quadratic sub-problems, using Osuna's theorem to ensure convergence.

Unlike the previous methods, SMO chooses to solve the smallest possible optimization problem at every step. For the standard SVM quadratic problem, the smallest possible optimization problem involves two Lagrange multipliers, because the Lagrange multipliers must obey a linear Equality constraint. At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values. The advantage of SMO lies in the fact that solving for two Lagrange multipliers can be done analytically. Thus, numerical quadratic optimization is avoided entirely. The inner loop of the algorithm can be expressed in a short amount of C code, rather than invoking an entire quadratic library routine. Even though more optimization sub-problems are solved in the course of the algorithm, each sub-problem is so fast that the overall quadratic problem is solved quickly.

In addition, SMO requires no extra matrix storage at all. Thus, very large SVM training problem scan fit inside of the memory of an ordinary personal computer or workstation. Because no matrix algorithms are used in SMO, it is less susceptible to numerical precision problems. There are two components to SMO an analytic method for solving for the two Lagrange multipliers, and a heuristic for choosing which multipliers to optimize.

4. Performance measure of SMO

The data set used to test SMO's speed was the UCI "adult" data set, which is available at [16]. There are 9888 examples in the "adult" training set. Two different SVMs were trained on the problem: a linear SVM, and a non-linear SVM. The timings for SMO versus chunking and Osuna for a linear SVM and non-linear SVM are shown in the table below:

Table 1: Timing of SMO versus chunking for a linear SVM and non-linear SVM.

Training Set Size	For Linear SVM		For Non-Linear SVM	
	SMO	Chunking	SMO	Chunking
2477	2.2	13.1	26.3	64.9
3470	4.9	16.1	44.1	110.4
4912	8.1	40.6	83.6	372.5
7366	12.7	140.7	156.7	545.5
9888	24.7	239.3	248.1	907.6

For the linear SVM on this data set, the SMO training time faster than Chunking. This experiment is another situation where SMO is superior to Osuna and chunking in computation time. In this case, the scaling for SMO is somewhat better than chunking: SMO is a factor of between two and six times faster than chunking. The non-linear test shows that SMO is still faster than chunking. For the real-world test sets, SMO can be a factor of 1200 times faster for linear SVMs and a factor of 15 times faster for non-linear SVMs. Because of its ease of use and better scaling with training set size, SMO is a strong candidate for becoming the standard SVM training algorithm.

5. Experiment results

To show the effectiveness of the proposed approach some experimental results are shown. The experiment has done by testing the system in different ways and then comparing their result to estimate the final result. Here we used the sample of word "Zero" 20 times for each speaker. We work on total 8 speakers. All the data samples are collected from an international database. Our experiment results are given below by using Chunking and SMO training algorithm in SVM. Now, we will see how the system acts when using Chunking training algorithm in SVM.

Table 2: Results obtain from using Chunking training algorithm.

Speaker ID	Total Sample	No. of correct samples	No. of false samples	Success rate in percentage (%)
1	20	16	4	75
2	20	20	0	90
3	20	20	0	90
4	20	17	3	80
5	20	20	0	90
6	20	16	4	75
7	20	18	2	90
8	20	20	0	90

Now, we will see how the system acts when using sequential minimum optimization (SMO) training algorithm in SVM.

Table 3: Results obtain from using SMO training algorithm.

Speaker ID	Total Sample	No. of correct samples	No. of false samples	Success rate in percentage (%)
1	20	18	2	90
2	20	20	0	100
3	20	20	0	100
4	20	19	1	95
5	20	20	0	100
6	20	18	2	90
7	20	18	2	90
8	20	19	1	95

IPR2026-00272
Krisp Technologies EX1010 Page 349

Here is the final result of our proposed approach

Table 4: Success rate in both SVM training methods.

Training algorithm	No. of correct samples	No. of false samples	Success rate in percentage (%)
Chunking	147	13	91.88
SMO	152	8	95

6. Conclusions

This paper successfully presents an approach based on neural networks for voice recognition and user identification. This approach based on MFCC-domain support vector machine by using SMO learning method. We have found a good result after testing the voice dependent system where there is a 91.88% success rate for using Chunking SVM training method and 95% for using SMO SVM training method. Thus we can conclude that, our approach have identified and verified human voice better than previous approaches.

The MFCC technique has been applied for speaker identification. VQ is used to minimize the data of the extracted feature. The study reveals that as number of centroids increases, identification rate of the system increases. It has been found that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has to be increased as the number of speakers increases. The study shows that the linear scale can also have a reasonable identification rate if a comparatively higher number of centroid is used. However, the recognition rate using a linear scale would be much lower if the number of speakers increases. Mel scale is also less vulnerable to the changes of speaker's vocal cord in course of time. The present study is still ongoing, which may include following further works. HMM may be used to improve the efficiency and precision of the segmentation to deal with crosstalk, laughter and uncharacteristic speech sounds. A more effective normalization algorithm can be adopted on extracted parametric representations of the acoustic signal, which would improve the identification rate further. Finally, a combination of features (MFCC, LPC, LPCC, Formant etc) may be used to implement a robust parametric representation for speaker identification.

References

[1] Lawrence Rabiner and Biing-Hwang Juang, Fundamental of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[2] Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu., Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification, in IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, January 1999.

[3] F. Soong, E. Rosenberg, B. Juang, and L. Rabiner, A Vector Quantization Approach to Speaker Recognition, AT&T Technical Journal, vol. 66, pp. 14-26, March/April 1987.

[4] Comp.speech Frequently Asked Questions WWW site, http://svr-www.eng.cam.ac.uk/comp.speech.

[5] Jr., J. D., Hansen, J., and Proakis, J., Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.

[6] Aravind Ganapathiraju, Support Vector Machines for Speech Recognition, Ph. D dissertation, Mississippi State University, Mississippi, January 2002.

[7] Burges, C. J. C., A Tutorial on Support Vector Machines for Pattern Recognition, submitted to Data Mining and Knowledge Discovery, http://svm.research.bell-labs.com/SVMdoc.html, 1998.

[8] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambridge, Massachusetts, USA, 1997.

[9] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1993.

[10] Osuna, E., Freund, R., Girosi, F., Improved Training Algorithm for Support Vector Machines, Proc. IEEE NNSP '97, 1997.

[11] Rabiner, L.R., A tutorial on Hidden Markov Models, Proceeding of IEEE, Vol.73, pp.1349-1387, 1989.

[12] Simon Haykin, McMaster University, Hamilton, Ontario, Canada, Neural Networks a Comprehensive Foundation, 2nd edition, pp.256-347.

[13] Squires, B and Sammut, Automatic Speaker Recognition: An application for Machine learning, Twelfth International Conference on Machine Learning, 1999.

[14] U. Bodenhausen, S. Manke, and A. Waibel, Connectionist Architectural Learning for High Performance Character and Speech Recognition, Proceedings of the International Conference on Acoustics Speech and Signal Processing, vol. 1, pp. 625-628, Minneapolis, MN, USA, April 1993.

[15] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer-Verlag, 1982.

[16] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult.

Appendix
KAMRUZZAMAN03

```
leader  01163nas a2200373 c 4500
001     991000000000525695
005     20260218111507.0
007     cr|||||||||||
008     070920c20079999pk u||p|o ||| 0||||1eng c
016     7#$2DE-101 $a985766905
016     7#$2DE-600 $a2388435-6
022     ##$a1993-6001
035     ##$a(CKB)1000000000525695
035     ##$a(OCoLC)643701617
035     ##$a(DE-101)985766905
035     ##$a(DE-599)ZDB2388435-6
035     ##$a(DE-599)2388435-6
040     ##$a8999 $bger $cDE-101 $d9999
041     ##$aeng
044     ##$cXB-PK
082     74$a620 $qDE-600 $222sdnb
245     00$aInternational journal of electrical and power engineering
246     ##$aInternational Journal of Electrical and Power Engineering
264     31$aFaisalabad $bMedWell $c2007-
300     ##$aOnline-Ressource
336     ##$aText $btxt $2rdacontent
337     ##$aComputermedien $bc $2rdamedia
338     ##$aOnline-Ressource $bcr $2rdacarrier
362     0#$a1.2007 -
363     01$81.1\x $a1 $i2007
906     ##$aJOURNAL
500     ##$aGesehen am 26.05.10
655     #7$0(DE-588)4067488-5 $0https://d-nb.info/gnd/4067488-5 $0(DE-101)040674886 $aZeitschrift $2gnd-content
776     08$x1990-7958
```



JOURNAL

International journal of electrical and power engineering

2007-

Available Online

TOP
SEND TO
SEARCH INSIDE
VIEW ONLINE
GET IT FROM OT...
DETAILS

Send to

EMAIL PERMALINK REFERENCE REFWORKS ENDNOTE EXPORT RIS EXPORT BIBTEX

Search inside

Search for articles within this journal

Article title or keyword

View Online

Full text availability

Free E- Journals

Available from 2007 volume: 1 issue: 1.

Get it from other locations

For more accurate results please sign in Sign in

Resource sharing is only allowed under specific terms and conditions

Details

Title	International journal of electrical and power engineering
Creation Date	2007-
Publisher	Faisalabad MedWell 1.2007 -
Genre	Zeitschrift
Other title	International Journal of Electrical and Power Engineering
Format	Online-Ressource
Source	Library Catalog

Links

Display Source Record

Appendix HECKMANN01

MARC View

LEADER 02400nas a2200529 a 4500				
001 50026713				
003 SIRSI				
005 20260114090047.0				
006 m o d				
007 cr n				
008 980610c19829999ne br pss 0 b0eng c				
010			a| 2003233217	
016	7		a| 1460279-9 2| DE-600	
016	7		a| 1460279 9. 2| DE-600	
016	7		a| 012435072 2| Uk	
016	7		a| 1460279 9 2| DE-600	
019			a| DO NOT EDIT this 360 MARC RECORD	
022			a| 1872-7182 l| 0167-6393 y| 0167-6393	
030			a| SCOMDH	
035			a| (WaSeSS)ssj0004882	
040			a| OH1 b| eng c| OH1 d| OCL d| OCLCQ d| GUA d| OCLCQ d| VRC d| U9S d| GUA d| OCLCQ d| AZU d| LGG d| OCLCQ d| OCLCO d| OCLCQ d| OCLCO d| OCLCF d| OCLCO d| OCLCQ d| BUF d| U3W d| WYU d| VT2 d| UHL d| LIP d| OCLCQ d| UKSSU d| UK7LJ d| OCLCQ d| OCLCO d| OCL d| OCLCQ d| UAB d| LDP d| UKUAL d| UKMGB d| DXU d| SFB d| OCLCL d| WaSeSS	
042			a| pcc	
050	1	4	a| BF455	
130	0		a| Speech communication (Online)	
222		0	a| Speech communication b| (Online)	
245	1	0	a| Speech communication h| [electronic resource].	
246	1		i| Also known as: a| Speech com	
260			a| [Amersterdam] : b| Elsevier Science	
310			a| Bimonthly, b| 2002-	
321			a| Frequency varies, b| <1995>-2001	
362	1		a| Began with vol. 1, issue 1 (May 1982).	
530			a| Also issued in print.	
538			a| Mode of access: World Wide Web.	
550			a| A Publication of the European Association for Signal Processing (EURASIP) and the International Speech Communication Association (ISCA).	
588	0		a| Vol. 16, issue 1 (Jan. 1995); title from contents screen (ScienceDirect, viewed October 17, 2003).	
588	1		a| Vol. 41, issues 2/3 (Oct. 2003), viewed October 17, 2003.	
650		0	a| Oral communication v| Periodicals.	
650		0	a| Communication.	
650		2	a| Communication	
655		0	a| Electronic journals.	
655		2	a| Periodical	
710	2		a| European Association for Signal Processing. 1| https://id.oclc.org/worldcat/entity/E39QH7JmrGHjqb8hFgKxpqcKtg.	
710	2		a| International Speech Communication Association. 1| https://id.oclc.org/worldcat/entity/E39QH7Jmp8HCf8gyJ94gjq7b9t.	
776	1		t| Speech communication x| 0167-6393 w| (DLC)sc 83001666 w| (OCoLC)8843982	
856	4	0	u| http://SK8ES4MC2L.search.serialssolutions.com/?V=1.0&L=SK8ES4MC2L&S=JCs&C=SPEECOM&T=marc	
910			a| CONSER online	
949			a| Electronic resource w| ASIS c| 1 i| 50026713-1001 l| ONLINE m| ONLINE r| Y s| Y t| ONLINE u| 1/17/2026 o| .STAFF. 360 MARC	

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES

Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES

Facebook
Twitter
Instagram

Speech communication [electronic resource].

Uniform Title	Speech communication (Online)
Additional Titles	Also known as: Speech com
Published	[Amersterdam] : Elsevier Science
Additional Creators	European Association for Signal Processing and International Speech Communication Association

Access Online

serialssolutions.com

Dates of Publication and/or Sequential Designation
Began with vol. 1, issue 1 (May 1982).

Subject(s)
Oral communication—Periodicals
Communication

Genre(s)
Electronic journals

ISSN
1872-7182

Publication Frequency
Bimonthly, 2002-
Frequency varies, <1995>-2001

Other Forms
Also issued in print.

Technical Details
Mode of access: World Wide Web.

Issuing Body
A Publication of the European Association for Signal Processing (EURASIP) and the International Speech Communication Association (ISCA).

View MARC record | catkey: 50026713

ASK A LIBRARIAN

PennState

Copyright ©2019 The Pennsylvania State University. All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Legal Statements | PSU Hotlines

PENN STATE UNIVERSITY LIBRARIES
Libraries Home (814) 865-6368
Libraries Intranet (Staff Only)
Accessibility Help
Website Feedback
Policies and Guidelines
Staff Directory

CONNECT WITH PENN STATE UNIVERSITY LIBRARIES
Facebook
Twitter
Instagram

Online Journals, Books and more

All | Journals Only | Books Only | Videos Only

Title begins with | Search

Show more search options ▼

Showing results 1 through 1 of 1

for the search: "SPEECOM"

Refine Results: All | Journals Only | Books Only | Videos Only

Limit by: Peer Reviewed | Open Access



Speech communication

ISSN: 0167-6393

Peer Reviewed

1982 to 1994 in Backfile Package - Computer Science (Legacy) [YCS]

05/01/1982 to Present in ScienceDirect Journals

Search inside this journal | Search

Back to Top

Copyright ©2025 The Pennsylvania State University.
All rights reserved. Except where otherwise noted, this work is subject to a Creative Commons Attribution 4.0 license. Details and exceptions.

Last Updated on 02-09-2026 © Ex Libris, a ProQuest Company, 2000-2026

Appendix
HECKMANN02

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 736–752
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A hierarchical framework for spectro-temporal feature extraction

Martin Heckmann [a,*], Xavier Domont [a,b], Frank Joublin [a], Christian Goerick [a]

[a] Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany
[b] Technische Universität Darmstadt, Control Theory and Robotics Lab, D-64283 Darmstadt, Germany

Available online 11 August 2010

Abstract

In this paper we present a hierarchical framework for the extraction of spectro-temporal acoustic features. The design of the features targets higher robustness in dynamic environments. Motivated by the large gap between human and machine performance in such conditions we take inspirations from the organization of the mammalian auditory cortex in the design of our features. This includes the joint processing of spectral and temporal information, the organization in hierarchical layers, competition between coequal features, the use of high-dimensional sparse feature spaces, and the learning of the underlying receptive fields in a data-driven manner. Due to these properties we termed the features as hierarchical spectro-temporal (HIST) features. For the learning of the features at the first layer we use Independent Component Analysis (ICA). At the second layer of our feature hierarchy we apply Non-Negative Sparse Coding (NNSC) to obtain features spanning a larger frequency and time region. We investigate the contribution of the different subparts of this feature extraction process to the overall performance. This includes an analysis of the benefits of the hierarchical processing, the comparison of different feature extraction methods on the first layer, the evaluation of the feature competition, and the investigation of the influence of different receptive field sizes on the second layer. Additionally, we compare our features to MFCC and RASTA-PLP features in a continuous digit recognition task in noise. On a wideband dataset we constructed ourselves based on the Aurora-2 task, as well as on the actual Aurora-2 database. We show that a combination of the proposed HIST features and RASTA-PLP features yields significant improvements and that the proposed features carry complementary information to RASTA-PLP and MFCC features.
© 2010 Elsevier B.V. All rights reserved.

Keywords: Spectro-temporal; Auditory; Robust speech recognition; Image processing; Learning; Competition; Hierarchical

1. Introduction

Humans have the astonishing ability to preserve stable representations even in a dynamic environment. In contrast to automatic speech recognition systems additional background noise and changes in the transmission channel have only minor effects on human performance (Lippmann, 1997; Sroka and Braida, 2005). Hence better understanding the underlying processes in humans bears the potential to yield better recognition systems.

Unfortunately, our knowledge of the auditory processing in the mammalian brain is still quite limited. The visual system is for example already much better understood (King and Nelken, 2009). This is also expressed in the wealth of corresponding computational models of the visual system. On the other hand, there are several studies highlighting important similarities between the two systems. Sur et al. (1988) showed that newborn ferrets whose retinal nerves were rerouted to the auditory part of the thalamus, sometimes called the gateway to the cortex (Crick, 1984), where later able to respond to visual stimuli via their auditory cortex. This at least demonstrates a high plasticity of these areas during development if not a strong similarity in functional organization. Despite significant differences between auditory and visual processing in the brain common, modality independent processing principles are usually assumed (Read et al., 2002; King and Nelken, 2009). Different authors have shown that at least at the level of the receptive fields in the primary visual and

* Corresponding author. Tel.: +49 69 8901 1755.
E-mail addresses: martin.heckmann@honda-ri.de (M. Heckmann), xavier.domont@rtr.tu-darmstadt.de (X. Domont), frank.joublin@honda-ri.de (F. Joublin), christian.goerick@honda-ri.de (C. Goerick).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.006

IPR2026-00272
Krisp Technologies EX1010 Page 360

auditory cortices these similarities can be found (Schreiner and Calhoun, 1994; de Charms et al., 1998; Shamma, 2001). Measurements in the primary auditory cortex of different animals revealed its spectro-temporal organization, i.e. the receptive fields are selective to modulations in the time-frequency domain. The corresponding receptive fields have, as in the visual cortex, Gabor-like shapes. The above mentioned findings suggest that modeling principles known from image processing can beneficially be transferred to auditory tasks.

Traditionally, speech features mainly took inspirations from psychoacoustic findings and thereby relied in most cases on independent spectral (Hermansky, 1990; Hermansky and Morgan, 1994; Flynn and Jones, 2008; Haque et al., 2009) or temporal representations (Hermansky and Sharma, 1998).

In recent years also features inspired by above mentioned similarities between visual and auditory processing, i.e. features capable of directly capturing spectro-temporal variations, were developed. In his seminal work Kleinschmidt (2002) introduced the usage of 2 D Gabor features for speech recognition in noise. This was followed by others, also employing Gabor features on similar tasks, including speech vs. non-speech discrimination (Mesgarani et al., 2006; Meyer and Kollmeier, 2008; Sherry and Zhao, 2008). In (Elhilali and Shamma, 2006) a spectro-temporal respresentation based on Gabor filters was used for source separation. Ezzat et al. (2007) randomly selected spectro-temporal patches of the target word from the training set and then used these as features for keyword spotting.

The framework for the extraction of spectro-temporal speech features we present here takes many inspirations from the visual object recognition system of Wersing and Körner (2003) and is an extension of our previous work (Domont et al., 2007, 2008). In contrast to the previously mentioned approaches and other models in the literature we integrated additional processing principles which are also inspired by the mammalian sensory cortex. The processing in the sensory cortices seems to be organized in a hierarchical fashion. This has been stated for the visual (Hubel and Wiesel, 1965; Felleman and Van Essen, 1991) and auditory cortex (Rauschecker, 1998; Read et al., 2002; Scott et al., 2003). Based on this principle we propose a hierarchical framework consisting of two layers.[1] Features in the first layer extract local information. In the second layer the results of these local features are combined to form more complex features. The hierarchical processing and the construction of more complex and at the same time more specific features leads to a substantial increase in the number of features. A trend also observed in the human brain where approximately 3500 inner hair cells are present in the cochlea and about 100,000,000 neurons in the auditory cortex (Dusan and Rabiner, 2005).

In general, when dealing with spectro-temporal features developing methods for the selection of the relevant features is a key issue. Kleinschmidt (2002) already proposed to use a so-called "feature finding neural network" to select the features yielding the best recognition rates. The unsupervised learning of sparse spectro-temporal representations was proposed in (Klein et al., 2003; Behnke, 2003). Cho and Choi (2005) investigated how such learned representations can be applied to the task of sound classification. We follow this idea in that we learn the receptive fields on both layers of our hierarchy with unsupervised learning rules. Another important property of the mammalian sensory cortices is the competition between coequal features. To model this we implemented a Winner-Take-Most competition between the features on the first layer.

The following sections will describe our framework in more detail and will evaluate its performance in comparison to conventional speech features. In Section 2 we give an overview on our framework. Section 3 describes the preprocessing we apply to the speech signal prior to the feature extraction. The learning of the receptive fields and the extraction of the spectro-temporal features is detailed in Section 4. Section 5 presents recognition results we obtain on a noisy digits task. Based on this task this section also evaluates the contribution of the different elements of our framework to the overall performance. Finally, in Section 6 we discuss the results we obtain and possible improvements.

2. Overview

The key elements of our hierarchical feature extraction framework are depicted in Fig. 1. The first step performs a preprocessing of the speech signal and mainly consists of a transformation into the frequency domain and an enhancement of the formant structure. Based on this we calculate local spectro-temporal features. This step is followed by a competition between these local features. Together these two steps constitute the first layer of our framework. On the second layer the local features are combined to form complex features, spanning larger time and frequency regions. The final steps are an orthogonalization of the features via a Principal Component Analysis (PCA) and recognition of the feature stream with a Hidden Markov Model (HMM) based recognizer. Strictly speaking we do not consider the last two steps as being part of our framework. However, they are necessary to evaluate the feature extraction. Due to their hierarchical organization we termed the features resulting from the proposed framework as hierarchical spectro-temporal (HIST) features.

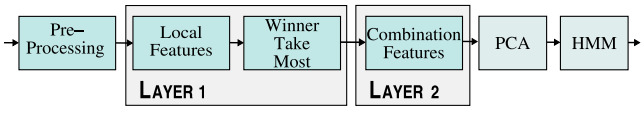


Fig. 1. Overview of the feature extraction process.

[1] Behnke (2003) already suggested a hierarchical speech feature extraction framework but did not evaluate the resulting features in respect to what information they extract and how this could be used for speech processing.

IPR2026-00272
Krisp Technologies EX1010 Page 361

3. Preprocessing

As most of the phonetic information is conveyed via the formant variations we perform a preprocessing of the signal after transformation into the spectral domain which aims at enhancing the formant structure.

The process of speech production is commonly modeled via a non-linear volume velocity source followed by a time-varying linear filter and radiation components (Fant, 1970). Hence, the speech signal we hear is the overlay of the excitation signal at the glottis, the time-varying resonance frequencies of the vocal tract, the radiation components of the mouth and lips, and influences of the room. The preprocessing we present in the following was mainly developed in (Gläser et al., 2010). It aims at compensating for the effects of the excitation signal and the radiation components and thereby focuses on the resonance frequencies of the vocal tract, commonly referred to as formants. We do not explicitly model the room effects or the excitation signal.

The spectral tilt introduced by excitation and radiation can be corrected via a pre-emphasis. Additionally, for voiced sounds the glottis converts the steady airflow produced by the lungs into a quasi-periodic train of flow pulses by which the transfer function of the vocal tract is sampled at multiples of the fundamental frequency. Consequently, spectrograms feature spectral peaks at the harmonics rather than the vocal tract resonance frequencies. This effect will be compensated for by a smoothing along the frequency axis.

3.1. Gammatone filterbank

We transform the speech signal into the spectro-temporal domain via the Patterson–Holdsworth auditory filterbank (Patterson et al., 1992). This filterbank is based on neurophysiological findings on the human auditory system and models the peripheral processing as carried out by the cochlea, where sound is transformed into spatio-temporal response patterns on the auditory nerve. It is implemented as a set of linear Gammatone filters, each of them tuned to a different frequency range. Hence, non-linear effects as suppression and level-dependent tuning curves are not modeled. The filterbank we use is composed of 128 filters covering the frequency range from 80 Hz to 8 kHz and follows the implementation suggested by Slaney (1993). Subsequently, the spectral envelope is calculated via rectification and low-pass filtering. Fig. 2(a) exemplary shows the envelope of the filter responses to a sequence of digits from the TIDigits database (Leonard et al., 1984).

3.2. Pre-emphasis

The source signal of voiced sounds is produced at the glottis. Fant (1979) suggested to use a second-order low-pass filter for the approximation of the glottal flow spectrum. This adequately approximates the most common phonation types, modal and creaky phonation (Childers and Lee, 1991). Thus, voiced excitation changes the spectral characteristic by −12 dB/oct.

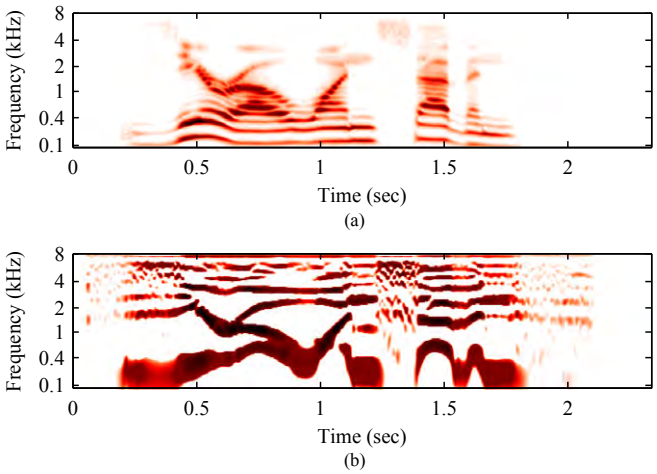


Fig. 2. In (a) the original spectrogram for the digit sequence "zero-one-seven" spoken by a male speaker is shown. The spectrogram after the application of the preprocessing for enhancing the formant structure is shown in (b).

Following (Stevens, 2000) the radiation components, i.e. the lip impedance, can be modeled via a first-order high-pass filter yielding a spectral change of +6 dB/oct. Overall this means that a pre-emphasis via amplification of frequency magnitudes by +6 dB/oct adequately eliminates the spectral influence of excitation and radiation.

3.3. Spectral filtering

After the above mentioned pre-emphasis we enhance the formant structure in the spectrogram by smoothing along the frequency axis following the same spirit as (Baer et al., 1993). We obtain this by a filtering with channel-dependent Difference-of-Gaussians (DoG) operators with standard deviations of the negative Gaussian components being twice as large as that of the corresponding positive ones:

$$\mathrm{DoG}_k(f) = \frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{(f-fc_k)^2}{2\sigma^2_{\mathrm{DoG}_k}}\right) - \frac{1}{2}\exp\left(-\frac{(f-fc_k)^2}{8\sigma^2_{\mathrm{DoG}_k}}\right)\right). \quad (1)$$

Here, $\mathbf{DoG}_k$ is the DoG operator of channel k featuring a center frequency fc_k. As standard deviations σ_{DoG_k} we use 70 Hz for filter channels with center frequencies in the range from 80 Hz to 5 kHz and 400 Hz filter channels with center frequencies in the range from 5 kHz to 8 kHz. We discretized the DoGs by sampling them at the logarithmically arranged Gammatone filterbank's channel center frequencies and normalized the resulting $\mathbf{DoG}_k$ vectors. Fig. 3 exemplarily depicts the DoGs of 8 filter channels. The smoothed spectrogram $\boldsymbol{S}_{\mathrm{Smooth}}$ is obtained via multiplying the matrix

IPR2026-00272
Krisp Technologies EX1010 Page 362

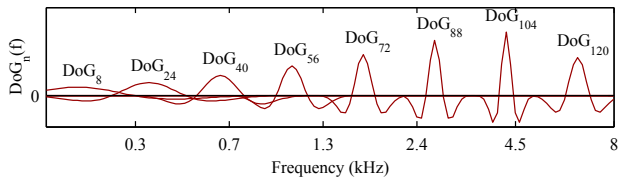


Fig. 3. The DoG operators of the filter channels used to enhance the formant structure in spectrograms vary in bandwidth and frequency resolution dependent on the channels' center frequencies. Here, the DoG operators of 8 exemplarily chosen filter channels are shown.

DoG containing in its rows the vectors $\mathbf{DoG}_k^T$ with the spectrogram following the pre-emphasis $\boldsymbol{S}_{\text{Pre-emphasis}}$:

$$\boldsymbol{S}_{\text{Smooth}} = \mathbf{DoG} \cdot \boldsymbol{S}_{\text{Pre-emphasis}}. \tag{2}$$

The smoothing with the DoG operator results in a suppression of the responses from neighboring frequency channels and thereby sharpens the peaks at the formants. We add a non-linear component to this suppression by suppressing all negative values. Additionally we apply a fourth root compression on the resulting signal to approximate the non-linear loudness perception:

$$\boldsymbol{S}_{\text{Compressed}} = \sqrt[4]{\max\left(\boldsymbol{S}_{\text{Smooth}}, 0\right)}. \tag{3}$$

In a final step we filter the spectrograms along the time direction with a low-pass filter having a cut-off frequency of 100 Hz and then perform a downsampling to 400 Hz. Fig. 2(b) continues the example depicted in (a) by showing the corresponding resulting spectrogram $\boldsymbol{S}$.

4. Hierarchical spectro-temporal features

The hierarchical feature extraction framework we present here consists of two layers. The first layer extracts local features, covering small regions in the spectro-temporal domain. In the second layer these local features are integrated to more complex features integrating information over longer time spans and frequency ranges (e.g. 40 ms × 8 kHz). Structuring it in this way was largely inspired by the visual object recognition system of Wersing and Körner (2003). Such hierarchical feature extraction schemes have shown to be beneficial in visual object recognition (Fukushima, 1980; Riesenhuber and Poggio, 1999; Fergus et al., 2003). In the following we will demonstrate how these approaches can be transferred to the auditory domain.

4.1. Extraction of local features

On the first layer $Q^{(1)}$ we extract features via a 2 D filtering of the preprocessed spectrograms $\boldsymbol{S}$ with a set $n^{(1)}$ of receptive fields $\boldsymbol{w}_l^{(1)}$. After filtering we only keep the absolute value of the response:

$$q_l^{(1)}(t,f) = \left|\left(\boldsymbol{S} * \boldsymbol{w}_l^{(1)}\right)(t,f)\right|, \tag{4}$$

The indices t and f indicate time and frequency. As a consequence of the filtering border effects occur when the receptive fields only partially overlap with the input spectrogram. We decided to keep the size of the $n^{(1)}$ spectrograms $\boldsymbol{q}_l^{(1)}$ after the filtering identical to the spectrogram $\boldsymbol{S}$ at the input and use an overlay-save implementation for the filtering in the time direction. This yields small border effects in the frequency direction and in most cases negligible effects in the time direction for the first and last block.

The filtering with the receptive fields, i.e. the filter kernels, is identical to a correlation between the spectrogram at the input and the frequency and time inverted receptive fields. Hence, the responses in the output spectrograms indicate the similarity of the spectro-temporal patches in the input spectrogram to the (time and frequency inverted) receptive fields.

Neurobiological experiments show that the processing in the primary visual cortex aims at reducing the redundancy of the stimuli. When applying corresponding unsupervised learning methods on natural images one obtains receptive fields similar in shape to Gabor functions (Olshausen et al., 1996; van Hateren and Ruderman, 1998).

We therefore decided to learn the filters of the first layer of our feature hierarchy with Independent Component Analysis (ICA) (Comon, 1994). For the learning we used 3500 randomly selected patches of size 50 ms × 20 channels, i.e. 20 × 20 pixels, taken from the training set of the TIDigits database (Leonard et al., 1984). We applied the FastICA fixed-point algorithm (Hyvärinen, 1999) for the calculation of the ICA. A tanh non-linearity was used, i.e. the negentropy of the data is approximated by the contrast function $G(u) = 1/a \log\cosh(au)$, with the parameter a in the range 1–2. For $n^{(1)} = 8$ the resulting receptive fields, depicted in Fig. 4, are rather difficult to interpret. They show some preferences for steady formants and upward and downward transitions. However, overall they are not very specific and only vaguely resemble Gabor functions.

Additionally, we also investigated the usage of genuine Gabor functions. The Gabor functions were not learned but selected manually. During this selection process we investigated different sets of filters. Thereby, we set the

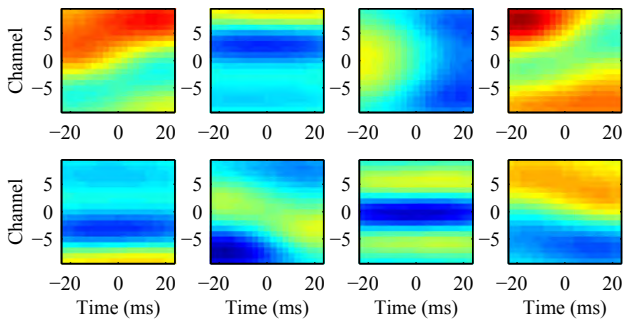


Fig. 4. Visualization of the receptive fields used for the extraction of local features learned via ICA. The frequency axis is given in channels as on the non-linear frequency grid the actual extend in frequency depends on the position on the frequency axis of the receptive field.

IPR2026-00272
Krisp Technologies EX1010 Page 363

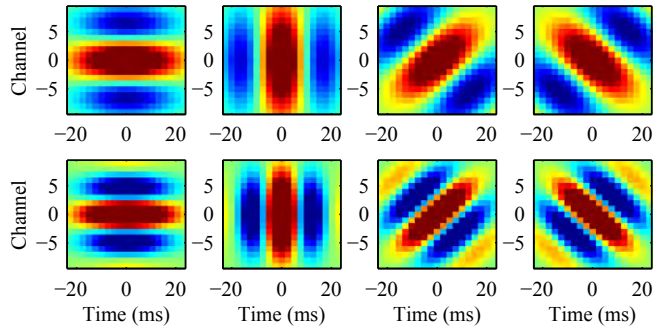


Fig. 5. Visualization of the Gabor based receptive fields used for the extraction of local features. The frequency axis is given in channels as on the non-linear frequency grid the actual extend in frequency depends on the position on the frequency axis of the receptive field.

modulation frequencies and orientations to resemble those of the learned ICA filters and to what one would expect to be a relevant feature given the widths and orientations of the formants in the enhanced spectrograms (see Fig. 5 for an example).

A discrete two dimensional complex Gabor function is a complex sinusoidal carrier $s(t,f)$ modulated with a Gaussian envelope $g(t,f)$. Important properties of Gabor functions are the shape of the Gaussian envelope, controlled via the variances σ_{G_t} and σ_{G_f}, and the frequency and orientation of the carrier, governed by the radian frequencies ω_t and ω_f. Further parameters are the center of mass in time t_0 and frequency f_0. The complex carrier $s(t,f)$ is defined as

$$s(t,f) = \exp\left[i\omega_t(t-t_0) + i\omega_f(f-f_0)\right] \tag{5}$$

and the Gaussian envelope $g(t,f)$ as

$$g(t,f) = \frac{1}{2\pi\sigma_{G_t}\sigma_{G_f}} \cdot \exp\left[\frac{-(t-t_0)_r^2}{2\sigma_{G_t}^2} + \frac{-(f-f_0)_r^2}{2\sigma_{G_f}^2}\right], \tag{6}$$

where the subscript r denotes the rotated coordinates

$$\begin{aligned}(t-t_0)_r &= (t-t_0)\cos\theta + (f-f_0)\sin\theta,\\ (f-f_0)_r &= -(t-t_0)\sin\theta + (f-f_0)\cos\theta\end{aligned} \tag{7}$$

with $\theta = \tan^{-1}(\omega_f/\omega_t)$, if $\omega_f \neq 0$, and $\theta = sign(\omega_f)\pi/2$ otherwise. A modulation frequency $\omega_t = 0$ yields purely spectral filters ($\theta = \pi/2$), a modulation frequency $\omega_f = 0$ purely temporal filters ($\theta = 0$), and settings in between spectro-temporal filters. The real part of the complex Gabor function yields an even filter and the imaginary part an odd filter. For our implementation we used the 8 odd filters depicted in Fig. 5.

4.2. Competition between local features

Ideally the neurons in the previous processing step will only respond to spectro-temporal patterns in the spectrogram matching their receptive field (e.g. an upward moving formant). However, the neurons are not sufficiently selective and most of the time several neurons respond with similar strength to a given pattern. To counterbalance this we introduce a competition mechanism between the different neurons similar to Wersing and Körner (2003). We independently apply a Winner-Take-Most (WTM) competition between the activities $q_l^{(1)}(t,f)$ of the l neurons for all points (t,f) in the spectrogram:

$$r_l^{(1)}(t,f) = \begin{cases} 0 & \text{if } \frac{q_l^{(1)}(t,f)}{M(t,f)} < \gamma^{(1)} \\ & \text{or } M(t,f) = 0 \\ \frac{q_l^{(1)}(t,f) - \gamma^{(1)} M(t,f)}{1-\gamma^{(1)}} & \text{else,} \end{cases} \tag{8}$$

$M(t,f) = \max_k q_k^{(1)}(t,f)$ is the maximal value at position (t,f) over the $n^{(1)}$ neurons and $0 \leqslant \gamma^{(1)} \leqslant 1$ is a parameter controlling the strength of the competition. By suppressing less active neurons the selectivity of the feature extraction is enhanced, a process commonly found in the auditory system (Young, 2008).

Furthermore, a non-linear transformation including a threshold $\vartheta^{(1)}$ is applied on all the $r_l^{(1)}(t,f)$:

$$s_l^{(1)}(t,f) = H(r_l^{(1)}(t,f) - \vartheta^{(1)}), \tag{9}$$

where $H(x)$ is the Heaviside step function.

To introduce robustness against small fluctuations in location and amplitude as well as further sharpen the contrast between the activities we perform a pooling followed by a non-linear transformation on all the $s_l^{(1)}(t,f)$. The pooling consist of a smoothing with a 2D Gaussian filter $\boldsymbol{g}^{(1)}$ and reduction of the resolution of the activations $\boldsymbol{s}_l(t,f)$ by a factor of four in both frequency and time dimension. As non-linear transformation we use a tanh function:

$$c_l^{(1)}(t,f) = \tanh\left(\boldsymbol{s}_l^{(1)} * \boldsymbol{g}^{(1)}\right)(4t,4f). \tag{10}$$

This yields 32 frequency channels and a sampling rate of 100 Hz. In the visual object recognition system of (Wersing and Körner, 2003) a binarization was performed after the tanh non-linearity. This results in further robustness against fluctuations. As prerequisite it is required that the dynamics of the input data does not vary too much. For images this can be guaranteed via a normalization of each

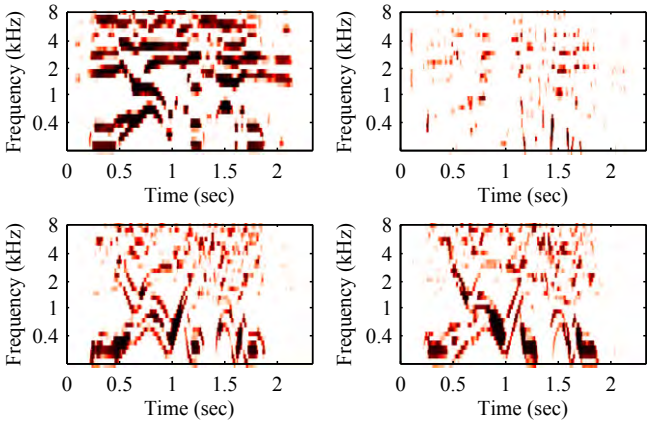


Fig. 6. The results of the first feature extraction layer obtained with the example from Fig. 2 are visualized. As local features the Gabor functions depicted in Fig 5 were used. Only the responses of the first four receptive fields (top row in Fig. 5) are shown.

IPR2026-00272
Krisp Technologies EX1010 Page 364

individual image. In the case of speech such a normalization is more difficult and has to be adaptive. For the time being we omitted the binarization but we are convinced that such an adaptive input normalization followed by stronger non-linearities will be very beneficial for the feature extraction process.

Fig. 6 visualizes the results of the first feature extraction layer with the example from Fig. 2. The preference of the receptive fields for certain orientations can clearly be seen. One can further notice that as a consequence of the Winner-Take-Most competition mainly only one feature is active at a given point in the spectrogram. In the following we will refer to these features as $\boldsymbol{c}^{(1)}$.

4.3. Extraction of combination features

In contrast to previous approaches which extract spectro-temporal information we only rely on a few initial neurons in the first layer $Q^{(1)}$ but introduce a second layer $Q^{(2)}$ which combines the responses of the neurons on the first layer to more complex spectro-temporal patterns.

Each of the $n^{(2)}$ combination patterns in $Q^{(2)}$ is composed of $n^{(1)}$ receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, i.e. one for each of the neurons in the previous stage $Q^{(1)}$. The coefficients of these receptive fields are non-negative and span all frequency channels. Similarly to (4) the activity $q_k^{(2)}(t)$ of the kth neuron at the time t is given by

$$q_k^{(2)}(t) = \sum_{l=1}^{n^{(1)}} \left(\boldsymbol{c}_l^{(1)} * \boldsymbol{w}_{l,k}^{(2)} \right)(t,f) \tag{11}$$

with the main difference that the final activity $q_k^{(2)}(t)$ is the result of a summation over all $n^{(1)}$ receptive fields. As the combination patterns span the whole frequency range the response of the neurons does not depend on f anymore. This means that, by computing the convolution, the patterns $\boldsymbol{w}_{l,k}^{(2)}$ are only shifted in the time direction. Note that the absolute value is not required in (11) as both the $\boldsymbol{c}_l^{(1)}$ and the $\boldsymbol{w}_{l,k}^{(2)}$ are non-negative.

The combination patterns were also learned in an unsupervised manner using Non-Negative Sparse Coding (NNSC) (Hoyer, 2004). NNSC differs from Non-negative Matrix Factorization (NMF) by the presence, in the cost function (12), of a sparsity enforcing term which aims at limiting the number of non-zero coefficients required for the reconstruction. Consequently, if a feature appears often in the data, it will be learned, even if it can be obtained by a combination of two or more other features. Therefore, the NNSC is expected to learn complex and global features appearing in the data.

We cut out patches of length $\Delta = 40$ ms of the first layer activations $\boldsymbol{c}_l^{(1)}$. From these patches we learned $n^{(2)} = 50$ combination features by minimizing the following cost function (Wersing and Körner, 2003):

$$E = \sum_i \left\| \boldsymbol{P}_i - \sum_{k=1}^{n^{(2)}} \alpha_{k,i} \boldsymbol{w}_k^{(2)} \right\|^2 + \beta \sum_i \sum_{k=1}^{n^{(2)}} |\alpha_{k,i}|, \tag{12}$$

where $\boldsymbol{P}_i$ is a tensor representing the $n^{(1)}$ layers of the ith patch, the $\boldsymbol{w}_k^{(2)}$ are $n^{(2)}$ non-negative tensors each of them containing the $n^{(1)}$ receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, the $\alpha_{k,i}$ are nonnegative reconstruction factors, and β is a parameter allowing to control the sparsity of the learned features.

4.4. Integration with an HMM

HMMs are still the dominant paradigm for speech recognition. Therefore, we also use them to assess the performance of the proposed feature extraction framework. Due to notorious shortages in training data it is advisable to use diagonal covariance matrices for the Gaussian mixtures. This entails that decorrelated features are better suited. Consequently, we perform a Principal Component Analysis (PCA) prior to transferring the features to the HMM. The PCA statistics were determined from the training part of the TIDigits database. Additionally, we also add Delta and Delta–Delta features (i.e. the first and second derivative of the features) to the feature vector. The integration of the Delta features is done prior to the calculation of the PCA. When in the following we refer to HIST features we refer to the features as obtained after the application of the PCA.

5. Results

To asses the performance of the proposed HIST features we use two different datasets. In both cases the task is the robust speaker-independent recognition of continuous digits under a comprehensive variety of additional background noise. This setup is inspired by the Aurora-2 framework (Pearce and Hirsch, 2000). The difference between the two datasets is that our first and main dataset consists of wideband, i.e. 16 kHz, signals whereas Aurora-2, our second dataset, consists of telephone quality signals, i.e. 8 kHz bandwidth with additional channel distortions. Via different tests on the wideband dataset we determine the role the different elements of the presented feature extraction framework play for the overall performance. The evaluation on Aurora-2 serves to further substantiate our results.

5.1. Speech databases

5.1.1. Wideband corpus

Similar to Aurora-2 we also derived our wideband corpus from TIDigits. TIDigits contains 326 speakers each pronouncing 77 digit sequences (Leonard et al., 1984). In contrast to Aurora-2 we use the full TIDigits set. To this data we added four types of noise from the Noisex database (Varga and Steeneken, 1993): White, Babble, Factory, and Car. Each noise type was added at Signal to Noise Ratios (SNRs) ranging from −5 dB. . . inf, i.e. we also kept the clean signal. The differences in our speech data to Aurora-2 are:

- We downsampled signals to 16 kHz instead of 8 kHz.
- When mixing the signals with noise using FaNT (Hirsch, 2005) we used the G.712 only for the noise and signal

IPR2026-00272
Krisp Technologies EX1010 Page 365

level estimation, i.e. the obtained signals have no channel distortions.
- We took four types of noise from the Noisex database.
- Each individual test condition contains significantly (12 times) more utterances than Aurora-2.

We have chosen a sampling rate of 16 kHz for our main corpus as our primary application domain is the speech based interaction with Honda's humanoid robot. In this case a restriction to 8 kHz is without justification. Additionally, also telephone speech has already partially moved to higher bandwidths.

The Hidden Markov Models were trained on clean signals with HTK using the same parameters as in the Aurora-2 framework. Whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state were used.

5.1.2. Aurora-2

Aurora-2 uses a subset of 8440 training sentences out of the 12,549 training sentences contained in TIDigits. These signals were downsampled to 8 kHz and convolved with a G.712 characteristic yielding telephone quality signals with an effective bandwidth of 0.3, ..., 3.4 kHz. Aurora consists of a clean and a multi condition training part, both of equal size. In the multi condition training part recordings from inside a subway, babble noise, car noise, and recordings in an exhibition hall are added to the signals. The first test data set contains signals degraded by the same types of noise as used in the multi condition training. In the second set recordings from a restaurant, in a street, at an airport, and at a train station are added to the signals. Finally, in the third set the signals are convolved with a MIRS instead of a G.712 characteristic. As noise types subway and street are used for this set. All noise types were added, as in our wideband dataset, at SNR levels ranging from -5 dB... inf. Each individual test contains only 1001 utterances, compared to 12,547 in our wideband dataset. As a consequence the statistical significance of the results obtained on Aurora-2 is notably inferior to that we obtained on our wideband dataset. If not stated otherwise, all following tests are performed on the wideband dataset.

5.2. Parameterization of the feature extraction framework

The proposed HIST feature extraction framework requires the proper setting of different parameters. With the exception of those experiments where we mention a different setting of the parameters the following parameterization is used for all subsequent tests.

For the extraction of local features on the first level of the hierarchy we used 8 filters learned via ICA. The corresponding receptive fields are depicted in Fig. 4. The sizes of the receptive fields were 20×20 at a sampling rate of 400 Hz. Hence they span 50 ms and 20 channels.

For the competition between the different local features we used $\gamma^{(1)} = 0.7$. The following non-linear compression uses a value of $\vartheta^{(1)} = 0.25$. In the pooling stage the dimensions in time and frequency are reduced by four.

We learned $n^{(2)} = 50$ combination features $(q_1^{(2)}(t), \ldots, q_{n^{(2)}}^{(2)}(t))^T$ all spanning the full frequency range. As temporal extend we used 40 ms wide receptive fields. At a sampling rate of 100 Hz this corresponds to four samples.

Delta (resp. Delta–Delta) features were computed using a 9th order FIR low-pass (resp. band-pass). After decorrelation of the feature vectors via PCA we retained from the 150 features the 39 with the highest eigenvalues. We have chosen this number as it corresponds to the dimensionality of the MFCC features we use as a benchmark.

Due to the several non-linearities in our processing the performance gain due to a modification at a certain stage in the processing is a very unreliable predictor for the performance of the complete hierarchy. For this reason we always report the performance of the whole hierarchy. An exception to this is of course the investigation on the contribution of the different layers.

5.3. Comparison to purely spectral features

First, we want to compare the results we achieve with the proposed HIST features to those of conventional spectral features. As benchmark we used Mel Frequency Cepstral Coefficients (MFCCS) (Davis and Mermelstein, 1980) and RelAtive SpecTral Perceptual Linear Predictive (RASTA-PLP) features (Hermansky and Morgan, 1994). In the case of MFCC features we applied Cepstral Mean Subtraction as it yielded in almost all cases significant performance improvements. For both feature types we also calculated Delta and Delta–Delta features in the same fashion as for the HIST features. When comparing the recognition scores in Fig. 7 one can see that the HIST features show inferior performance than either the MFCC or RASTA-PLP features in clean and low noise conditions (2.5% for HIST vs. 1.5% for RASTA-PLP in clean).[2] From this it is evident that some important information in the speech signal is not well represented by the HIST features. However, with increasing noise level the purely spectral features deteriorate quickly whereas the performance of the HIST features is more robust. One can also observe that the MFCC s yield better results in clean than RASTA-PLP features (1.2% vs. 1.5%) but deteriorate faster when additional background noise is present.

Based on the observation that the performance of the proposed HIST features and the conventional features is complementary, i.e. conventional features showing good performance at low noise levels and HIST features in high noise levels, we combined the HIST features with

[2] We also observed that the HIST features have a strong tendency to produce erroneous word insertions. To counterbalance this we used a word insertion penalty of `-p -70` in the HTK decoding for all results we report for the HIST features alone. All remaining tests, including the combination of HIST with other features, were performed with the standard setting `-p 0`.

IPR2026-00272
Krisp Technologies EX1010 Page 366

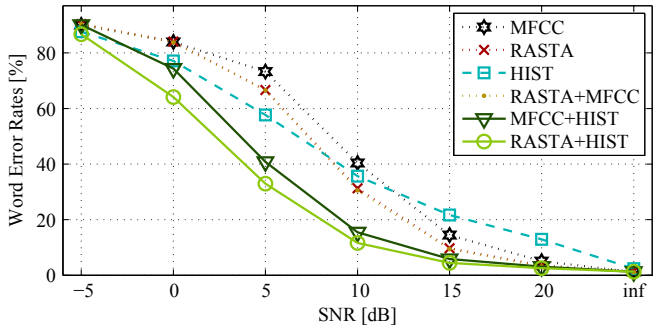


Fig. 7. Word error rates (WERs) when factory noise was added to the speech data.

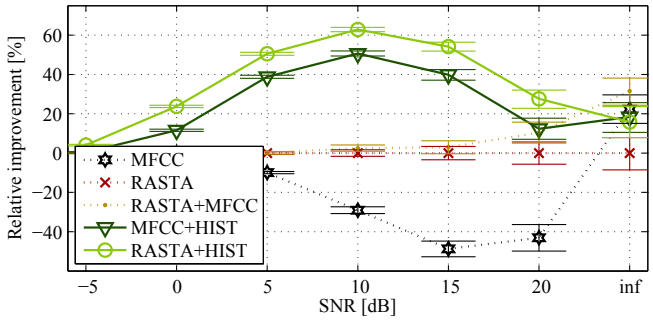


Fig. 8. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

Table 1
Word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf.

Features	Noise type			
	White	Factory	Babble	Car
MFCC	49.1	44.1	35.6	25.4
RASTA-PLP	43.1	41.0	35.0	19.5
$c^{(1)}$	52.7	51.7	63.0	21.8
HIST	43.3	42.1	53.9	17.9
RASTA-PLP + MFCC	43.5	40.7	32.8	20.4
MFCC + HIST	33.6	33.0	36.4	11.9
RASTA-PLP + $c^{(1)}$	30.6	31.6	37.3	12.0
RASTA-PLP + HIST	26.7	29.1	48.7	10.6
$\text{HIST}_{\text{Gabor}}$	50.4	40.9	57.4	24.1
RASTA-PLP + $\text{HIST}_{\text{Gabor}}$	46.9	39.4	57.4	21.2
$\text{HIST}_{\text{No WTM}}$	58.6	70.3	89.5	47.7
RASTA-PLP + $\text{HIST}_{\text{No WTM}}$	33.8	35.1	45.0	23.7
$\text{HIST}_{20\text{ms}}$	44.2	42.7	51.4	18.9
RASTA-PLP + $\text{HIST}_{20\text{ms}}$	34.9	35.5	40.7	13.9
$\text{HIST}_{80\text{ms}}$	54.0	48.1	54.5	17.1
RASTA-PLP + $\text{HIST}_{80\text{ms}}$	41.7	40.7	60.9	11.7

RASTA-PLP features. Thereby we concatenated at each frame the 45 RASTA-PLP features and the 39 HIST features to an 84 dimensional feature vector. For comparison we also combined MFCC and HIST features and MFCC and RASTA-PLP features in the same way. As can be seen from Fig. 7 the combination of RASTA-PLP and MFCC features gives a significant improvement for clean (1.0%) but does not help when additional background noise is present. In fact, it is hard to distinguish the two curves in Fig. 7. The combination of HIST and RASTA-PLP features shows a very similar performance to the RASTA-PLP features alone in low noise conditions and gives substantial improvements when the noise level increases. The same is true for the combination of HIST and MFCC features although the overall performance is not as good, most likely due to the inferior performance of the MFCC features with additional background noise. Table 1 displays the recognition results for each of the 4 noise types tested.

To better quantify the performance of the combination of the proposed HIST features with RASTA-PLP features compared to conventional features we visualized the relative improvements over RASTA-PLP features in Fig. 8. Improvements of the combination of HIST and RASTA-PLP features are moderate for clean speech (from 1.5% to 1.3% on clean) and reach levels of 40–60% at medium SNR levels. Over most SNR levels we obtain an SNR gain of approximately 5 dB, i.e. we obtain from the combination of HIST and RASTA-PLP features almost the same word error rates as at 5 dB better SNR levels when using only RASTA-PLP features. It can also be seen that the combination of MFCC and RASTA-PLP features does not lead to a significant improvement. The averaged relative improvements are given in Table 2.

Despite the significant improvements we see from the combination of HIST and RASTA-PLP features in most cases the performance with additional babble noise is quite poor. We attribute this to the preprocessing and the strong non-linearities. The preprocessing very effectively enhances the formant structure. This is also the case when only babble noise is present. Due to the following non-linearities this formant structure is then further emphasized such that it does not much differ anymore from speech at a normal level and generates word or phone hypotheses. This can also be seen from a deeper analysis of the errors. With 20 dB babble noise RASTA-PLP features show 3.3% word errors, 0.9% insertions and deletions and additional 2.4% confusions. On the other hand the combination of HIST and RASTA-PLP features yields 14.7% word errors, 12.6% insertions and deletions and only 2.1% confusions. This means that the combination of HIST and RASTA-PLP features does indeed in more cases recognize the correct word but on the other hand produces tremendously more erroneous word insertions.

The largest improvements we see are for scenarios with high significance for real applications. For the rather simple task of continuous digit recognition word error rates should certainly not exceed 10%. When adding the HIST features to the RASTA-PLP features at an SNR level of 10 dB in white noise, factory noise, and car noise we see improvements form 13.5% to 7.0%, from 31.3% to 11.6%, and from 8.0% to 3.7%. Hence more than a reduction by two down to tolerable error rates.

IPR2026-00272
Krisp Technologies EX1010 Page 367

Table 2
Relative word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf. The error rates obtained with RASTA-PLP features were used as a baseline.

Features	Noise type			
	White	Factory	Babble	Car
MFCC	−29.1	−15.4	2.5	−13.1
RASTA-PLP + MFCC	3.4	6.9	17.7	16.4
MFCC + HIST	19.8	24.6	−11.8	38.8
RASTA-PLP + $\boldsymbol{c}^{(1)}$	30.1	26.2	−45.1	31.9
RASTA-PLP + HIST	36.7	34.1	−109.0	41.0
RASTA-PLP + $\text{HIST}_{\text{Gabor}}$	−51.1	10.4	−267.9	11.8
RASTA-PLP + $\text{HIST}_{\text{No WTM}}$	21.4	18.4	−88.9	−12.7
RASTA-PLP + $\text{HIST}_{\text{20ms}}$	22.7	21.4	−53.7	31.4
RASTA-PLP + $\text{HIST}_{\text{80ms}}$	5.4	7.3	−223.0	39.2

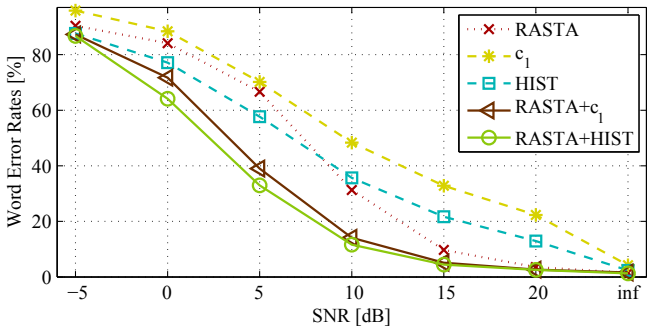


Fig. 9. Word error rates (WERs) when factory noise was added to the speech data.

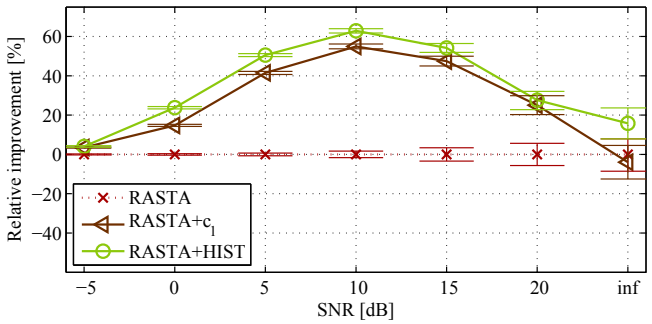


Fig. 10. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

Due to the inferior performance of the MFCC features by themselves and in combination with the HIST features we will in the following only use the RASTA-PLP features.

5.4. Contribution of the hierarchical processing

Next, we want to investigate the contribution of the two hierarchical levels to the overall performance. We calculated for the features $\boldsymbol{c}^{(1)}$ after the first layer of our framework (compare Eq. (10)) a PCA in the same way as for the complete HIST features. With the resulting features and

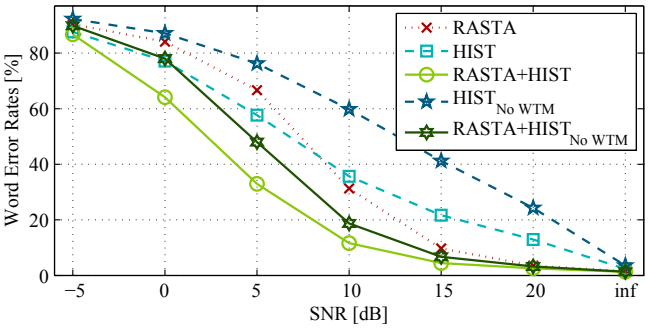


Fig. 11. Word error rates (WERs) when factory noise was added to the speech data.

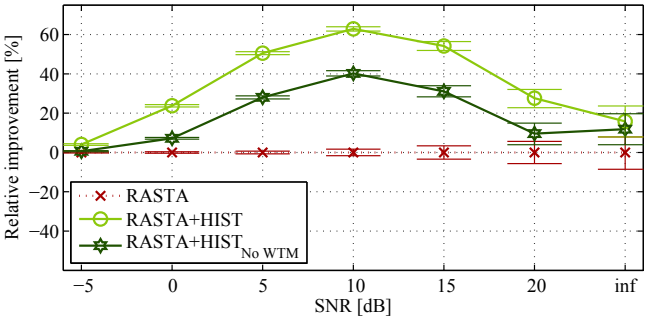


Fig. 12. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

with the concatenation of these features and RASTA-PLP features we trained an HMM. The results indicated with $\boldsymbol{c}^{(1)}$ and RASTA-PLP + $\boldsymbol{c}^{(1)}$ in Figs. 9 and 10 as well as in Tables 1 and 2 show that the features on the first layer alone already yield good performance. Nevertheless the second layer further improves this performance in all cases, except for babble noise. The results based on the $\boldsymbol{c}^{(1)}$ features show lower word insertions than the HIST features when babble noise is added but at the same time yield more confusions (7.1% insertions and deletions and 2.2% confusions compared to 12.6% insertions and deletions and 2.1% confusions at 20 dB SNR).

5.5. Contribution of the feature competition

Now we want to analyze the influence of the competition between the features on the first layer introduced in Eq. (8) on the performance. The results we obtain without the Winner-Take-Most (WTM) competition are displayed in Figs. 11 and 12 as well as Tables 1 and 2. As one can see the impact of the competition depends on the noise type and noise level. In the case of car noise the competition seems to be especially beneficial. However, with additional babble noise the results without the competition are better than with it. A closer look reveals again that without the competition the number of erroneous insertions decreases but at the same time the number of confusions increases. Overall the competition leads to notable improvements.

IPR2026-00272
Krisp Technologies EX1010 Page 368

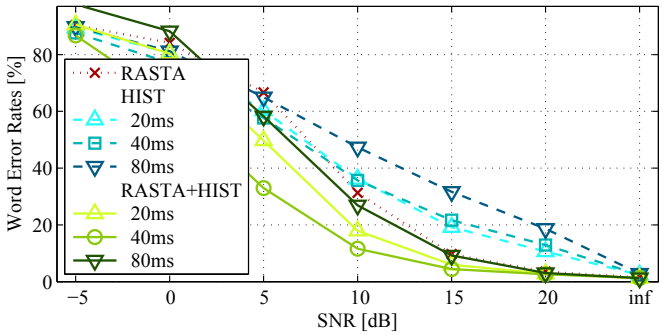


Fig. 13. Word error rates (WERs) when factory noise was added to the speech data.

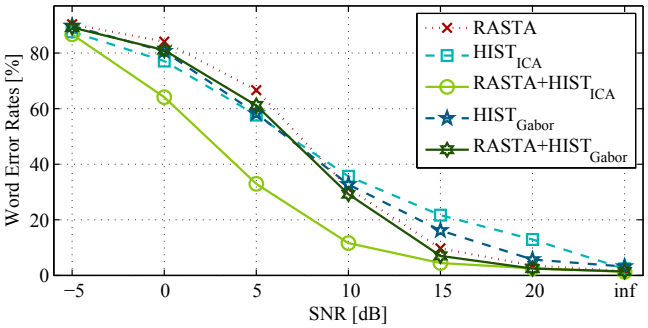


Fig. 15. Word error rates (WERs) when factory noise was added to the speech data.

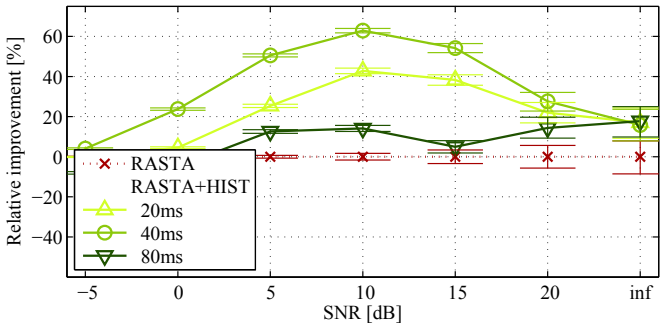


Fig. 14. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

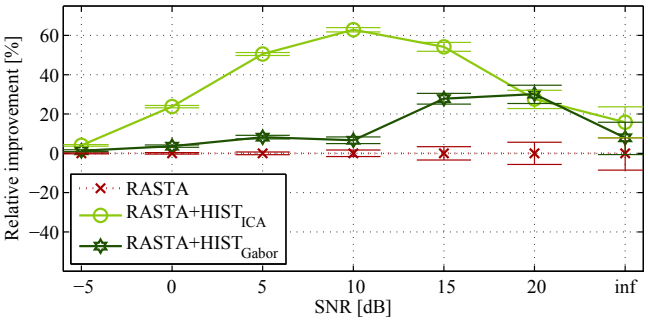


Fig. 16. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

5.6. Contribution of the size of the combination features

One possible source for the improvements we see could be that our features span a longer time window than RASTA-PLP and MFCC features (40 ms compared to 25 ms). To further investigate this we varied the size, i.e. the length, of the combination features. We used 20, 40, and 80 ms. Taking significantly longer time windows would harbour the risk of a strong adaptation of the features to the lexical content. As can be seen from Tables 1 and 2 and Figs. 13 and 14 a width of 40 ms for the second layer seems to be optimal. With smaller and larger widths the performance decreases. Nevertheless with a width of 20 ms the performance is largely better than that of RASTA-PLP features alone. Keeping in mind that already the filters at the first layer have a width of 50 ms we nevertheless conclude from the results that it is not only the longer temporal window which is responsible for the performance improvements we see.

5.7. Choice of the first layer

In this experiment we want to evaluate the role of the receptive field shape on the first layer. In the tests so far we used the features learned via ICA (compare Fig. 4). In Figs. 15 and 16 as well as Tables 1 and 2 we compare them to features derived from the Gabor functions displayed in Fig. 5. Even tough the shape of the Gabor filters seems to match the shapes one observes in the spectrograms much better than the ICA features the results we obtain are significantly inferior to those with the features learned via ICA. During these tests we varied the sizes and frequencies of the Gabor filters to a large extend but did not find a setting which yielded better results as those reported above. This could change when one uses more than just 8 filters.

5.8. Adaptation to HMMs

Standard HMMs yield best results with independent features which statistics can well be modeled as the overlay of a few Gaussian distributions (Hermansky, 1998). However, as can be seen from the feature covariances in Fig. 17(a) and (b) the correlation between the different dimensions of the HIST features is much stronger than for RASTA-PLP features. The PCA step effectively eliminates this problem and hence is indispensable for the successful integration with an HMM featuring diagonal covariance matrices (compare Fig. 17(c)).

In addition, we use the PCA also to reduce the dimensionality of the features. In Figs. 18 and 19 we compare the results we obtain when we either use all 150 dimensions of the HIST features and combine them with RASTA-PLP

IPR2026-00272
Krisp Technologies EX1010 Page 369

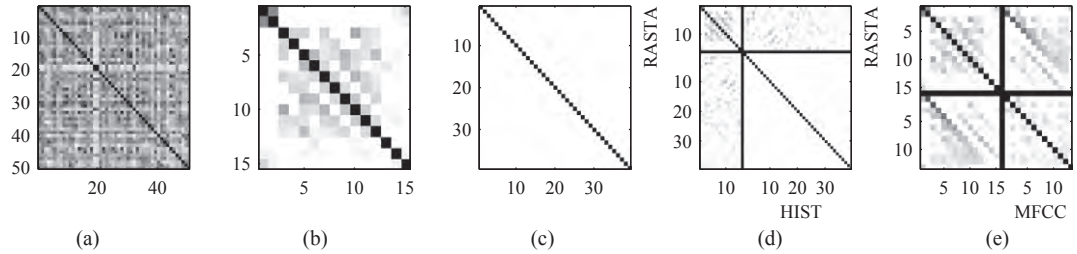


Fig. 17. Covariance matrices of HIST prior to PCA (a), RASTA-PLP (b), HIST after PCA (c), RASTA-PLP + HIST (d), and RASTA-PLP + MFCC (e) features. Prior to the calculation of the covariance matrices we removed the mean and performed a variance normalization of each feature individually. Delta and Delta-Delta features were excluded, except for the HIST features after the PCA in (c) and in combination with RASTA-PLP features in (d). For the combination of RASTA-PLP + HIST and RASTA-PLP + MFCC the separation between the two feature sets is highlighted with a black bar.

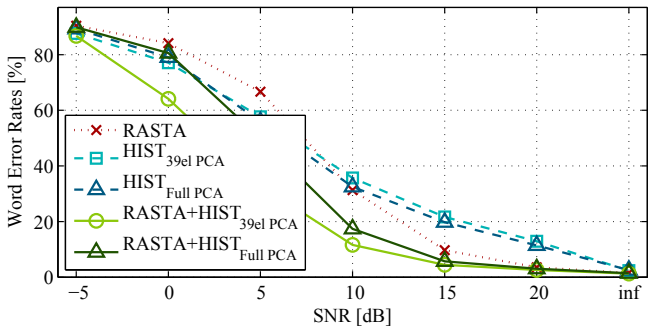


Fig. 18. Word error rates (WERs) when factory noise was added to the speech data.

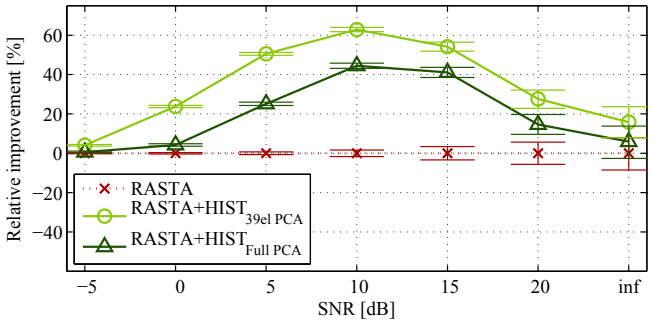


Fig. 19. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

features to those when we reduce the dimensionality to 39 as we did in the previous experiments. As can be seen from these figures the performance slightly improves for the HIST features alone but does decrease for the combination of HIST and RASTA-PLP features. We attribute this to the weakness of the HMMs to identify the relevant features in this combination. When we use all 150 HIST features the additional features resulting from eigenvectors with low eigenvalues and putatively low significance seem to partly mask the 45 RASTA-PLP features. Hence, the preselection of the relevant HIST features by retaining only the 39 features with the highest eigenvalues seems to be beneficial when combining them with lower dimensional features in an HMM recognition system.

5.9. Complementarity of the features

We also want to use the covariance matrices between the different feature types to shed some light on the causes for the improvements we see. Fig. 17(d) and (e) show that the correlation between the HIST features and the RASTA-PLP features is much lower than that between the MFCC features and the RASTA-PLP features. From this it follows that the HIST features deliver information complementary to RASTA-PLP features whereas MFCC and RASTA-PLP features extract very similar information. This explains on one hand why we do not see noticeable improvements when combining RASTA-PLP and MFCC features and on the other hand why we do observe them when combining either RASTA-PLP or MFCC features with HIST features.

Table 3
Word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf when training was performed on noisy data.

Features	Noise type			
	White	Factory	Babble	Car
RASTA-PLP	27.5	26.4	22.0	6.5
HIST	24.8	27.7	40.0	8.4
$HIST_{adapted}$	26.2	28.7	49.8	16.5
RASTA-PLP + HIST	23.9	24.9	22.9	5.3
RASTA-PLP + $HIST_{adapted}$	21.9	23.8	22.8	6.2

5.10. Matched training

In all previous tests the features as well as the HMM models were trained on clean data and tested with different levels of additional background noise. In most scenarios one has some a priori information on the type and strength of the background noise one will face. Training the HMM models with speech corrupted by this additional background noise significantly increases performance when later tested in similar conditions. Therefore, we additionally performed a test where we added all four noise types at SNR levels of 20 and 10 dB to the training set of

IPR2026-00272
Krisp Technologies EX1010 Page 370

Table 4
Relative word error rates averaged for an individual type of additional noise over SNR values ranging from −5 dB... inf when training was performed on noisy data. The error rates obtained with RASTA-PLP features were used as a baseline.

Features	Noise type			
	White	Factory	Babble	Car
RASTA-PLP + HIST	21.7	9.7	2.1	17.6
RASTA-PLP + $HIST_{adapted}$	30.3	19.1	4.1	11.0

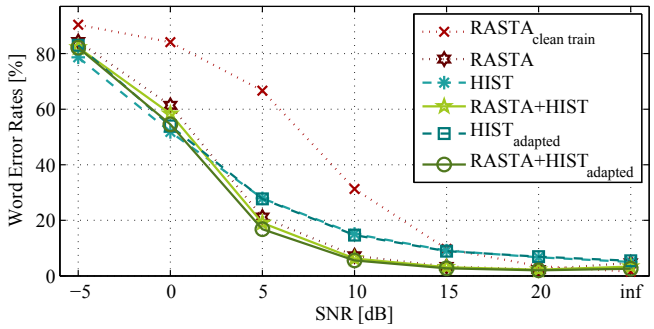


Fig. 20. Word error rates (WERs) when factory noise was added to the speech data. Training was performed on noisy data. Results for the original HIST features or HIST features adapted to the training condition are displayed.

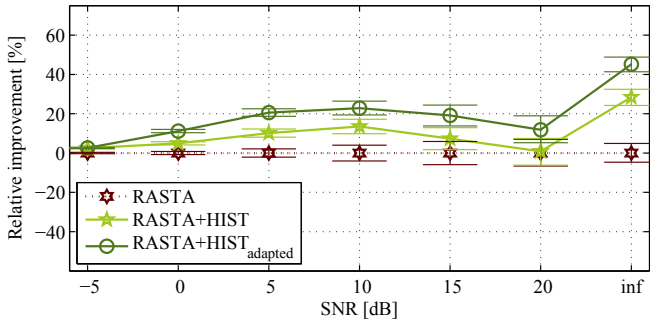


Fig. 21. Word error rates (WERs) when factory noise was added to the speech data relative to those obtained with RASTA-PLP features alone. Training was performed on noisy data. Results for the original HIST features or HIST features adapted to the training condition are displayed. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

TIDigits. We combined this with the clean signals and trained the HMM models. For computational reasons we used in all cases only a third of the training set. During recognition we used the same data as in the previous tests.

As can be seen from Tables 3 and 4 as well as Figs. 20 and 21 this matched training reduces error rates in well matching scenarios to a large extend. This is true for RASTA-PLP features as well as for the combination of RASTA-PLP and HIST features. On the other hand the performance on the clean signal severely decreased (from 1.5% to 4.7% with RASTA-PLP and 3.4% for the combination of HIST and RASTA-PLP). What we also observed is that the strong tendency of the HIST features to generate insertion errors in the case of babble noise disappeared. Now we see an improvement from the combination of RASTA-PLP and HIST features for all cases tested. We attribute this to the fact that in this case the HMMs learned to reject weak speech hypotheses generated by the HIST features. Overall the improvement from the combination of the features is smaller than in the case when training was performed only on clean signals. The reason for this seems to be two-fold. First, the weaker generalization capabilities of the RASTA-PLP features is remedied by the better adaptation of the HMMs to the scenarios. Second, in this test we used so far features which were learned on clean signals.

In a further test we also learned the features on the noisy data used to train the HMMs. We want to refer to the resulting features as $HIST_{adapted}$. The results in Tables 3 and 4 as well as Figs. 20 and 21 show that this decreased the performance of the HIST features when taken alone but improves performance when used in combination with RASTA-PLP features for almost all cases. We see improvements of around 20%. In the case of clean the improvement even amounts to 45% up to a word error rate of 2.6%. An exception is the case of car noise. We saw before that car noise has a much smaller effect on the recognition performance than the other noise types. Therefore, we presume that including the other noise types in the training leads for tests with additional car noise to a stronger mismatch between training and testing than in the case of training in clean and testing in noise.

5.11. Evaluation on Aurora-2

The tests so far were all performed on the original TIDigits corpus where we added noise at different SNR levels, i.e. our wideband corpus with a 16 kHz sampling rate. Despite the fact that we compared our features to the commonly used RASTA-PLP and MFCC features we wanted to further substantiate our results with additional tests on Aurora-2.

For the tests on Aurora-2 we only changed the parameters in our feature hierarchy directly related to the new sampling rate of 8 kHz. Thereby the different characteristic of the telephone quality signals also in the low frequency range is not taken into account. We adapted the Gammatone filter bank such that the filter center frequencies are distributed in the range of $0,\ldots,4$ kHz instead of $0,\ldots,8$ kHz as before. The DoG filters from Eq. (1) were adapted accordingly. All remaining parameters remained unchanged.

In Table 5 the results of these test are depicted. Additionally, in Fig. 22 the absolute word error rates and in Fig. 23 the relative word error rates compared to the RASTA-PLP features for subway noise in the clean training condition are given. Figs. 24 and 25 show the corresponding results for multi condition training.

The results show that we obtain substantial improvements for some noise types but there are also problematic

IPR2026-00272
Krisp Technologies EX1010 Page 371

Table 5
Word error rates (WERs) and relative improvements for the Aurora-2 dataset. The training was carried out either with clean speech, or with speech mixed with various noise types (multi condition training). Tests were always performed using a mixture of noisy and clean signals. The presented scores are averages over all SNRs levels (i.e. −5 dB... inf).

		Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train-station	Subway (MIRS)	Street (MRIS)	Average
Clean condition training	RASTA-PLP	35.2	33.8	37.4	38.6	34.4	35.5	31.3	35.2	35.2	35.8	35.2
	HIST	52.5	68.3	51.6	57.1	61.2	42.5	52.0	52.9	36.1	36.4	51.1
	RASTA-PLP + HIST	30.0	34.9	34.0	35.8	36.0	30.5	29.0	30.9	30.9	32.9	32.5
	Rel. improvement	22.3	−11.6	7.5	11.0	−2.9	19.6	6.6	14.8	20.5	8.3	9.6
Multi condition training	RASTA-PLP	18.2	20.9	22.7	21.3	20.4	20.8	18.9	22.2	18.4	20.8	20.5
	HIST	25.0	37.5	24.4	27.0	35.5	27.9	28.6	29.2	22.9	27.9	28.6
	RASTA-PLP + HIST	17.1	20.4	21.8	20.4	19.9	19.0	18.2	20.1	17.0	20.0	19.4
	Rel. improvement	13.2	1.1	10.5	14.8	10.2	16.4	3.2	14.3	16.1	11.9	11.2

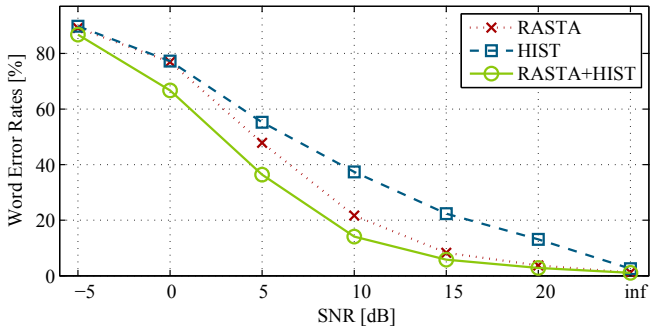


Fig. 22. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the clean training set and testing with additional subway noise.

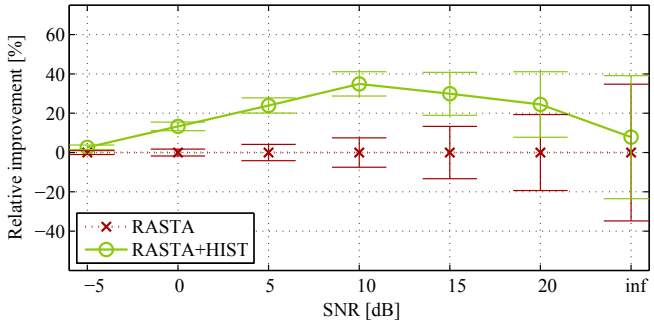


Fig. 23. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the clean training set and testing with additional subway noise relative to those obtained with MFCCS alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

noise types where the combination of HIST and RASTA-PLP features yields only small improvements or is even not beneficial. A closer analysis of the noise used in Aurora-2 reveals that additional speech is present in almost all noise types. Only subway, car, and train-station noise do not contain any additional background speech. Hence we assume that the partly unfavorable behavior of our HIST features for most of the remaining noise types is related to the same problem we saw already for babble noise. The results obtained on the multi condition training, depicted in Table 4 as well as Figs. 24 and 25 support this

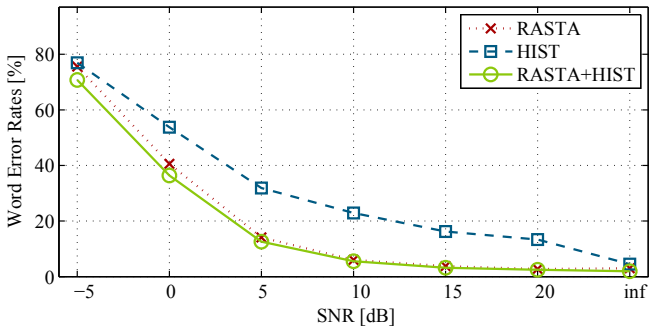


Fig. 24. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the multi condition training set and testing with additional subway noise.

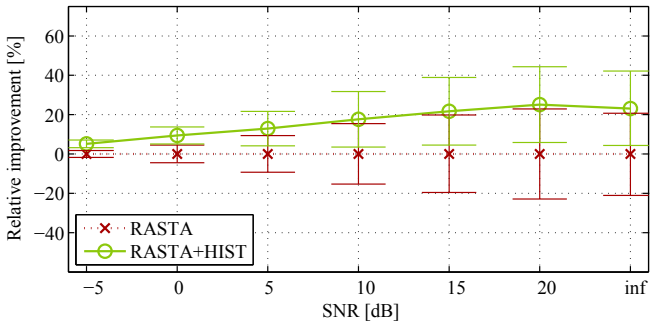


Fig. 25. Word error rates (WERs) obtained on the Aurora-2 dataset with training on the multi condition training set and testing with additional subway noise relative to those obtained with MFCCS alone. Bars indicate the 95% confidence intervals calculated according to Vilar (2008).

hypothesis. In this case both the HIST features and the HMM models were learned form the multi condition training part of Aurora-2. As we saw already previously, e.g. in Fig. 20, this multi condition training improves performance of the RASTA-PLP features in noisy conditions substantially and at the same time the benefit from the combination with the HIST features is reduced. However, the unfavorable behavior of the HIST features when speech is also present in the noise also gets smaller. In this case

IPR2026-00272
Krisp Technologies EX1010 Page 372

we now see again improvements for all types. When comparing these results to the previous ones one hast also to take into account their much weaker statistical significance (this can e.g. be assessed when comparing the confidence intervals depicted in Figs. 8 and 22.

5.12. Computational complexity

The proposed feature extraction framework involves several time consuming 2 D convolutions. Nevertheless, we were able to integrate the feature extraction into an online speech recognition system (Heckmann et al., 2009). We used a Linux workstation with eight Intel Xeon X5355 CPUs working at 2.66 GHz with 4 MB cache and 4 GB RAM per processor. Using only one core of one of the processors, the extraction of the HIST features for 1 s of speech lasts approximatively 280 ms, i.e. 3.5 times faster than real time. Additionally, the higher dimensionality of the resulting feature vector (84 dimensions for HIST + RASTA-PLP compared to only 45 for RASTA-PLP alone) increases the computational load during decoding. However, this can easily be handled by current computers for most speech recognition tasks.

6. Discussion

Similar to other approaches we apply pattern matching techniques inspired by image processing to the task of speech feature extraction. This was motivated not only by the observation that speech shows distinct patterns in the spectrogram but also by recent findings in neurobiology on similarities in the processing in the visual and auditory cortex. We extended this idea by introducing additional concepts from models of visual object perception as unsupervised feature learning, feature competition, hierarchical processing, and high-dimensional sparse representations.

With the results obtained from the first layer of our feature extraction framework we could replicate previous findings by, e.g., Kleinschmidt and Gelbart (2002) as well as Meyer and Kollmeier (2008), that adding spectro-temporal information improves the recognition in noise. We cannot fully rule out the hypothesis that the main reason for the improvements we see lies in the nature of our preprocessing. However, in (Gläser et al., 2010) we compared a preprocessing very similar to the one presented here on a formant tracking task to a Linear Predicitive Coding (LPC) based preprocessing. There we saw that our preprocessing shows under some conditions, especially for the tracking of the second formant in white noise, superior performance but in most cases yields similar results as the LPC based preprocessing. Therefore, we assume that the improvements we see here should be attributed to the representation of spectro-temporal information via the proposed hierarchy. In light of the good results (Kim and Stern, 2010) obtained by applying a 15th-root on the envelope signal we will in the future also investigate if such a strong non-linearity is also beneficial to further improve our preprocessing.

At least for the small number of features we use the actual shape of the features in the first layer seems to play an important role. Despite numerous variations we always achieved significantly better results when learning the features via ICA than when using Gabor shaped features. Hence, the unsupervised data-driven learning of the features seems to be superior to a predefined set of features.

When investigating the feature competition we saw consistent improvements. However, we applied the framework without further modification when calculating the results reported above without the competition. It is well possible that tuning of the parameters will reduce this difference in performance. Nevertheless, we saw in all cases we tested a gain from the competition. On the other hand we also observed that the competition is so far not ideally adapted to the task. In some cases the receptive fields responses largely varied with minor displacements in the spectrogram. As a consequence different features dominate at closely neighboring points thereby breaking up larger structures in the spectrogram and dispersing them over the different features. Modifying the purely local competition such that also surrounding regions are taken into account seems to be a promising approach to further improve the results of this competition step.

One key aspect of our feature extraction framework is the processing in two hierarchical layers. In contrast to purely spectral features and previous spectro-temporal features the proposed second layer, which we termed combination layer, is able to represent complete formant configurations and model non-stationary patterns. We showed that adding this second layer further improves the recognition performance on clean data and in noise. This underlines that the information represented by this second layer is indeed relevant.

In addition to the recognition experiments we demonstrated via a correlation analysis that conventional spectral features as MFCCS and RASTA-PLP features capture similar information. On the other hand, the proposed HIST features extract additional, i.e. complementary, information. From this analysis we also saw that without an additional PCA the different feature dimensions of the HIST features are much stronger correlated than it is the case for RASTA-PLP or MFCC features. Such a behavior is particularly inapt for the use with an HMM. In general one can expect that due to a "coevolution" of spectral features and HMMs an HMM is not very well suited to be used with such novel features (Morgan et al., 2005). Similar to (Meyer and Kollmeier, 2008; Wang et al., 2008; Hermansky and Sharma, 1998; Chen et al., 2004) which use less conventional features and hence features less well suited to HMMs we also expect further improvements when applying alternative recognition backends as e.g. Multilayer Perceptrons or TANDEM modelling (Hermansky et al., 2000; Morgan and Bourlard, 1995). Such alternative recognition backends will be of particular

IPR2026-00272
Krisp Technologies EX1010 Page 373

importance when we continue to pursue our goal to obtain sparse and high-dimensional representations and increase the number of elements on the different layers.

We evaluated our features on two different databases each with a comprehensive set of background noise types and SNR levels. On the dataset containing the TIDigits data at a sampling rate of 16 kHz we saw, except for the case of babble noise, consistent improvements from the combination of the HIST features and RASTA-PLP or MFCC features. In particular we obtain strong improvements in scenarios most relevant for actual applications. In the case of babble noise the tendency of the HIST features for high insertions is certainly unfavorable. Also for the tests we performed on the Aurora-2 database we saw unfavorable behavior for those noise types which also included speech. A remedy to this is on one hand the use of additional mechanisms to detect speech on- and offsets and, as we showed, the use of speech corrupted by babble noise already in the training phase. Furthermore, we expect that improvements in our preprocessing will also be able to alleviate this problem.

The improvements we obtained when combining the HIST features with purely spectral features on Aurora-2 were notably lower than those we obtained on TIDigits. One of the reasons for this is certainly that during the development of the HIST features we took only 16 kHz wideband speech data into account and did only change the parameters of the Gammatone filter bank and the DoG filters when applying it to Aurora-2 which uses telephone quality speech sampled at 8 kHz and with an efficient bandwidth of $0.3, \ldots, 3.4$ kHz. Adaptation of the preprocessing, the filter sizes and the different non-linearities will certainly yield to better performance also on telephone quality speech.

When taken alone our features show for good and moderate SNR levels clearly inferior performance to conventional spectral features. From this it follows that some important information captured very well by RASTA-PLP and MFCC features is not represented by the HIST features. We speculate that the smoothing along the time and frequency axis, the preprocessing as well as the pooling stage, could be at the root of this. The information retained might be especially robust against additional distortions but on the other hand not detailed enough to provide precise recognition in undistorted speech. Further experiments will be necessary to validate this hypothesis and develop mechanisms to better capture the relevant information such that the presented approach can serve as a true alternative to conventional features.

So far we only reported results on a continuous digit recognition task. It will be interesting to see how the proposed HIST features perform on more complex tasks and if the features learn only characteristics of the specific task or of speech in general. In a recent experiment we could show that the performance of the HIST features when learned on Timit (Garofolo et al., 1993), a database consisting of phonetically rich American sentences not containing numbers, and then applied on TIDigits the performance only degraded little compared to the case when learning the features on the same dataset as later used for the recognition experiments (Heckmann, 2010). We see this as a first indication that the features are not only learning information limited to the task but more general speech properties.

The additional performance gain we observed via an adaptation of the features to the noise in the matched training condition illustrates that such an adaptation to the current situation can be beneficial. For mammals such task specific plasticity of receptive fields in the auditory cortex seems to play an important role (Fritz et al., 2003). Given that all the learning steps involved in the presented feature extraction framework are unsupervised such an online adaptation seems to the least possible.

Despite the significant improvements we obtained compared to spectral features, many questions on how to optimally combine the different elements we presented here are still open. This is also manifest in the observation we made that by different settings of parameters one could tune the features to have different properties like better performance on clean at the cost of inferior performance in noise, better performance when used alone compared to better performance when combined with RASTA-PLP features and so on. Also for this reason we always reported the performance of the whole hierarchy. Overall, we think that we could show that the different ideas borrowed from neurobiology and vision research we introduced as well as the framework as a whole enable more robust speech recognition. Nevertheless, we see this rather as early steps with many more to follow.

Acknowledgments

First of all we want to thank Heiko Wersing for providing us the algorithms of his object recognition system and many advise on how to use it. Next, we want to thank Stephan Hasler for supporting us with the different learning algorithms. We also want to thank Claudius Gläser for assisting us with the preprocessing and many fruitful discussions. Furthermore, we want to thank Mark Dunn, Bram Bolder, Antonello Ceravola and, Marcus Stein for their help with the computer and software infrastructure. Finally, we want to thank the anonymous reviewers which helped with their constructive comments to improve the paper.

References

Baer, T., Moore, B., Gatehouse, S., 1993. Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. J. Rehabil. Res. Develop. 30, 49.

Behnke, S., 2003. Discovering hierarchical speech features using convolutional non-negative matrix factorization. In: Proc. Internat. Joint Conf. on Neural Networks (IJCNN), Vol. 4, pp. 2758–2763.

Chen, B., Zhu, Q., Morgan, N., 2004. Learning long-term temporal features in LVCSR using neural networks. In: Proc. 8th Internat. Conf. on Spoken Language (ICSLP). ISCA.

IPR2026-00272
Krisp Technologies EX1010 Page 374

Childers, D., Lee, C., 1991. Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Amer. 90, 2394.
Cho, Y., Choi, S., 2005. Nonnegative features of spectro-temporal sounds for classification. Pattern Recognition Lett. 26 (9), 1327–1336.
Comon, P., 1994. Independent component analysis: a new concept? Signal Process. 36, 287–314.
Crick, F., 1984. Function of the thalamic reticular complex: the searchlight hypothesis. Proc. Natl. Acad. Sci. 81 (14), 4586–4590.
Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Proc. 28 (4), 357–366.
de Charms, R., Blake, D., Merzenich, M., 1998. Optimizing sound features for cortical neurons. Science 280 (5368), 1439–1443.
Domont, X., Heckmann, M., Joublin, F., Goerick, C., 2008. Hierarchical sectro-temporal features for robust speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Las Vegas, Nevada, pp. 4417–4420.
Domont, X., Heckmann, M., Wersing, H., Joublin, F., Menzel, S., Sendhoff, B., Goerick, C., 2007. Word recognition with a hierarchical neural network. Advances in Nonlinear Speech Processing. Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp. 142–151.
Dusan, S., Rabiner, L., 2005. On integrating insights from human speech perception into automatic speech recognition. In: 9th Eur. Conf. on Speech Communication and Technology (EUROSPEECH). ISCA, Lisbon, Portugal.
Elhilali, M., Shamma, S., 2006. A biologically-inspired approach to the cocktail party problem. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Toulouse, France.
Ezzat, T., Bouvrie, J., Poggio, T., 2007. Spectro-temporal analysis of speech using 2-D Gabor filters. In: Proc. INTERSPEECH. ISCA, Antwerp, Belgium.
Fant, G., 1970. Acoustic Theory of Speech Production. Mouton De Gruyter.
Fant, G., 1979. Glottal source and excitation analysis. Speech Transmiss. Lab. Q. Prog. Stat. Rep. 1, 70–85.
Felleman, D., Van Essen, D., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1 (1), 1–47.
Fergus, R., Perona, P., Zisserman, A., 2003. Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Vol. 2.
Flynn, R., Jones, E., 2008. Combined speech enhancement and auditory modelling for robust distributed speech recognition. Speech Comm. 50 (10), 797–809.
Fritz, J., Shamma, S., Elhilali, M., Klein, D., 2003. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat. Neurosci. 6 (11), 1216–1223.
Fukushima, K., 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernet. 36 (4), 193–202.
Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. DARPA TIMIT acoustic–phonetic continuous speech corpus CD-ROM. Philadelphia.
Gläser, C., Heckmann, M., Joublin, F., Goerick, C., 2010. Combining auditory preprocessing and Bayesian estimation for robust formant tracking. IEEE Trans. Audio Speech Lang. Process. 18 (2), 224–236.
Haque, S., Togneri, R., Zaknich, A., 2009. Perceptual features for automatic speech recognition in noisy environments. Speech Comm. 51 (1), 58–75.
Heckmann, M., 2010. Supervised vs. unsupervised learning of spectro temporal speech features. In: Accepted for ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA).
Heckmann, M., Brandl, H., Domont, X., Bolder, B., Joublin, F., Goerick, C., 2009. An audio–visual attention system for online association learning. In: Proc. INTERSPEECH. ISCA, Brighton, UK.
Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.
Hermansky, H., 1998. Should recognizers have ears? Speech Comm. 25 (1–3), 3–27.
Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMMsystems. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 3. IEEE, Istanbul, Trukey.
Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Proc. 2 (4), 578–589.
Hermansky, H., Sharma, S., 1998. TRAPS-classifiers of temporal patterns. In: 5th Internat. Conf. on Spoken Language Processing (ICSLP). ISCA, Sydney, Australia.
Hirsch, G., 2005. FaNT filtering and noise adding tool. Tech. rep., Niederrhein University of Applied Sciences.
Hoyer, P., 2004. Non-negative matrix factorization with sparseness constraints. J. Machine Learn. Res. 5, 1457–1469.
Hubel, D., Wiesel, T., 1965. Receptive fields and functional architecture in two nonstriate visual aeas (18 and 19) of the cat. J. Neurophysiol. 28 (2), 229–289.
Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Networks 10, 626–634.
Kim, C., Stern, R., 2010. Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Dallas, TX, pp. 4574–4577.
King, A., Nelken, I., 2009. Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. 12 (6), 698–701.
Klein, D., Konig, P., Kording, K., 2003. Sparse spectrotemporal coding of sounds. EURASIP J. Appl. Signal Process. 2003 (7), 659–667.
Kleinschmidt, M., 2002. Methods for capturing spectro-temporal modulations in automatic speech recognition. Acta Acust. Acust. 88 (3), 416–422.
Kleinschmidt, M., Gelbart, D., 2002. Improving word accuracy with Gabor feature extraction. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP). ISCA, Denver, CO.
Leonard, R., Incorporated, T., Dallas, T., 1984. A database for speaker-independent digit recognition. In: Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 9. IEEE, San Diego, CA.
Lippmann, R., 1997. Speech recognition by machines and humans. Speech Comm. 22 (1), 1–15.
Mesgarani, N., Slaney, M., Shamma, S., 2006. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans. Audio Speech Lang. Proc. 14 (3), 920–930.
Meyer, B., Kollmeier, B., 2008. Optimization and evaluation of Gabor feature sets for ASR. In: Proc. INTERSPEECH. ISCA, Brisbane, Australia.
Morgan, N., Bourlard, H., 1995. Continuous speech recognition. IEEE Signal Process. Mag. 12 (3), 24–42.
Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., et al., 2005. Pushing the envelope-aside. Signal Process. Mag. IEEE 22 (5), 81–88.
Olshausen, B. et al., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.
Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., C.Zhang, Allerhand, M.H., 1992. Complex sounds and auditory images. In: Cazals, Y., Demany, L., Horner, K. (Eds.), Auditory Physiology and Perception, Proc. 9th Internat. Symposium on Hearing. Pergamon, Oxford, pp. 429–446.
Pearce, D., Hirsch, H., 2000. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP). ISCA, Bejing, China.
Rauschecker, J., 1998. Cortical processing of complex sounds. Curr. Opin. Neurobiol. 8 (4), 516–521.
Read, H.L., Winer, J.A., Schreiner, C.E., 2002. Functional architecture of auditory cortex. Curr. Opin. Neurobiol. 12 (4), 433–440.
Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2 (11), 1019–1025.

IPR2026-00272
Krisp Technologies EX1010 Page 375