

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

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

The Internet Archive and the Wayback Machine

4. 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.

Liu

5. I have reviewed Exhibit #####, a document entitled VoiceFixer: Toward General Speech Restoration with Neural Vocoder by Haohe Liu, et al.
6. Attached hereto as Appendix LIU01 is a screen capture of the Internet Archive Wayback Machine entry for <http://arxiv.org/abs/2109.13731>. I secured these screen captures myself from <https://web.archive.org/web/20211124092448/http://arxiv.org/abs/2109.13731>. This record was captured as of November 24, 2021.
7. Appendix LIU01 is an Internet Archive capture of a web page entitled VoiceFixer: Toward General Speech Restoration with Neural Vocoder as hosted on the arXiv.org website. Appendix LIU01 is a capture of a web page describing VoiceFixer: Toward General Speech Restoration with Neural Vocoder, including the paper's authors, an abstract and paper submission history. According to this page, VoiceFixer: Toward General Speech Restoration with Neural Vocoder was first published in its current form as of October 5, 2021. This page also offers a copy of the paper available for download by clicking the 'PDF' link visible on page 2 of Appendix LIU01.

8. Attached hereto as Appendix LIU02 is a screen capture of the Internet Archive Wayback Machine entry for <https://arxiv.org/pdf/2109.13731.pdf>. I secured these screen captures myself from <https://web.archive.org/web/20211124092457/https://arxiv.org/pdf/2109.13731.pdf>. This record was captured as of November 24, 2021.
9. Appendix LIU02 is an Internet Archive capture of a file entitled "VoiceFixer: Toward General Speech Restoration with Neural Vocoder" as hosted on the arXiv.org website. I secured this PDF copy by clicking the link entitled 'PDF' visible on page 2 of Appendix LIU01. In comparing Appendix LIU02 to Exhibit #####, it is my determination that Exhibit ##### is a true and correct copy of "VoiceFixer: Toward General Speech Restoration with Neural Vocoder" as hosted on the arXiv.org website. As the Internet Archive first captured this file as of November 24, 2021, it is my determination that "VoiceFixer: Toward General Speech Restoration with at least as early as Neural Vocoder" was available to the public at least as early as November 24, 2021 via the Internet Archive and likely earlier via the original arXiv.org website.

Liu II

10. I have reviewed Exhibit #####, a document entitled VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration by Haohe Liu, et al as published in the Proceedings of INTERSPEECH 2022. This paper contains significant differences from the previous edition of Liu discussed in this declaration.

11. Attached hereto as Appendix LIU03 is a screen capture of the Internet Archive Wayback Machine entry for <https://arxiv.org/abs/2204.05841>. I secured these screen captures myself from <https://web.archive.org/web/20220413011549/https://arxiv.org/abs/2204.05841>. This record was captured as of April 13, 2022.

12. Appendix LIU03 is an Internet Archive capture of a page entitled VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration by Haohe Liu, et al as published in the Proceedings of INTERSPEECH 2022 as hosted on the arXiv.org website. Appendix LIU03 is a capture of a web page describing VoiceFixer: Toward General Speech Restoration with Neural Vocoder by Haohe Liu, et al as published in the Proceedings of INTERSPEECH 2022, including the paper's authors, an abstract and paper submission history. According to this page, the INTERSPEECH 2022 edition of VoiceFixer: Toward General Speech Restoration with Neural

Vocoder was first published in its current form as of April 12, 2022. This page also offers a copy of the paper available for download by clicking the 'PDF' link visible on page 2 of Appendix LIU03.

13. Attached hereto as Appendix LIU04 is a screen capture of the Internet

Archive Wayback Machine entry for <https://arxiv.org/pdf/2204.05841.pdf>. I secured these screen captures myself from

<https://web.archive.org/web/20220413012645/https://arxiv.org/pdf/2204.05841.pdf>. This record was captured as of April 13, 2022.

14. Appendix LIU04 is an Internet Archive capture of a file entitled VoiceFixer:

A Unified Framework for High-Fidelity Speech Restoration by Haohe Liu, et al as published in the Proceedings of INTERSPEECH 2022 as hosted on the arXiv.org website. I secured this PDF copy by clicking the link entitled 'PDF' visible on page 2 of Appendix LIU03. In comparing Appendix LIU04 to Exhibit ####, it is my determination that Exhibit #### is a true and correct copy of VoiceFixer: Toward General Speech Restoration with Neural Vocoder by Haohe Liu, et al as published in the Proceedings of INTERSPEECH 2022 as hosted on the arXiv.org website. As the Internet Archive first captured this file as of April 13, 2022, it is my determination that the INTERSPEECH 2022 edition of VoiceFixer: Toward General

Speech Restoration with Neural Vocoder was available to the public at least as early as April 13, 2022 via the Internet Archive and likely earlier via the original arXiv.org website.

Sonsaat

15.I have reviewed Exhibit 1005, “L2-ARCTIC: A Non-Native English Speech Corpus” by Guanlong Zhao, et al.

16.Attached hereto as Appendix ZHAO01 is a screen capture of the Internet Archive Wayback Machine entry for <https://psi.engr.tamu.edu/l2-arctic-corpus/>. I secured these screen captures myself from <https://web.archive.org/web/20191208033554/https://psi.engr.tamu.edu/l2-arctic-corpus/>. This record was captured as of December 8, 2019.

17.Appendix ZHAO01 is an Internet Archive capture of a web page entitled L2-ARCTIC: A Non-Native English Speech Corpus as hosted on the Tamu Lab website. Appendix ZHAO01 is a page describing the research performed in the Zhao paper as well as a link labelled ‘publication’ allowing site visitors to download a PDF copy of this paper. As the Internet Archive first captured this page as of December 8, 2019, it is my determination that “L2-ARCTIC: A

Non-Native English Speech Corpus” was first made available at least as early as of December 8, 2019 via the original Tamu Lab website.

18. Attached hereto as Appendix ZHAO02 is a screen capture of the Internet Archive Wayback Machine entry for https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1110.pdf. I secured these screen captures myself from https://web.archive.org/web/20210122095202/https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1110.pdf. This record was captured as of January 22, 2021.

19. Appendix ZHAO02 is an Internet Archive capture of a file entitled "L2-ARCTIC: A Non-Native English Speech Corpus" as hosted on the ISCA Speech website. I secured this file by clicking the link entitled ‘publication’ as found in Appendix ZHAO01. In comparing Appendix ZHAO02 to Exhibit 1005, it is my determination that Exhibit 1005 is a true and correct copy of the "L2-ARCTIC: A Non-Native English Speech Corpus" web page as hosted on the ISCA Speech website. As the Internet Archive first captured this page as of January 22, 2021, it is my determination that “L2-ARCTIC: A Non-Native English Speech Corpus” was first made available at least as early as of

January 22, 2021 via the Internet Archive and likely earlier via the original ISCA Speech website.

20. Attached hereto as Appendix ZHAO03 is a screen capture of the Internet Archive Wayback Machine entry for https://isca-speech.org/archive/Interspeech_2018/. I secured these screen captures myself from https://web.archive.org/web/20190920100632/https://isca-speech.org/archive/Interspeech_2018/. This record was captured as of September 20, 2019.

21. Appendix ZHAO03 is an Internet Archive capture of a web page entitled Interspeech 2018 as hosted on the ISCA Speech website. Appendix ZHAO03 is a table of contents for the Proceedings of the 2018 Interspeech conference. Page 27 of Appendix ZHAO03 contains an entry for “L2-ARCTIC: A Non-Native English Speech Corpus”, indicating this paper was included within the conference proceedings for the 2018 Interspeech Conference. This entry also contains a link to download a PDF copy of this paper. As the Internet Archive first captured this page as of September 20, 2019, it is my determination that Interspeech 2018 and therefore “L2-ARCTIC: A Non-Native English Speech Corpus” was first made available to the public at least as early as of September 20, 2019 via the ISCA Speech website.

22.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.

23.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: 3/24/2026

A handwritten signature in black ink, appearing to read 'June Ann Munford', written in a cursive style.

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. Wolfsped (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 Eglia 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)

- Wolfsped

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 LIU01



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN UPLOAD

Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER



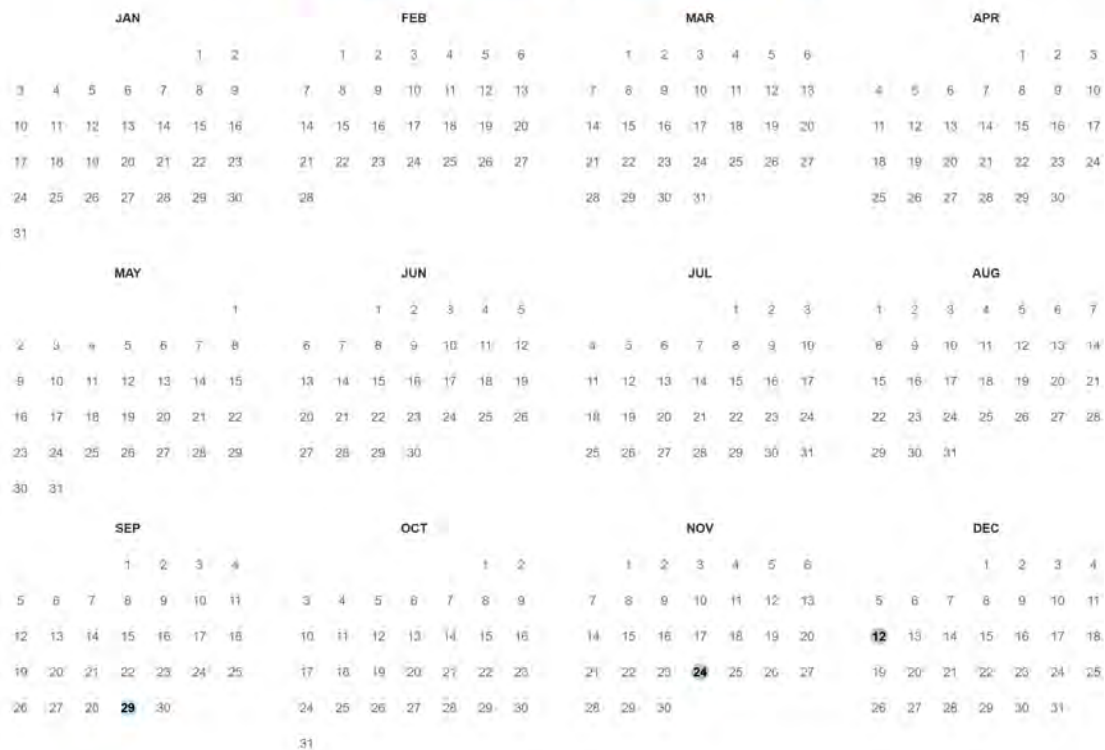
Explore more than 1 trillion web pages saved over time

[Calendar](#) · [Collections](#) · [Changes](#) · [Summary](#) · [Site Map](#) · [URLs](#)

Saved 25 times between September 29, 2021 and March 6, 2026.



Wed, 29 Sep 2021 01:17:11 GMT (why: archiveteam, archiveteam_urls)



Note

This calendar view maps the number of times <https://arxiv.org/abs/2109.13731> 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) nonprofit 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](#).



Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > cs > arXiv:2109.13731

Search...

All fields

Search

[Help](#) | [Advanced Search](#)

Computer Science > Sound

[Submitted on 28 Sep 2021 (v1), last revised 5 Oct 2021 (this version, v3)]

VoiceFixer: Toward General Speech Restoration with Neural Vocoder

Haohu Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, Yuxuan Wang

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on single-task speech restoration (SSR), such as speech denoising or speech clipping. However, SSR systems only focus on one task and do not address the general speech restoration problem. In addition, previous SSR systems show limited performance in some speech restoration tasks such as speech super-resolution. To overcome those limitations, we propose a general speech restoration (GSR) task that attempts to remove multiple distortions simultaneously. Furthermore, we propose VoiceFixer, a generative framework to address the GSR task. VoiceFixer consists of an analysis stage and a synthesis stage to mimic the speech analysis and comprehension of the human auditory system. We employ a ResUNet to model the analysis stage and a neural vocoder to model the synthesis stage. We evaluate VoiceFixer with additive noise, room reverberation, low-resolution, and clipping distortions. Our baseline GSR model achieves a 0.499 higher mean opinion score (MOS) than the speech enhancement SSR model. VoiceFixer further surpasses the GSR baseline model on the MOS score by 0.256. Moreover, we observe that VoiceFixer generalizes well to severely degraded real speech recordings, indicating its potential in restoring old movies and historical speeches. The source code is available at [this https URL](#).

Subjects: **Sound (cs.SD)**; Audio and Speech Processing (eess.AS)Cite as: [arXiv:2109.13731 \[cs.SD\]](#)(or [arXiv:2109.13731v3 \[cs.SD\]](#) for this version)

Submission history

From: Haohu Liu [[view email](#)][\[v1\]](#) Tue, 28 Sep 2021 13:51:16 UTC (24,169 KB)[\[v2\]](#) Mon, 4 Oct 2021 16:58:03 UTC (24,172 KB)[\[v3\]](#) Tue, 5 Oct 2021 15:52:27 UTC (24,324 KB)

Bibliographic Tools

[Code & Data](#)[Related Papers](#)[About arXivLabs](#)

Bibliographic and Citation Tools

- ☐ Bibliographic Explorer ([What is the Explorer?](#))
- ☐ Litmaps ([What is Litmaps?](#))
- ☐ scite Smart Citations ([What are Smart Citations?](#))

[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))

Download:

- PDF
- Other formats (license)

Current browse context:

cs.SD

< prev | next >
new | recent | 2109

Change to browse by:

cs
eess
eess.AS

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

Bookmark

[About](#)[Help](#)[Contact](#)[Subscribe](#)[Copyright](#)[Privacy Policy](#)[Web Accessibility Assistance](#)[arXiv Operational Status](#)Get status notifications via [email](#) or [slack](#)

Appendix LIU02



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN UPLOAD

Search

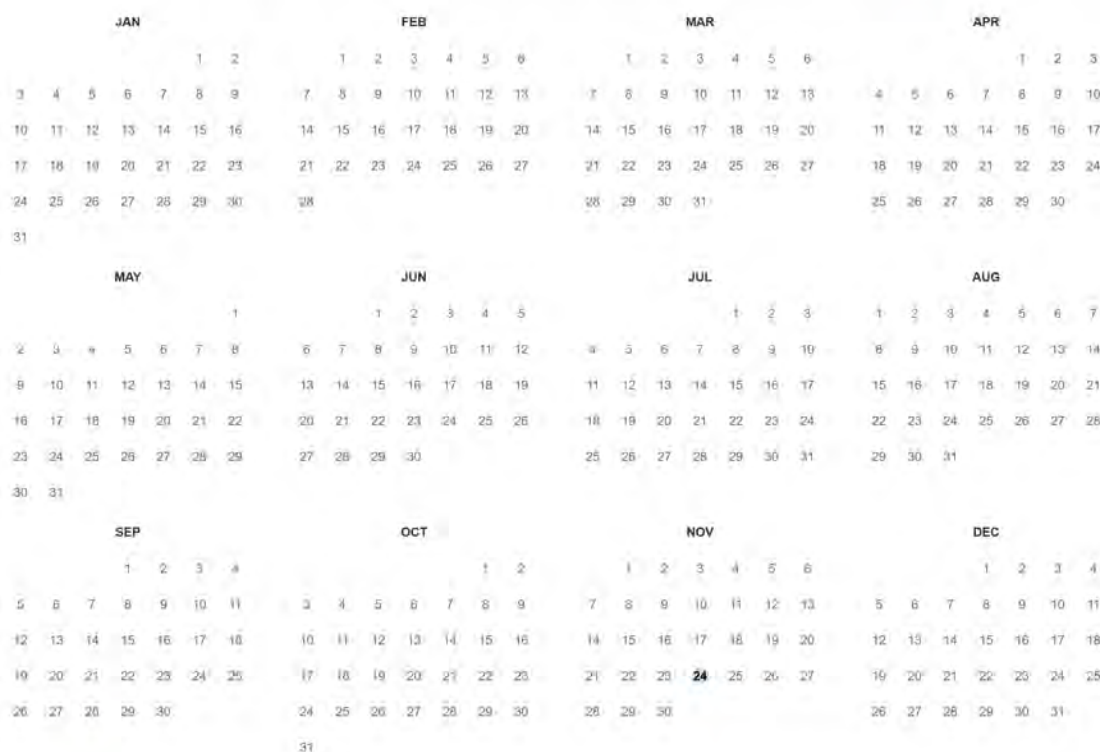
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER



Explore more than 1 trillion web pages saved over time

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

Saved 1 time November 24, 2021



Note

This calendar view maps the number of times <https://arxiv.org/pdf/2109.13731.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](#).

VOICEFIXER: TOWARD GENERAL SPEECH RESTORATION WITH NEURAL VOCODER

Haohe Liu^{1,2*}, Qiuqiang Kong¹, Qiao Tian¹, Yan Zhao¹,
DeLiang Wang², Chuanzeng Huang¹, Yuxuan Wang¹

¹ Speech, Audio and Music Intelligence (SAMI), ByteDance

² Department of Computer Science and Engineering, The Ohio State University

ABSTRACT

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on *single-task speech restoration* (SSR), such as speech denoising or speech declipping. However, SSR systems only focus on one task and do not address the general speech restoration problem. In addition, previous SSR systems show limited performance in some speech restoration tasks such as speech super-resolution. To overcome those limitations, we propose a *gen-*

SDJ 5 Oct 2021



VOICEFIXER: TOWARD GENERAL SPEECH RESTORATION WITH NEURAL VOCODER

Haohe Liu^{1,2*}, Qiuqiang Kong¹, Qiao Tian¹, Yan Zhao¹,
DeLiang Wang², Chuanzeng Huang¹, Yuxuan Wang¹

¹ Speech, Audio and Music Intelligence (SAMI), ByteDance

² Department of Computer Science and Engineering, The Ohio State University

ABSTRACT

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on *single-task speech restoration* (SSR), such as speech denoising or speech declipping. However, SSR systems only focus on one task and do not address the general speech restoration problem. In addition, previous SSR systems show limited performance in some speech restoration tasks such as speech super-resolution. To overcome those limitations, we propose a *general speech restoration* (GSR) task that attempts to remove multiple distortions simultaneously. Furthermore, we propose *VoiceFixer*¹, a generative framework to address the GSR task. *VoiceFixer* consists of an *analysis stage* and a *synthesis stage* to mimic the speech analysis and comprehension of the human auditory system. We employ a ResUNet to model the analysis stage and a neural vocoder to model the synthesis stage. We evaluate *VoiceFixer* with additive noise, room reverberation, low-resolution, and clipping distortions. Our baseline GSR model achieves a 0.499 higher mean opinion score (MOS) than the speech denoising SSR model. *VoiceFixer* further surpasses the GSR baseline model on the MOS score by 0.256. Moreover, we observe that *VoiceFixer* generalizes well to severely degraded real speech recordings, indicating its potential in restoring old movies and historical speeches. The source code is available at https://github.com/haoheliu/voicefixer_main.

1 INTRODUCTION

Speech restoration is a process to restore degraded speech signals to high-quality speech signals. Speech restoration is an important research topic due to speech distortions are ubiquitous. For example, speech is usually surrounded by background noise, blurred by room reverberations, or recorded by low-quality devices (Godsill et al., 2002). Those distortions degrade the perceptual quality of speech for human listeners. Speech restoration has a wide range of applications such as online meeting (Defossez et al., 2020), hearing aids (Van den Bogaert et al., 2009), and audio editing (Van Winkle, 2008). Still, speech restoration remains a challenging problem due to the large variety of distortions in the world.

Previous works in speech restoration mainly focus on *single task speech restoration* (SSR), which deals with only one type of distortion at a time. For example, speech denoising (Loizou, 2007), speech dereverberation (Naylor & Gaubitch, 2010), speech super-resolution (Kuleshov et al., 2017), or speech declipping (Záviška et al., 2020). However, in the real world, speech signal can be degraded by several different distortions simultaneously, which means previous SSR systems oversimplify the speech distortion types (Kashani et al., 2019; Lin et al., 2021; Kuleshov et al., 2017;

[†]Work done while interning at ByteDance.

¹Restoration samples: <https://haoheliu.github.io/demopage-voicefixer>

Birnbaum et al., 2019). The mismatch between the training data used in SSR and the testing data from the real world degrades the speech restoration performance.

To address the mismatch problem, we propose a new task called general speech restoration (GSR), which aims at restoring multiple distortions in a single model. A numerous studies (Cutler et al., 2021; Cauchi et al., 2014; Han et al., 2015) have reported the benefits of jointly training multiple speech restoration tasks. Nevertheless, performing GSR using one-stage systems still suffer from the problems in each SSR task. For example, on generative tasks such as speech super-

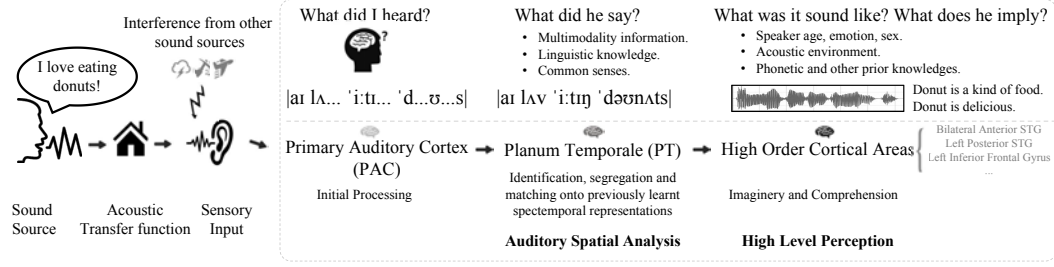


Figure 1: The neural and cognitive model of how human brain understand and restore distorted speech.

The design of *VoiceFixer* is motivated by the biological mechanisms of human hearing when restoring distorted speech. Intuitively, if a person tries to identify a strongly distorted voice, his/her brain can do the recovery by utilizing both the distorted speech signal and the prior knowledge of the language. As shown in Figure 1, the speech distortion perception is modeled by neuroscientists as a two-stage process, including an auditory scene analysis stage (Bregman, 1994), and a high level comprehension/synthesis stage (Griffiths & Warren, 2002). In the analysis stage, the sound information is first transformed into acoustic features by primary auditory cortex (PAC). Then planum temporale (PT), the cortical area posterior to the auditory cortex, acts as a computational hub by segregating and matching the acoustic features to low level spectrotemporal representations. In the synthesis stage, a high order cortical area is hypothesised to perform the high level perception tasks (Griffiths & Warren, 2002; Kennedy-Higgins, 2019). Our proposed *VoiceFixer* systems model the analysis stage with spectral transformations and a deep residual UNet, and the synthesis stage with a convolutional vocoder trained using adversarial losses. One advantage of the two-stage *VoiceFixer* is that the analysis and synthesis stages can be trained separately. Two-stage methods have also been successfully applied to the speech synthesis task (Wang et al., 2016; Ren et al., 2019) where acoustic models and vocoders are trained separately.

VoiceFixer is the first GSR model that is able to restore a wide range of low-resolution speech sampled from 2 kHz to 44.1 kHz, which is different from previous studies working on constant sampling rates (Lim et al., 2018; Wang & Wang, 2021; Lee & Han, 2021). To the best of our knowledge, *VoiceFixer* is the first model that jointly performs speech denoising, speech dereverberation, speech super-resolution, and speech declipping in a unified model.

The rest of this paper is organized as follows. Section 2 introduces the formulations of speech distortions. Section 3 describes the design of *VoiceFixer*. Section 4 discusses the evaluation results. Appendixes introduce related works and show speech restoration demos.

2 PROBLEM FORMULATIONS

We denote a segment of a speech signal as $s \in \mathbb{R}^L$, where L is the samples number in the segment. We model the distortion process of the speech signal as a function $d(\cdot)$. The degraded speech $x \in \mathbb{R}^L$ can be written as:

$$x = d(s). \quad (1)$$

Speech restoration is a task to restore high-quality speech $\hat{s}$ from x :

$$\hat{s} = f(\mathbf{x}) \quad (2)$$

where $f(\cdot)$ is the restoration function and can be viewed as a reverse process of $d(\cdot)$. The target is to estimate $\mathbf{s}$ by restoring $\hat{s}$ from the observed speech $\mathbf{x}$. Recently, several deep learning based one-stage methods have been proposed to model $f(\cdot)$ such as fully connected neural networks, recurrent neural networks, and convolutional neural networks. Detailed introductions can be found in Appendix A.

Distortion modeling is an important step to simulate distorted speech when building speech restoration systems. Several previous works model distortions in a sequential order (Vincent et al., 2017; Tan et al., 2020; Zhao et al., 2019). Similarly, we model the distortion $d(\cdot)$ as a composite function:

$$d(\mathbf{x}) = d_1 \circ d_2 \circ \dots \circ d_Q(\mathbf{x}), d_q \in \mathbb{D}, q = 1, 2, \dots, Q, \quad (3)$$

where $\circ$ stands for function composition and Q is the number of distortions to consist $d(\cdot)$. Set $\mathbb{D} = \{d_v(\cdot)\}_{v=1}^V$ is the set of distortion types where V is the total number of types. Equation 3 describes the procedure of compounding different distortions from $\mathbb{D}$ in a sequential order. We introduce four speech distortions as follows.

Additive noise is one of the most common distortion and can be modeled by the addition between speech $\mathbf{s}$ and noise $\mathbf{n} \in \mathbb{R}^L$:

$$d_{\text{noise}}(\mathbf{s}) = \mathbf{s} + \mathbf{n}. \quad (4)$$

Reverberation is caused by the reflections of signal in a room. Reverberation makes speech signals sound distant and blurred. It can be modeled by convolving speech signals with a room impulse response filter (RIR) $\mathbf{r}$:

$$d_{\text{rev}}(\mathbf{s}) = \mathbf{s} * \mathbf{r}, \quad (5)$$

where $*$ stands for convolution operation.

Low-resolution distortions refer to audio recordings that are recorded in low sampling rates or with limited bandwidth. There are many causes for low-resolution distortions. For example, when microphones have low responses in high-frequency, or audio recordings are compressed to low sampling rates, the high frequencies information will be lost. We follow the description in Wang & Wang (2021) to produce low-resolution distortions but add more filter types (Sulun & Davies, 2020). After designing a low pass filter $\mathbf{h}$, we first convolve it with $\mathbf{s}$ to avoid the aliasing phenomenon. Then we perform resampling on the filtered result from the original sampling rate o to a lower sampling rate u :

$$d_{\text{low_res}}(\mathbf{s}) = \text{Resample}(\mathbf{s} * \mathbf{h}, o, u), \quad (6)$$

Clipping distortions refer to the clipped amplitude of audio signals, which are usually caused by low-quality microphones. Clipping can be modeled by restricting signal amplitudes within $[-\eta, +\eta]$:

$$d_{\text{clip}}(\mathbf{s}) = \max(\min(\mathbf{s}, \eta), -\eta), \eta \in [0, 1]. \quad (7)$$

In the frequency domain, the clipping effect produces harmonic components in the high-frequency part and degrades speech intelligibility accordingly.

3 METHODOLOGY

3.1 ONE-STAGE SPEECH RESTORATION MODELS

Previous deep learning based speech restoration models are usually in one stage. That is, a model predicts restored speech $\hat{s}$ from input $\mathbf{x}$ directly:

$$f : \mathbf{x} \rightarrow \hat{s}. \quad (8)$$

The mapping function $f(\cdot)$ can be modeled by time domain speech restoration systems such as one-dimensional convolutional neural networks (Luo & Mesgarani, 2019) or frequency domain systems such as mask-based (Narayanan & Wang, 2013) methods:

$$\hat{S} = (F_{\text{sp}}(|\mathbf{X}|; \theta) \odot (|\mathbf{X}| + \epsilon))e^{j\angle \mathbf{X}}. \quad (9)$$

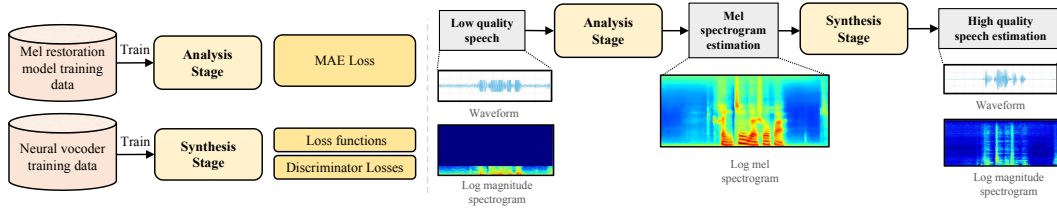


Figure 2: Overview of the proposed *VoiceFixer* system.

where $\mathbf{X}$ is the short-time fourier transform (STFT) of x and ϵ is a small positive constant. $\mathbf{X}$ has a shape of $T \times F$ where T is the number of frame and F is the number of frequency bins. The output of the mask estimation function $F(\cdot; \theta)$ is multiplied by the magnitude spectrogram $|\mathbf{X}|$ to produce the target spectrogram estimation $\hat{\mathbf{S}}$. Then, inverse short-time fourier transform (iSTFT) is applied on $\hat{\mathbf{S}}$ to obtain $\hat{s}$. The one-stage speech restoration models are typically optimized by minimizing the mean absolute error (MAE) loss between the estimated spectrogram $\hat{\mathbf{S}}$ and the target spectrogram $\mathbf{S}$:

$$\mathcal{L} = \left\| |\hat{\mathbf{S}}| - |\mathbf{S}| \right\|_1 \quad (10)$$

Previous one-stage models usually build on high-dimensional features such as time samples and the STFT spectrograms. However, Kuo & Sloan (2005) point out that the high-dimensional features will lead to exponential growth in search space. The model can work on the high-dimensional features under the premise of enlarging the model capacity but may also fail in challenging tasks. Therefore, it would be beneficial if we could build a system on more delicate low-dimensional features.

3.2 VOICEFIXER

In this study, we propose *VoiceFixer*, a two-stage speech restoration framework. Multi-stage methods have achieved state-of-the-art performance in many speech processing tasks (Jarrett et al., 2009; Tan et al., 2020). In speech restoration, our proposed *VoiceFixer* breaks the conventional one-stage system into a two-stage system:

$$f : x \mapsto z, \quad (11)$$

$$g : z \mapsto \hat{s}. \quad (12)$$

Equation 11 denotes the analysis stage of *VoiceFixer* where a distorted speech x is mapped into a representation z . Equation 12 denotes the synthesis stage of *VoiceFixer*, which synthesize z to the restored speech $\hat{s}$. Through the two-stage processing, *VoiceFixer* mimics the human perception of speech described in Section 1.

3.2.1 ANALYSIS STAGE

The goal of the analysis stage is to predict the intermediate representation z , which can be used later to recover the speech signal. In our study, we choose mel spectrogram as the intermediate representation. Mel spectrogram has been widely used in many speech processing tasks (Shen et al., 2018; Kong et al., 2019; Narayanan & Wang, 2013). The frequency dimension of mel spectrogram is usually much smaller than that of STFT thus can be regarded as a way of feature dimension reduction. So, the objective of the analysis stage becomes to restore mel spectrograms of the target signals. The mel spectrogram restoration process can be written as the following equation,

$$\hat{\mathbf{S}}_{\text{mel}} = f_{\text{mel}}(\mathbf{X}_{\text{mel}}; \alpha) \odot (\mathbf{X}_{\text{mel}} + \epsilon), \quad (13)$$

where $\mathbf{X}_{\text{mel}}$ is the mel spectrogram of x . It is calculated by $\mathbf{X}_{\text{mel}} = |\mathbf{X}| \mathbf{W}$ where $\mathbf{W}$ is a set of mel filter banks with shape of $F \times F'$. The columns of $\mathbf{W}$ are not divided by the width of their mel bands, i.e., area normalization, because this will make the restoration model difficult to recover the high-frequency part. The mapping function $f_{\text{mel}}(\cdot; \alpha)$ is the mel restoration mask estimation model parameterized by α . The output of f_{mel} is multiplied by $\mathbf{X}_{\text{mel}}$ to predict the target mel spectrogram.

We use ResUNet (Kong et al., 2021a) to model the analysis stage as shown in Figure 3a, which is an improved UNet (Ronneberger et al., 2015). The ResUNet consists of several encoder and

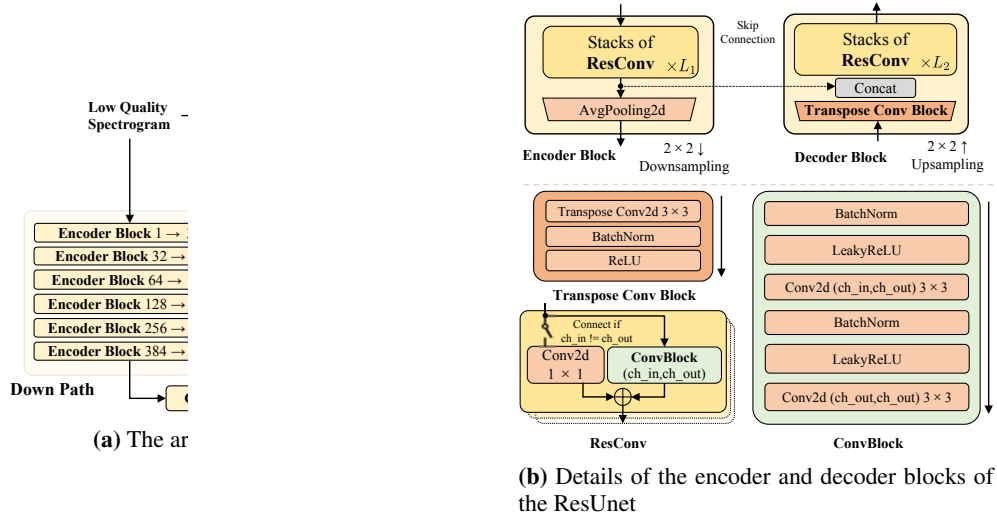


Figure 3: The architecture of ResUNet, which output has the same size as input.

decoder blocks. There are skip connections between encoder and decoder blocks at the same level. As is shown in Figure 3b, both encoder and decoder block share the same structure, which is a series of residual convolutions (ResConv). Each convolutional layer in ResConv consists of a batch normalization (BN), a leakyReLU activation, and a linear convolutional operation. The encoder blocks apply average pooling for downsampling. The decoder blocks apply transpose convolution for upsampling. In addition to ResUNet, we implement the analysis stage with fully connected deep neural network (DNN), and bidirectional gated recurrent units (BiGRU) (Chung et al., 2014) for comparison. The DNN consists of six fully connected layers. The BiGRU has similar structures with DNN except for replacing the last two layers of DNN into bi-directional GRU layers.

The details of these three models are discussed in Appendix B.1. We will refer to ResUNet as UNet later for abbreviation. We optimize the model in the analysis stage using the MAE loss between the estimated mel spectrogram $\hat{S}_{\text{mel}}$ and the target mel spectrogram S_{mel} :

$$\mathcal{L}_{\text{ana}} = \left\| \hat{S}_{\text{mel}} - S_{\text{mel}} \right\|_1 \quad (14)$$

3.2.2 SYNTHESIS STAGE

The synthesis stage is realized by a neural vocoder that synthesizes the mel spectrogram into waveform as denoted in Equation 15:

$$\hat{s} = g(\mathbf{X}_{\text{mel}}; \beta), \quad (15)$$

where $g(\cdot; \beta)$ stands for the vocoder model parameterized by β . We employ a recently proposed non-autoregressive model, time and frequency domain based generative adversarial network (TFGAN), as the vocoder.

Figure 4 shows the detailed architecture of *TFGAN*, in which the input mel spectrogram $\mathbf{X}_{\text{mel}}$ will first pass through a condition network *CondNet*, which contains N_1 one-dimensional convolution layers with exponential linear unit activations (Clevert et al., 2015). Then, in *UpNet*, the feature is upsampled N_2 times with ratios of $s_0, s_1, \dots$, and s_{N_2-1} using *UpsampleBlock* and *ResStacks*. Within the *UpsampleBlock*, the input is first passed through a leakyReLU activation and then fed into a sinusoidal function, which output is added to its input to remove periodic artifacts in breathing part of speech. Then, the output is bifurcated into two branches for upsampling. One branch repeats the samples s_n times followed by a one-dimensional convolution. The other branch uses a stride s_n transpose convolution. The output of the repeat and transpose convolution branches are added together as the output of *UpsampleBlock*. *ResStacks* module contains two dilated convolution layers with leakyReLU activations. The exponentially growing dilation in *ResStack* enable the model to capture long range dependencies. The *TFGAN* in our synthesis model applies $N_2 = 4$. After four

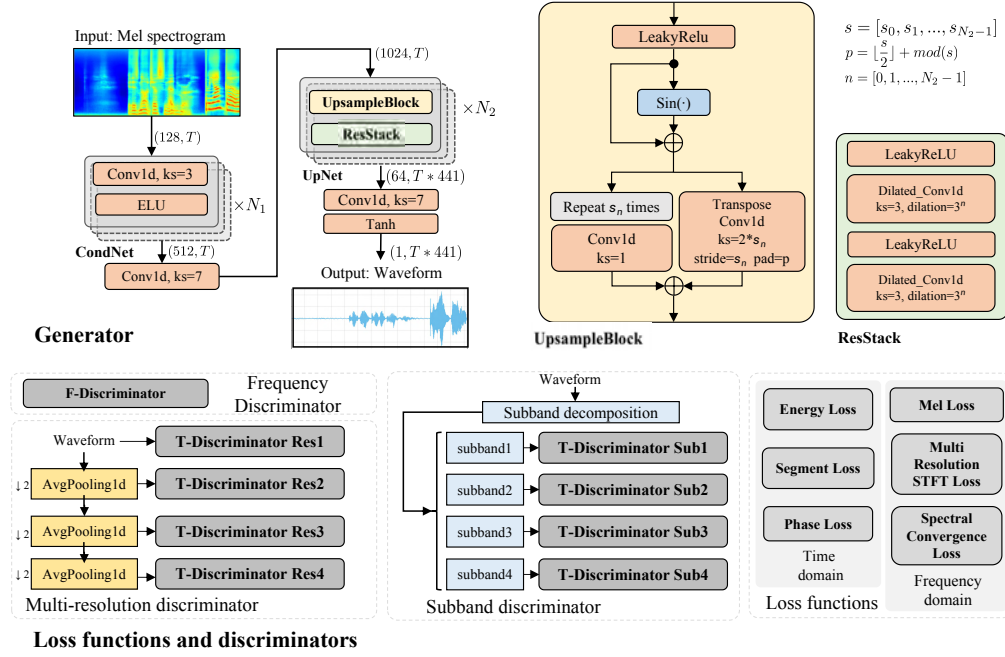


Figure 4: The architecture and training scheme of *TFGAN*, whose generator is later used as vocoder. The generator takes mel spectrogram as input and upsamples it into waveform. Both output waveform and its STFT spectrogram are used to compute loss. We employ both time and frequency discriminators for discriminative training.

UpsampleBlock blocks with ratios $[7, 7, 3, 3]$, each frame of the mel spectrogram is transformed into a sequence with 441 samples corresponding to 10 ms of audio sampled at 44.1 kHz.

The training criteria of the vocoder consist of frequency domain loss $\mathcal{L}_F$, time domain loss $\mathcal{L}_T$, and weighted discriminator loss $\mathcal{L}_D$:

$$\mathcal{L}_{\text{syn}} = \mathcal{L}_F + \mathcal{L}_T + \lambda_D \mathcal{L}_D, \quad (16)$$

The frequency domain loss $\mathcal{L}_F$ is the combination of a mel loss $\mathcal{L}_{\text{mel}}$ and multi-resolution spectrogram losses:

$$\mathcal{L}_F(\hat{s}, s) = \lambda_{\text{mel}} \mathcal{L}_{\text{mel}}(\hat{s}, s) + \sum_{k=1}^{K_F} (\lambda_{\text{sc}} \mathcal{L}_{\text{sc}}^{(k)}(\hat{s}, s) + \lambda_{\text{mag}} \mathcal{L}_{\text{mag}}^{(k)}(\hat{s}, s)) \quad (17)$$

where $\mathcal{L}_{\text{sc}}$ and $\mathcal{L}_{\text{mag}}$ are the spectrogram losses calculated in the linear and log scale, respectively. There are K_F different window sizes ranging from 64 to 4096 to calculate $\mathcal{L}_{\text{sc}}$ and $\mathcal{L}_{\text{mag}}$ so that the trained vocoder is tolerant over phase mismatch (Yamamoto et al., 2020; Juvela et al., 2019). Table 2 in Appendix B.2 shows the detailed configurations.

Time domain loss is complementary to frequency domain loss to address problems such as periodic artifacts. Time domain loss combines segment loss $\mathcal{L}_{\text{seg}}^{(k)}$, energy loss $\mathcal{L}_{\text{energy}}^{(k)}$ and phase loss $\mathcal{L}_{\text{phase}}^{(k)}$:

$$\mathcal{L}_T(\hat{s}, s) = \sum_{k=1}^{K_T} (\lambda_{\text{seg}} \mathcal{L}_{\text{seg}}^{(k)}(\hat{s}, s) + \lambda_{\text{energy}} \mathcal{L}_{\text{energy}}^{(k)}(\hat{s}, s) + \lambda_{\text{phase}} \mathcal{L}_{\text{phase}}^{(k)}(\hat{s}, s)) \quad (18)$$

where segment loss $\mathcal{L}_{\text{seg}}^{(k)}$, energy loss $\mathcal{L}_{\text{energy}}^{(k)}$ and phase loss $\mathcal{L}_{\text{phase}}^{(k)}$ are described in Equation 24, 25, and 26 of Appendix B.2. There are K_T different window sizes ranging from 1 to 960 to calculate time domain loss at different resolutions. The details of window sizes are shown in Table 3 of Appendix B.2. The energy loss and phase loss have the advantage of alleviating metallic sounds.

Discriminative training is an effective way to train neural vocoders (Kong et al., 2020; Kumar et al., 2019). In our study, we utilize a group of discriminators, including a multi-resolution time discrim-

inator D_T , a subband discriminator D_{sub} , and frequency discriminator D_F :

$$D(\mathbf{s}) = \sum_{r=1}^{R_T} D_T^{(r)}(\mathbf{s}) + D_{sub}(\mathbf{s}) + D_F(\mathbf{s}) \quad (19)$$

$$\mathcal{L}_D(\mathbf{s}, \hat{\mathbf{s}}) = \min_g \max_D (\mathbb{E}_{\mathbf{s}}(\log(D(\mathbf{s}))) + \mathbb{E}_{\hat{\mathbf{s}}}(\log(1 - D(\hat{\mathbf{s}})))) \quad (20)$$

The multi-resolution discriminators D_T take signals from R_T kinds of time resolutions after average pooling as input. The subband discriminator D_{sub} performs subband decomposition (Liu et al., 2020) on the waveform, producing four subband signals to feed into four *T-discriminators*, respectively. Frequency discriminator D_F takes the linear spectrogram as input and outputs real or fake labels. The bottom part of Figure 4 shows the main idea of *T-discriminator* and *F-discriminator*. Appendix B.2 describes the detailed discriminator architectures.

There are two advantages of using neural vocoder in the synthesis stage. First, neural vocoder trained using a large amount of speech data contains prior knowledge on the structural distribution of speech signals, which is crucial to the restoration of distorted speech. The amount of training data of vocoder is more than that used in conventional SSR methods with limited speaker numbers. Second, the neural vocoder typically takes the mel spectrogram as input, resulting in fewer feature dimensions than the STFT features. The reduction in dimension helps to lower computational costs and achieve better performance in the analysis stage.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION METRICS

Training sets The training speech datasets we use including VCTK (Yamagishi et al., 2019), AISHELL-3 (Shi et al., 2020), and HQ-TTS. We call the noise datasets used for training as VD-Noise. To simulate the reverberations, we employ a set of RIRs to create an RIR-44k dataset. We use VCTK, VD-Noise, and the training part of RIR-44k to train the analysis stage. AISHELL-3, VCTK, and HQ-TTS datasets are used to train the vocoder in the synthesis stage. The details of those datasets and the RIRs simulation configurations are discussed in Appendix C.1.

Test sets We employ VCTK-Demand (Valentini-Botinhao et al., 2017) as the denoising test set and name it as DENOISE. We call our speech super-resolution, declipping, and dereverberation evaluation test sets as SR, DECLI, and DEREV, respectively. In addition, we create an ALL-GSR test set containing all distortions. We introduce the details of how we build these test sets in Appendix C.3.

Evaluation metrics The metrics we adopt include log-spectral distance (LSD) (Erell & Weintraub, 1990), wide band perceptual evaluation of speech quality (PESQ-wb) (Rix et al., 2001), structural similarity (SSIM) (Wang et al., 2004), and scale-invariant signal to noise ratio (SiSNR) (Le Roux et al., 2019). We use mean opinion scores (MOS) to subjectively evaluate different systems.

The output of the vocoder is not strictly aligned on sample level with the target, as is often the case in generative model (Kumar et al., 2020). This effect will degrade the metrics, especially for those calculated on time samples such as the SiSNR. So, to compensate the SiSNR, we design a similar metrics, scale-invariant spectrogram to noise ratio (SiSPNR), to measure the discrepancies on the spectrograms. Details of the metrics are described in Appendix C.4.

4.2 DISTORTIONS SIMULATION

For the SSR task, we perform only one type of distortion for evaluation. For the GSR task, we first assume that $\mathbb{D} = \{d_{\text{noise}}, d_{\text{rev}}, d_{\text{low_res}}, d_{\text{clip}}\}$ because those distortions are the most common distortions in daily environment (Ribas et al., 2016). Second, we assume that $Q \leq 4$ in Equation 3. In other words, each distortion in $\mathbb{D}$ appears at most one time. Then, we generate the distortions following a specific order $d_{\text{rev}}, d_{\text{clip}}, d_{\text{low_res}},$ and d_{noise} . These distortions are added randomly using random configurations.

4.3 BASELINE SYSTEMS

Table 5 in Appendix D summarizes all the experiments in this study. We implement several SSR and GSR systems using one-stage restoration models. For the GSR, we train a ResUNet model called *GSR-UNet* with all distortions. For the SSR models, we implement a *Denoise-UNet* for additive noise distortion, a *Dereverb-UNet* for reverberation distortion, a *SR-UNet* for low-resolution distortion, and a *Declip-UNet* for clipping distortion. For the SR task, we also include two state-of-the-art models, *NuWave* (Lee & Han, 2021) and *SEANet* (Li et al., 2021) for comparison. For declipping task, we compare with a state-of-the-art synthesis-based method *SSPADE* (Záviška et al., 2019a). To explore the impact of model size of the mel restoration model, we setup ResUNets with two sizes, *UNet-S* and *UNet*, which have one and four *ResConv* blocks in each encoder and decoder block, respectively.

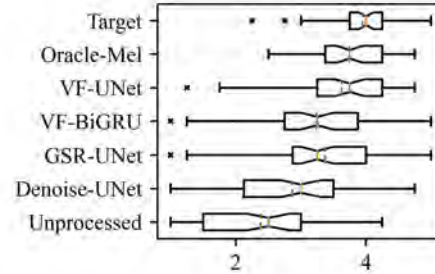
4.4 EVALUATION RESULTS

Neural vocoder To evaluate the performance of the neural vocoder, we compare two baselines. The *Target* system denotes using the perfect s for evaluation. The *Unprocessed* system denotes using distorted speech x for evaluation. The *Oracle-Mel* system denotes using the mel spectrogram of s as input to the vocoder, which marks the performance of the vocoder. As shown in Table 1, the *Oracle-Mel* system achieves a MOS score of 3.74, which is close to the *Target* MOS of 3.95, indicating that the vocoder performs well in the synthesis task.

Models	PESQ	LSD	SiSPNR	SSIM	MOS
Unprocessed	1.94	2.00	7.20	0.64	2.38
Oracle-Mel	2.52	0.91	11.73	0.74	3.74
Target	4.64	0.01	110.55	1.00	3.95
GSR-UNet	2.67	1.01	12.19	0.79	3.37
Denoise-UNet	2.33	1.98	9.65	0.65	2.87
Dereverb-UNet	1.97	1.81	8.50	0.59	/
VF-DNN	1.55	1.18	10.13	0.68	/
VF-BiGRU	1.92	1.02	10.98	0.71	3.24
VF-UNet-S	2.01	1.02	11.09	0.71	/
VF-UNet	2.05	1.01	11.14	0.71	3.62

Table 1: Average PESQ, LSD, SiSPNR, SSIM and MOS scores on the general speech restoration test set, ALL-GSR, which includes random distortions.

Figure 5: Box plot of the MOS scores on general speech restoration task. Red solid line and green dashed line represent median and mean value.



General speech restorations Table 1 shows the evaluation results on ALL-GSR test set. Figure 5 shows the box plot of the MOS scores of these systems. The *GSR-UNet* outperforms the two SSR models, *Denoise-UNet* and *Dereverb-UNet* by a large margin. It surpasses *Denoise-UNet* model by 0.5 on MOS score, which suggests the GSR model is more powerful than the SSR model on this test set. For convenience, we denote *VoiceFixer* as *VF* in tables and figures. We observe that the *VF-UNet* model achieves the highest MOS score and LSD score. Specifically, *VF-UNet* obtains 0.256 higher MOS score than that of *GSR-UNet*. This result indicates that *VoiceFixer* is better than ResUNet based one-stage model on overall quality. What's more, we notice that the MOS score of *VF-UNet* is only 0.11 lower than the *Oracle-Mel*, demonstrating the good performance of the analysis stage. Among the *VoiceFixer* analysis models, the *UNet* front-end achieves the best. The *VF-BiGRU* model achieves similar subjective metrics with the *VF-UNet* model but has much lower MOS scores. This phenomenon shows that the improvement in subjective metrics in *VoiceFixer* is not always consistent with objective evaluation results.

Super-resolution Table 6 in Appendix D.1 shows the evaluation results on the super-resolution test set *SR*. For the 2 kHz, 4 kHz, and 8 kHz to 44.1 kHz super-resolution tasks, *VF-UNet* achieves a significantly higher LSD, SiSPNR and SSIM scores than other models. The LSD value of *VF-UNet* in 2 kHz sampling rate is still higher than the 8 kHz sampling rate score of *GSR-UNet*, *SR-UNet*, *NuWave*, and *SEANet*. This demonstrates the strong performance of *VoiceFixer* on dealing with low sampling rate cases. The *VF-BiGRU* model outperforms *VF-UNet-S* model on average scores for its better performance on low upsample-ratio cases. MOS box plot in Figure 6b shows that *VF-UNet* performs the best on 8 kHz to 44.1 kHz test set. Figure 6e shows the MOS score of *Unprocessed*

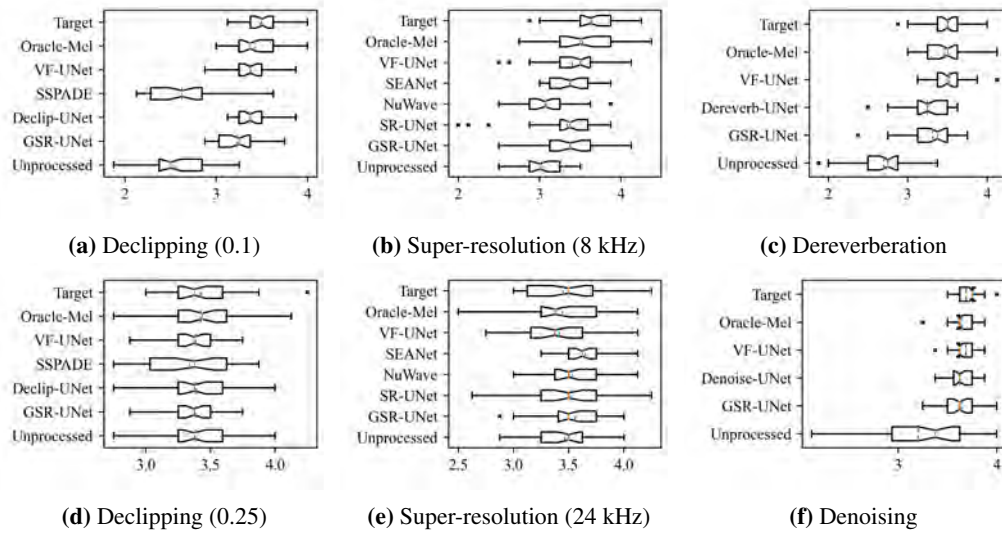


Figure 6: Box plot of the MOS scores on speech super-resolution, declipping, dereverberation and denoising.

is close to *Target* on 24 kHz to 44.1 kHz test set, meaning limited perceptual difference between the two sampling rates. On this test set, *SEANet* even achieves a higher MOS score than *Target*. That’s due to the high-frequency part it generate contains more energy than that of *Target*, making the results sound clearer.

Denoising We evaluate the speech denoising performance on the DENOISE test set and show results in Table 7 in Appendix D.1. We find that *GSR-UNet* preserves more details in the high-frequency part and has better PESQ and SiSPNR values than the denoising only SSR model *Denoise-UNet*. The reason might be that speech data augmentations and jointly performing super-resolution task can increase the generalization and inpainting ability of the model (Hao et al., 2020). The PESQ score of *VF-UNet* reaches 2.43, higher than *SEGAN*, *WaveUNet*, and the model trained with weakly labeled data in Kong et al. (2021b). The MOS evaluations in Figure 6f on speech denoising task also demonstrate that the result of *VF-UNet* sound comparable with one-stage speech denoising models.

Declipping and dereverberation Table 9 and Table 8 in Appendix D.1 show similar performance trends on the speech declipping and speech dereverberation. In both tasks, the SSR model *Dereverb-UNet* and *Declip-UNet* achieve the highest scores. The performance of *GSR-UNet* is slightly worse, but it is acceptable considering that *GSR-UNet* does not need extra training for each task. *SSPADE* performs better on SiSNR, but the PESQ and STOI scores are lower, especially in the 0.1 threshold case. The MOS score in Figure 6d shows that the clipping effect in the 0.25 threshold case is not easy to perceive, leading to high MOS scores across all methods. In Figure 6a, both *Declip-UNet* and *VF-UNet* achieve the highest objective scores on the 0.1 threshold clipping test set. On the dereverberation test set DEREV, *VF-UNet* achieves the highest MOS score 3.52.

5 CONCLUSIONS

In this work, we propose *VoiceFixer*, an effective approach for general speech restoration. *VoiceFixer* consists of an analysis stage modeled by a ResUNet and a synthesis stage modeled by a neural vocoder. The evaluation results show that *VoiceFixer* achieves leading performance across general speech restoration, speech super-resolution, speech denoising, speech dereverberation, and speech declipping tasks. In the future, we will extend *VoiceFixer* to restore general audio signals, including music and general sounds.

REPRODUCIBILITY STATEMENT

We make our code and datasets downloadable for painless reproducibility. Our pre-trained *Voice-Fixer* and inference code are presented in <https://github.com/haoheliu/voicefixer>. The code for performing the experiments discussed in Section 4 is downloadable in https://github.com/haoheliu/voicefixer_main. The experiment code can conduct evaluations and generate reports on the metrics mentioned in Section 4.1 automatically. The *NuWave* is realized using the code open-sourced by Lee & Han (2021): <https://github.com/mindslab-ai/nuwave>. We reproduce *SSPADE* using the toolbox provided by Závíška et al. (2020) at <https://rajmic.github.io/declipping2020>. Besides, we upload the speech and noise training set, VCTK and VD-Noise, to <https://zenodo.org/record/5528132>, the RIR-44k dataset to <https://zenodo.org/record/5528124>, and the test sets to <https://zenodo.org/record/5528144>. The AISHELL-3 dataset is open-sourced at http://www.aishelltech.com/aishell_3. The HQ-TTS is a collection of datasets from openslr.org, as is described in Table 4 of Appendix C.1.

REFERENCES

- Fanhu Bie, Dong Wang, Jun Wang, and Thomas Fang Zheng. Detection and reconstruction of clipped speech for speaker recognition. *Speech Communication*, pp. 218–231, 2015.
- Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei Koh, and Stefano Ermon. Temporal film: Capturing long-range sequence dependencies with feature-wise modulations. *arXiv preprint arXiv:1909.06628*, 2019.
- Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.
- Benjamin Cauchi, Ina Kodrasi, Robert Rehr, Stephan Gerlach, Ante Jukic, Timo Gerkmann, Simon Doclo, and Stefan Goetze. Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme. In *Proc. REVERB Challenge Workshop*, pp. 1–8, 2014.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units. *arXiv preprint arXiv:1511.07289*, 2015.
- Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Sten Sootla, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, Robert Aichner, et al. Acoustic echo cancellation challenge. In *INTERSPEECH*, 2021.
- Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the wave-form domain. *arXiv preprint arXiv:2006.12847*, 2020.
- Adoram Erell and Mitch Weintraub. Estimation using log-spectral-distance criterion for noise-robust speech recognition. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 853–856, 1990.
- Simon Godsill, Peter Rayner, and Olivier Cappé. Digital audio restoration. In *Applications of Digital Signal Processing to Audio and Acoustics*, pp. 133–194. Springer, 2002.
- Timothy D Griffiths and Jason D Warren. The planum temporale as a computational hub. *Trends in Neurosciences*, pp. 348–353, 2002.
- Archit Gupta, Brendan Shillingford, Yannis Assael, and Thomas C Walters. Speech bandwidth extension with wavenet. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 205–208, 2019.
- Kun Han, Yuxuan Wang, DeLiang Wang, William S Woods, Ivo Merks, and Tao Zhang. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 982–992, 2015.

- Xiang Hao, Xiangdong Su, Shixue Wen, Zhiyu Wang, Yiqian Pan, Feilong Bao, and Wei Chen. Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 6959–6963, 2020.
- Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020.
- Kevin Jarrett, Koray Kavukcuoglu, Marc’Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *International Conference on Computer Vision*, pp. 2146–2153, 2009.
- Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, and Paavo Alku. GELP: GAN-Excited linear prediction for speech synthesis from mel-spectrogram. *arXiv preprint arXiv:1904.03976*, 2019.
- Thomas Kailath. Lectures on Wiener and Kalman filtering. In *Lectures on Wiener and Kalman Filtering*, pp. 1–143. 1981.
- Hamidreza Baradaran Kashani, Ata Jodeiri, Mohammad Mohsen Goodarzi, and Shabnam Gholam-dokht Firooz. Image to image translation based on convolutional neural network approach for speech declipping. *arXiv preprint arXiv:1910.12116*, 2019.
- Dan Kennedy-Higgins. *Neural and cognitive mechanisms affecting perceptual adaptation to distorted speech*. PhD thesis, University College London, 2019.
- Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Emanuel Habets, Reinhold Haeb-Umbach, Volker Leutnant, Armin Sehr, Walter Kellermann, and Roland Maas. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.
- Sran Kitić, Nancy Bertin, and Rémi Gribonval. Sparsity and cosparsity for audio declipping: a flexible non-convex approach. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 243–250, 2015.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020.
- Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang, and Mark D Plumbley. Sound event detection and time–frequency segmentation from weakly labelled data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 777–787, 2019.
- Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. In *The International Society for Music Information Retrieval*, 2021a.
- Qiuqiang Kong, Haohe Liu, Xingjian Du, Li Chen, Rui Xia, and Yuxuan Wang. Speech enhancement with weakly labelled data from audioset. *arXiv preprint arXiv:2102.09971*, 2021b.
- Juho Kontio, Laura Laaksonen, and Paavo Alku. Neural network-based artificial bandwidth expansion of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 873–881, 2007.
- Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*, 2017.
- Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestein, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019.
- Rithesh Kumar, Kundan Kumar, Vicki Anand, Yoshua Bengio, and Aaron Courville. NU-GAN: High resolution neural upsampling with gan. *arXiv preprint arXiv:2010.11362*, 2020.

- Frances Y Kuo and Ian H Sloan. Lifting the curse of dimensionality. *Notices of the American Mathematical Society*, pp. 1320–1328, 2005.
- Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR–half-baked or well done? In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 626–630, 2019.
- Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321*, 2021.
- Kehuang Li and Chin-Hui Lee. A deep neural network approach to speech bandwidth expansion. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 4395–4399, 2015.
- Yunpeng Li, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek. Real-time speech frequency bandwidth extension. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 691–695, 2021.
- Teck Yian Lim, Raymond A Yeh, Yijia Xu, Minh N Do, and Mark Hasegawa-Johnson. Time-frequency networks for audio super-resolution. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 646–650, 2018.
- Ju Lin, Yun Wang, Kaustubh Kalgaonkar, Gil Keren, Didi Zhang, and Christian Fuegen. A two-stage approach to speech bandwidth extension. *INTERSPEECH*, pp. 1689–1693, 2021.
- Haohe Liu, Lei Xie, Jian Wu, and Geng Yang. Channel-wise subband input for better voice and accompaniment separation on high resolution music. *arXiv preprint arXiv:2008.05216*, 2020.
- Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.
- Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, pp. 1256–1266, 2019.
- Craig Macartney and Tillman Weyde. Improved speech enhancement with the Wave-U-Net. *arXiv preprint arXiv:1811.11307*, 2018.
- Wolfgang Mack and Emanuël AP Habets. Declipping speech using deep filtering. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 200–204, 2019.
- Rainer Martin. Spectral subtraction based on minimum statistics. *Power*, 1994.
- Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. *arXiv preprint arXiv:1807.09840*, 2018.
- Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, pp. 1877–1884, 2016.
- Anna K Nábělek, Tomasz R Letowski, and Frances M Tucker. Reverberant overlap-and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, pp. 1259–1265, 1989.
- Yoshihisa Nakatoh, Mineo Tsushima, and Takeshi Norimatsu. Generation of broadband speech from narrowband speech based on linear mapping. *Electronics and Communications in Japan*, pp. 44–53, 2002.
- Arun Narayanan and DeLiang Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 7092–7096, 2013.
- Patrick A Naylor and Nikolay D Gaubitch. *Speech dereverberation*. Springer Science & Business Media, 2010.

- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- Santiago Pascual, Antonio Bonafonte, and Joan Serra. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.
- Wei Ping, Kainan Peng, Kexin Zhao, and Zhao Song. Waveflow: A compact flow-based model for raw audio. In *International Conference on Machine Learning*, pp. 7706–7716, 2020.
- Adam Polyak, Lior Wolf, Yossi Adi, Ori Kabeli, and Yaniv Taigman. High fidelity speech regeneration with application to speech enhancement. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 7143–7147, 2021.
- Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 3617–3621, 2019.
- Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*, 2019.
- Lucas Rencker, Francis Bach, Wenwu Wang, and Mark D Plumbley. Sparse recovery and dictionary learning from nonlinear compressive measurements. *IEEE Transactions on Signal Processing*, pp. 5659–5670, 2019.
- Dayana Ribas, Emmanuel Vincent, and José Ramón Calvo. A study of speech distortion conditions in real scenarios for speech processing applications. In *IEEE Spoken Language Technology Workshop*, pp. 13–20, 2016.
- Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752, 2001.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.
- Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, pp. 629–632, 1996.
- Boaz Schwartz, Sharon Gannot, and Emanuël AP Habet. Online speech dereverberation using kalman filter and em algorithm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 394–406, 2014.
- Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural TTS synthesis by conditioning wavenet on Mel spectrogram predictions. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 4779–4783, 2018.
- Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.
- Xiaofeng Shu, Yehang Zhu, Yanjie Chen, Li Chen, Haohe Liu, Chuanzeng Huang, and Yuxuan Wang. Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation. *arXiv preprint arXiv:2107.09298*, 2021.
- Serkan Sulun and Matthew EP Davies. On filter generalization for music bandwidth extension using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, pp. 132–142, 2020.
- Ke Tan, Yong Xu, Shi-Xiong Zhang, Meng Yu, and Dong Yu. Audio-visual speech separation and dereverberation with a two-stage multimodal network. *IEEE Journal of Selected Topics in Signal Processing*, pp. 542–553, 2020.

- Qiao Tian, Yi Chen, Zewang Zhang, Heng Lu, Linghui Chen, Lei Xie, and Shan Liu. TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis. *arXiv preprint arXiv:2011.12206*, 2020.
- Cassia Valentini-Botinhao et al. Noisy speech database for training speech enhancement algorithms and TTS models. 2017.
- Tim Van den Bogaert, Simon Doclo, Jan Wouters, and Marc Moonen. Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids. *The Journal of the Acoustical Society of America*, pp. 360–371, 2009.
- Charles Van Winkle. Audio analysis and spectral restoration workflows using adobe audition. In *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice*, 2008.
- Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, pp. 535–557, 2017.
- Heming Wang and DeLiang Wang. Towards robust speech super-resolution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- Hong Wang and Fumitada Itakura. Dereverberation of speech signals based on sub-band envelope estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, pp. 3576–3583, 1991.
- Wenfu Wang, Shuang Xu, and Bo Xu. First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention. In *INTERSPEECH*, pp. 2243–2247, 2016.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, pp. 600–612, 2004.
- Donald S Williamson and DeLiang Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1492–1501, 2017.
- Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.
- Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 6199–6203, 2020.
- Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al. Durian: Duration informed attention network for multimodal synthesis. *arXiv preprint arXiv:1909.01700*, 2019.
- Pavel Závíška, Pavel Rajmic, Ondřej Mokř, and Zdeněk Průša. A proper version of synthesis-based sparse audio declipper. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 591–595, 2019a.
- Pavel Závíška, Pavel Rajmic, and Jiří Schimmel. Psychoacoustically motivated audio declipping based on weighted l1 minimization. In *2019 42nd International Conference on Telecommunications and Signal Processing*, pp. 338–342, 2019b.
- Pavel Závíška, Pavel Rajmic, Alexey Ozerov, and Lucas Rencker. A survey and an extensive evaluation of popular audio declipping methods. *IEEE Journal of Selected Topics in Signal Processing*, pp. 5–24, 2020.
- Yan Zhao, Zhong-Qiu Wang, and DeLiang Wang. Two-stage deep learning for noisy-reverberant speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 53–62, 2019.

A APPENDIX A

A.1 SPEECH RESTORATION TASKS

Speech super-resolution A lot of early studies (Nakatoh et al., 2002; Kontio et al., 2007) break super-resolution (SR) into spectral envelop estimation and excitation generation from the low-resolution part. At that time, the direct mapping from the low-resolution part to the high-resolution feature is not widely explored since the dimension of the high-resolution part is relatively high. Later, deep neural network (Li & Lee, 2015; Kuleshov et al., 2017) is introduced to perform SR. These approaches show better subjective quality comparing with traditional methods. To increase the modeling capacity, *TFilm* (Birnbaum et al., 2019) is proposed to model the affine transformation among each time block. Similarly, *WaveNet* also shows effectiveness in extending the bandwidth of a band-limited speech (Gupta et al., 2019). To utilize the information both from the time and frequency domain, Wang & Wang (2021) propose a time-frequency loss that can yield a balanced performance on different metrics. Recently, *NU-GAN* (Kumar et al., 2020) and *NU-Wave* (Lee & Han, 2021) pushed the target sample rate in SR to high fidelity, namely 48 kHz.

Although employing deep neural networks in BWE shows promising results, the generalization capability of these methods is still limited. For example, previous approaches (Kuleshov et al., 2017) usually train and test models with a fixed setup, i.e., fixing the initial and target sampling rates. However, in real-world applications, speech bandwidth is not usually constant. In addition, since the real high-low quality speech pair is hard to collect, typically, we produces low-quality audio with lowpass or bandpass filters during training. In this case, systems tend to overfit specific filters. As discussed in Sulun & Davies (2020), when the kind of filter used during training and testing differ, the performance can fall considerably. To alleviate filter overfitting, Sulun & Davies (2020) propose to train models with multiple kinds of lowpass filters, by which the unseen filters can be handled properly.

Speech declipping The methods for speech declipping can be categorized as supervised methods and unsupervised methods. The unsupervised, or blind methods usually perform declipping based on some generic regularization and assumption of what natural audio should look like, such as ASPADE (Kitić et al., 2015), dictionary learning (Rencker et al., 2019), and psychoacoustically motivated l1 minimization (Záviška et al., 2019b). The supervised models, mostly based on deep neural network (DNN) (Bie et al., 2015; Mack & Habets, 2019), are usually trained on clipped and unclipped data pairs. For example, Kashani et al. (2019) treat the declipping as an image-to-image translation problem and utilize the *UNet* to perform spectral mapping. Currently, most of the state-of-the-art methods are unsupervised (Záviška et al., 2020) because they are usually designed to work on different kinds of audio, while the supervised model mainly specialized on data similar to its training data. However, Záviška et al. (2020) believes supervised models still have the potential for better declipping performance.

Speech denoising Conventional methods are efficient and effective on dealing with stationary noise, such as spectral subtraction (Martin, 1994) and Wiener and Kalman filtering (Kailath, 1981). By comparison, deep learning based models such as *Conv-TasNet* (Luo & Mesgarani, 2019) show higher subjective score and robustness on complex cases. Recently, new schemes have emerged for training speech denoising models. *SEGAN* (Pascual et al., 2017) tried a generative way. Kong et al. (2021b) achieved a denoising model using only weakly labeled data. And Polyak et al. (2021) realized a denoising model using a regeneration approach.

Speech dereverberation Some of the early methods in speech dereverberation, such as inverse filtering (Naylor & Gaubitch, 2010) and subband envelope estimation (Wang & Itakura, 1991), aiming at deconvolving the reverberant signal by estimating an inverse filter. However, the inverse filter is hard and not robust to estimate accurately. Other methods, like spectral subtraction, is based on an overlap-masking (Nábělek et al., 1989) effect of reverberation. Schwartz et al. (2014) performed dereverberation using Kalman filter and expectation-maximization algorithm. Recently, deep learning based dereverberation methods have emerged as the state-of-the-art. Han et al. (2015) used a fully connected DNN to learn a spectral mapping from reverberant speech to clean speech. In Williamson & Wang (2017), similar to the masking-based denoising methods, they proposed to perform dereverberation using a time-frequency mask.

A.1.1 JOINT RESTORATION AND SYNTHETIC RESTORATION

Joint restoration Many works have adopted the joint restoration approach to improving models. To make the acoustic echo cancellation (AEC) result sound cleaner, *MC-TCN* (Shu et al., 2021) proposed to jointly perform AEC and noise suppression at the same time. *MC-TCN* achieved a mean opinion score of 4.41, outperforming the baseline of AEC Challenge (Cutler et al., 2021) by 0.54. Moreover, in the REVERB challenge (Kinoshita et al., 2013), the test set has both reverberation and noise. Therefore, the proposed methods in the challenge should perform both denoising and dereverberation. In Han et al. (2015), the authors proposed to perform dereverberation and denoising within a single DNN and substantially outperform related methods regarding quality and intelligibility. However, previous joint processing usually involved only two sub tasks. In our study, we joint perform four or more tasks to achieve general restoration.

Synthetic restoration Directly estimate the source signal from the observed mixture is hard especially when the SNR is low. Some studies adopted a regeneration approach. In Polyak et al. (2021), the authors utilized an ASR model, a pitch extraction model, and a loudness model to extract semantic level information from the speaker. Then these features were fed to an encode-decoder network to regenerate the speech signal. To maintain the consistency of speaker characteristics, it used an auxiliary identity network to compute the identity feature. Similar to synthetic speech restoration, TTS can be treated as the regeneration of speech from texts.

A.1.2 NEURAL VOCODER

Vocoder, which maps the encoded speech features to the waveforms, is an indispensable component in speech synthesis. The most widely used input feature for vocoder is mel spectrogram. In recent years, since the emergence of *WaveNet* (Oord et al., 2016), neural network based vocoders demonstrate clear advantages over traditional parametric methods (Morise et al., 2016). Comparing with conventional methods, the quality of *WaveNet* is more closer to the human voice. Later, *WaveRNN* (Yu et al., 2019) is proposed to model the waveform with a single GRU. In this way, *WaveRNN* has much lower complexity comparing with *WaveNet*. However, the autoregressive nature of these models and deep structure make their inference process hard to be paralleled. To address this problem, non-autoregressive models like *WaveGlow* (Prenger et al., 2019) and *WaveFlow* (Ping et al., 2020) were proposed. Afterward, non-autoregressive GAN-based models such as *MelGAN* (Kumar et al., 2019) push the synthesis quality to a comparable level with autoregressive models. Recently, *TFGAN* (Tian et al., 2020) demonstrated strong capability in vocoding. Directed by multiple discriminators and loss functions, *TFGAN* was able to leverage information from both time domain and frequency domain. As a result, the synthesis quality of *TFGAN* is more natural and less metallic comparing with other GAN-based non-autoregressive models. In this work, we realize a universal vocoder based on *TFGAN*, which can reconstruct waveform from mel spectrogram of an arbitrary speaker with good perceptual quality. The pre-trained vocoder is available online to facilitate future studies and reproduce our work.

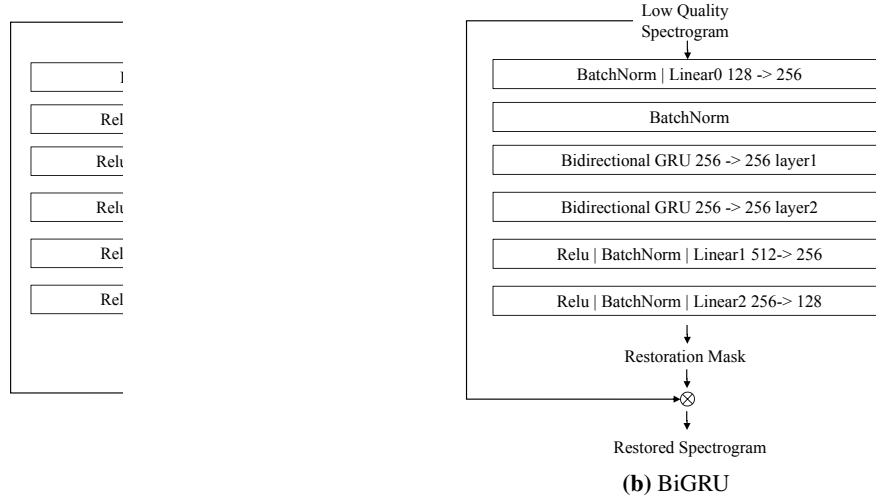


Figure 7: The architecture of DNN and Bi-GRU

The DNN and BiGRU we use are shown in Figure 7. DNN is a six layers fully connected network with BatchNorm and ReLU activations. The DNN accept each time step of the low-quality spectrogram as the input feature and output the restoration mask. Similarly, for the BiGRU model, we substitute some layers in DNN to a two-layer bidirectional GRU to capture the time dependency between time steps. To increase the modeling capacity of BiGRU, we expanded the input dimension of GRU to twice the mel frequency dimension with full connected networks.

The detailed architecture of ResUNet is shown in Figure 3a. In the down-path, the input low-quality mel spectrogram will go through 6 encoder blocks, which includes a stack of L_1 *ResConv* and a 2×2 average pooling. In *ResConv*, the outputs of *ConvBlock* and the residual convolution are added together as the output. *ConvBlock* is a typical two layers convolution with BatchNorm and leakyReLU activation functions. The kernel size of residual convolution and the convolution in *ConvBlock* is 1×1 and 3×3 . Correspondingly, the decoder blocks have the symmetric structure of the encoder blocks. It first performs a transpose convolution with 2×2 stride and 3×3 kernels, which result is concatenated with the output of the encoder at the same level to form the input of the decoder. The decoder also contain L_2 layers of *ResConv*. The output of the final decoder block is passed to a final *ConvBlock* to fit the output channel.

We use Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a $3e-4$ learning rate to optimize the analysis stage of *VoiceFixer*. We treat the first 1000 steps as the warmup phase, during which the learning rate grows linearly from 0 to $3e-4$. We decay the learning rate by 0.9 every 400 hours of training data. We perform an evaluation every 200 hours of training data. If we observe three consecutive evaluations with no improvement, we will interrupt the experiment.

For all the STFT and iSTFT, we use the hanning window with a window length of 2048 and a hop length of 441. As all the audio we use is at the 44.1 kHz sample rate, the corresponding spectrogram size in this setting will be $T \times 1025$, where T is the dimension of time frames. For mel spectrogram, the dimensions of the linear spectrogram are transformed into $T \times 128$.

B.2 DETAILS OF THE SYNTHESIS STAGE

As shown in Table 3, we use 7 kinds of STFT resolutions and 4 kinds of time resolution during the calculation of $\mathcal{L}_F$ and $\mathcal{L}_T$. So $K_F = 7$ in Equation 17 and $K_T = 4$ in Equation 18.

The mel loss $\mathcal{L}_{\text{mel}}$, spectral convergence loss $\mathcal{L}_{\text{sc}}$, STFT magnitude loss $\mathcal{L}_{\text{mag}}$, segment loss $\mathcal{L}_{\text{seg}}$, energy loss $\mathcal{L}_{\text{energy}}$, and phase loss $\mathcal{L}_{\text{phase}}$ are defined in Equation 21-26. The function $v(\cdot)$ is the

k	1	2	3	4	5	6	7
win-length	4096	2048	1024	512	256	128	64
hop-length	2048	1024	512	256	128	64	32
fft-size	8192	4096	2048	1024	512	256	128

Table 2: STFT setup for different k in $\mathcal{L}_F$.

k	1	2	3	4
frame-length	1	240	480	960
hop-length	1	120	240	480

Table 3: Windowing setup for different k in $\mathcal{L}_T$.

windowing function that divide time sample into w windows and compute mean value within each window, $v(s)_{1 \times w} = (\text{mean}(s_0), \text{mean}(s_1), \dots, \text{mean}(s_{w-1}))$. Each s_w stand for windowed s . Δ stand for first difference.

$$\mathcal{L}_{\text{mel}}(\hat{s}, s) = \left\| \hat{S}_{\text{mel}} - S_{\text{mel}} \right\|_2 \quad (21)$$

$$\mathcal{L}_{\text{sc}}(\hat{s}, s) = \frac{\left\| |\hat{S}| - |S| \right\|_F}{\left\| |\hat{S}| \right\|_F} \quad (22)$$

$$\mathcal{L}_{\text{mag}}(\hat{s}, s) = \left\| \log(|\hat{S}|) - \log(|S|) \right\|_1, \quad (23)$$

$$\mathcal{L}_{\text{seg}}(\hat{s}, s) = \|v(\hat{s}_w) - v(s_w)\|_1, \quad (24)$$

$$\mathcal{L}_{\text{energy}}(\hat{s}, s) = \|v(\hat{s}_w^2) - v(s_w^2)\|_1, \quad (25)$$

$$\mathcal{L}_{\text{phase}}(\hat{s}, s) = \|\Delta v(\hat{s}_w^2) - \Delta v(s_w^2)\|_1, \quad (26)$$

Table 9 and Table 8 show the structure of frequency and time domain discriminators. The sub-band discriminators D_{sub} and multi-resolution time discriminators $D_T^{(r)}(s)$ use the structure of *T-discriminator*, which is a stack of one dimensional convolution with grouping and large kernel size. The frequency discriminator D_F use the similar module *ResConv* similar to *ResUNet* shown in Figure 3b.

T-discriminator
Conv1d(1, 128, kernel_size=16), LeakyRelu(0.2)
Conv1d(128, 128, kernel_size=41, stride=4, padding=20, groups=8), LeakyRelu(0.2)
Conv1d(128, 128, kernel_size=41, stride=4, padding=20, groups=16), LeakyRelu(0.2)
Conv1d(128, 128, kernel_size=41, stride=4, padding=20, groups=32), LeakyRelu(0.2)
Conv1d(128, 1, kernel_size=3, stride=1, padding=1), LeakyRelu(0.2)

Figure 8: The structure of *T-discriminator*.

F-discriminator
Conv2d(1, 32, kernel_size=(3, 3))
ResConv(32, 32, stride=1, kernel_size=(3, 3))
ResConv(32, 32, stride=1, kernel_size=(3, 3))
ResConv(32, 64, stride=2, kernel_size=(3, 3))
ResConv(64, 64, stride=1, kernel_size=(3, 3))
ResConv(64, 32, stride=2, kernel_size=(3, 3))
ResConv(32, 32, stride=1, kernel_size=(3, 3))
ResConv(32, 32, stride=2, kernel_size=(3, 3))
ResConv(32, 32, stride=1, kernel_size=(3, 3))

Figure 9: The structure of *F-discriminator*.

For the training of vocoder, we setting up the λ_D to λ_{seg} value in Equation 16, Equation 17, and Equation 18 as $\lambda_D = 4.0$, $\lambda_{\text{mel}} = 50$, $\lambda_{\text{sc}} = 5.0$, $\lambda_{\text{mag}} = 5.0$, $\lambda_{\text{energy}} = 100.0$, $\lambda_{\text{phase}} = 100.0$, and $\lambda_{\text{seg}} = 200.0$

C APPENDIX C

C.1 DATASETS

Clean speech CSTR VCTK corpus (Yamagishi et al., 2019) is a multi-speaker English corpus containing 110 speakers with different accents. We split it into a training part VCTK-Train and a testing part VCTK-Test. The version of VCTK we used is 0.92. To follow the data preparation strategy of Lee & Han (2021), only the *mic1* microphone data is used for experiments, and *p280* and *p315* are omitted for the technical issues. For the remaining 108 speakers, the last 8 speakers, *p360,p361,p362,p363,p364,p374,p376,s5* are splitted as test set VCTK-Test. Within the other 100 speakers, *p232* and *p257* are omitted because they are used later in the test set DENOISE, the remaining 98 speakers are defined as VCTK-Train. Except for the training of *NuWave*, all the utterances are resampled at the 44.1 kHz sample rate. AISHELL-3 is an open-source Hi-Fi mandarin speech corpus, containing 88035 utterances with a total duration of 85 hours. HQ-TTS dataset contains 191 hours of clean speech data collected from a serial of datasets on openslr.org. In Table 4, we include the details of *HQ-TTS*, including the URL and language types of each subset.

Table 4: The components of HQ-TTS dataset.

URL	Languages	URL	Languages
http://www.openslr.org/32/	Afrikaans, Sesotho, Setswana and isiXhosa	http://www.openslr.org/70/	Nigerian English
http://www.openslr.org/37/	Bangladesh Bengali and Indian Bengali	http://www.openslr.org/71/	Chilean Spanish
http://www.openslr.org/41/	Javanese	http://www.openslr.org/72/	Colombian Spanish
http://www.openslr.org/42/	Khmer	http://www.openslr.org/73/	Peruvian Spanish
http://www.openslr.org/43/	Nepali	http://www.openslr.org/74/	Puerto Rico Spanish
http://www.openslr.org/44/	Sundanese	http://www.openslr.org/75/	Venezuelan Spanish
http://www.openslr.org/61/	Spanish	http://www.openslr.org/76/	Basque
http://www.openslr.org/63/	Malayalam	http://www.openslr.org/77/	Galician
http://www.openslr.org/64/	Marathi	http://www.openslr.org/78/	Gujarati
http://www.openslr.org/65/	Tamil	http://www.openslr.org/79/	Kannada
http://www.openslr.org/66/	Telugu	http://www.openslr.org/80/	Gujarati
http://www.openslr.org/69/	Catalan		

Noise data One of the noise dataset we use come from VCTK-Demand (VD) (Valentini-Botinhao et al., 2017), a widely used corpus for speech denoising and noise-robust TTS training. This dataset contains a training part VD-Train and a testing part VD-Test, in which both contain two noisy set VD-Train-Noisy, VD-Test-Noisy and two clean speech set VD-Train-Clean, VD-Test-Clean. To obtain the noise data from this dataset, we minus each noisy data from VD-Train-Noisy with its corresponding clean part in VD-Train-Clean to get the final training noise dataset VD-Noise. The noise data are all resampled to 44.1 kHz. Another noise dataset we adopt is the TUT urban acoustic scenes 2018 dataset (Mesaros et al., 2018), which is originally used for the acoustic scene classification task of DCASE 2018 Challenge. The dataset contains 89 hours of high-quality recording from 10 acoustic scenes such as airport and shopping mall. The total amount of audio is divided into development DCASE-Dev and evaluation DCASE-Eval parts. Both of them contain audio from all cities and all acoustic scenes.

Room impulse response We randomly simulated a collection of Room Impulse Response filters to simulate the 44.1 kHz speech room reverberation using an open-source tool ². The meters of height, width, and length of the room are sampled randomly in a uniform distribution $\mathcal{U}(1, 12)$. The placement of the microphone is then randomly selected within the room space. For the placement of the sound source, we first determined the distance to the microphone, which is randomly sampled in a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, $\mu = 2, \sigma = 4$. If the sampled value is negative or greater than five meters, we will sample the distance again until it meets the requirement. After sampling the distance between the microphone and sound source, the placement of the sound source is randomly selected within the sphere centered at the microphone. The RT60 value we choose come from the uniform distribution $\mathcal{U}(0.05, 1.0)$. For the pickup pattern of the microphone, we randomly choose from omnidirectional and cardioid types. Finally, we simulated 43239 filters, in which we randomly split out 5000 filters as the test set RIR-Test and named other 38239 filters as RIR-Train.

²https://github.com/sunits/rir_simulator_python

C.2 TRAINING DATA

We describe the simulation of training data in Algorithm 1. $\mathbb{S} = \{s^{(0)}, s^{(1)}, \dots, s^{(i)}\}$, $\mathbb{N} = \{n^{(0)}, n^{(1)}, \dots, n^{(i)}\}$, and $\mathbb{R} = \{r^{(0)}, r^{(1)}, \dots, r^{(i)}\}$ are the speech dataset, noise dataset, and RIR dataset. We use several helper function to describe this algorithm. `randomFilterType( $\cdot$ )` is a function that randomly select a type of filter within butterworth, chebyshev, bessel, and elliptic. `Resample( $x, o_1, u$ )` is a resampling function that resample the one dimensional signal x from a original samplerate o_1 to the target u samplerate. `buildFilter( $t, c, o_2$ )` is a filter design function that return a type t filter with cutoff frequency c and order o_2 . `max( $\cdot$ )`, `min( $\cdot$ )`, and `abs( $\cdot$ )` is the element wise maximum, minimum, and absolute value function. `mean( $\cdot$ )` calculate the mean value of the input.

We first select a speech utterance s , a segment of noise n and a RIR filter r randomly from the dataset. Then with p_1 probability, we add the reverberation effect using r . And with p_2 probability, we add clipping effect with a clipping ratio η , which is sampled in a uniform distrubution $\mathcal{U}(\eta_{low}, \eta_{high})$. To produce low-resolution effect, after determining the filter type t , we randomly sample the cutoff frequency c and order o from the uniform distribution $\mathcal{U}(C_{low}, C_{high})$ and $\mathcal{U}(O_{low}, O_{high})$. Then we perform convolution between x and the type t order o lowpass filter with cutoff frequency c . Finally the filtered data will be resampled twice, one is resample to $c * 2$ samplerate and another is resample back to 44.1 kHz. We also perform the same lowpass filtering to the noise signal randomly. This operation is necessary because, if not, the model will overfit the pattern that the bandwidth of noise signal is always different from speech. In this case, the model will fail to remove noise when the bandwidth of noise and speech are similar. For the simulation of noisy environment, we randomly add the noise n into the speech signal x using a random SNR $s \sim \mathcal{U}(S_{low}, S_{high})$. To fit the model with all energy levels, we randomly conduct a $q \sim \mathcal{U}(Q_{low}, Q_{high})$ scaling to the input and target data pair.

In our work, we choose the following parameters to perform this algorithm, $p_1 = 0.25$, $p_2 = 0.25$, $p_3 = 0.5$, $\eta_{low} = 0.06$, $\eta_{high} = 0.9$, $C_{low} = 750$, $C_{high} = 22050$, $O_{low} = 2$, $O_{high} = 10$, $S_{low} = -5$, $S_{high} = 40$, $Q_{low} = 0.3$, $Q_{high} = 1.0$.

Algorithm 1: Add random distortions to high-quality speech signal s

In: $s \leftarrow \mathbb{S}$; $n \leftarrow \mathbb{N}$; $r \leftarrow \mathbb{R}$

Out: The high-quality speech s and its randomly distorted version x

```

 $x = s$ ;
with  $p_1$  probability:
     $x = x * r$ ; /* Convolute with RIR filter */
with  $p_2$  probability:
     $\theta \sim \mathcal{U}(\Theta_{low}, \Theta_{high})$ ; /* Choose clipping ratio */
     $x = \max(\min(x, \theta), -\theta)$ ; /* Hard clipping */
with  $p_3$  probability:
     $t = \text{randomFilterType}()$ ;
     $c \sim \mathcal{U}(C_{low}, C_{high})$ ;
     $o \sim \mathcal{U}(O_{low}, O_{high})$ ; /* Random cutoff and order */
     $x = x * \text{buildFilter}(t, c, o)$ ; /* Low pass filtering */
     $x = \text{Resample}(\text{Resample}(x, 44100, c * 2), c * 2, 44100)$ ; /* Resample */
with  $p_4$  probability:
     $n = n * \text{buildFilter}(t, c, o)$ ; /* Low pass filtering on noise */
     $n = \text{Resample}(\text{Resample}(n, c * 2), 44100)$ ; /* Resample */
with  $p_5$  probability:
     $s \sim \mathcal{U}(S_{low}, S_{high})$ ;
     $q \sim \mathcal{U}(Q_{low}, Q_{high})$ ; /* Random SNR and scale */
     $n = \frac{n}{\text{mean}(\text{abs}(n)) / \text{mean}(\text{abs}(x))}$ ; /* Normalize the energy of noise */
     $x = (x + \frac{n}{10^{s/20}})$ ; /* Add noise */
 $s = qs$ ; /* Scaling */
 $x = qx$ ; /* Scaling */

```

C.3 TESTING DATA

Testing data is crucial for the evaluation for each kind of distortion. The testing data we use either come from open-sourced test set or simulated by ourselves.

Super-resolution The simulation of the SR test set follows the work of (Kuleshov et al., 2017; Wang & Wang, 2021). The low-resolution and target data pairs are obtained by transforming 44.1 kHz sample rate utterances in target speech data VCTK-Test to a lower sample rate u . To achieve that, we first convolve the speech data with an order 8 Chebyshev type I lowpass filter with the $\frac{u}{2}$ cutoff frequency. Then we subsample the signal to u sample rate using polyphase filtering. In this work, to test the performance on different sampling rate settings, u are set at 2 kHz, 4 kHz, 8 kHz, 16 kHz, and 24 kHz. We denote the corresponding five testing set as VCTK-4k, VCTK-4k, VCTK-8k, VCTK-16k, and VCTK-24k, respectively.

Denoising For the denoising task, we adopt the open-sourced testing set DENOISE described in Appendix C.1. This test set contains 824 utterances from a female speaker and a male speaker. The type of noise data comprises a domestic noise, an office noise, noise in the transport scene, and two street noises. The test set is simulated at four SNR levels, which are 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. The original data is sampled at 48 kHz. We downsample it to 44.1 kHz to fit our experiments.

Dereverberation The test set for dereverberation, DEREV, is simulated using VCTK-Test and RIR-Test. For each utterance in VCTK-Test, we first randomly select an RIR from RIR-Test, then we calculate the convolution between the RIR and utterance to build the reverberant speech. Finally, we build 2937 reverberant and target data pairs.

Declicking DECLI, the evaluation set for declicking, is also constructed based on VCTK-Test. We perform clipping on VCTK-Test following the equation in Section 2 and choose 0.25, 0.1 as the two setups for the clipping ratio. This result in two declicking test sets with different levels, each containing 2937 clipped speech and target audios.

General speech restoration To evaluate the performance on GSR, we simulate a test set ALL-GSR comprising of speech with all kinds of distortion. The clean speeches and noise data used to build ALL-GSR is VCTK-Test and DCASE-Eval. The simulation procedure of ALL-GSR is almost the same to the training data simulation described in Section 4.2. In total, 501 three seconds long utterances are produced in this test set.

MOS Evaluation We select a small portion from the test sets to carry out MOS evaluation for each one. In SR, DECLI, and DEREV, we select 38 utterances out for human ratings. In DENOISE and ALL-GSR, we randomly choose 42 and 51 utterances.

C.4 EVALUATION METRICS

Log-spectral distance LSD is a commonly used metrics on the evaluation of super-resolution performance (Kumar et al., 2020; Lee & Han, 2021; Wang & Wang, 2021). For target signal s and output estimate $\hat{s}$, LSD can be computed as Equation 27, where S and $\hat{S}$ stand for the magnitude spectrogram of s and $\hat{s}$.

$$\text{LSD}(S, \hat{S}) = \frac{1}{T} \sum_{t=1}^T \sqrt{\frac{1}{F} \sum_{f=1}^F \log_{10} \left(\frac{S(f, t)^2}{\hat{S}(f, t)^2} \right)^2} \quad (27)$$

Perceptual evaluation of speech quality PESQ is widely used in speech restoration literature as their evaluation metrics (Pascual et al., 2017; Hu et al., 2020). It was originally developed to model the subjective test commonly used in telecommunication. PESQ provides a score ranging from -0.5 to 4.5 and the higher the score, the better quality a speech has. In our work, we used an open-sourced implementation of PESQ to compute these metrics. Since PESQ only works on a 16 kHz sampling rate, we performed a 16 kHz downsampling to the output 44.1k audio before evaluation.

Structural similarity SSIM (Wang et al., 2004) is a metrics in image super-resolution. It addresses the shortcoming of pixel-level metrics by taking the image texture into account. We match the implementation of SSIM in (Wang et al., 2004) with ours and compute SSIM as Equation 28, where μ_S and σ_S is the mean and standard deviation of S . $\text{Cov}_{S\hat{S}}$ is the Covariance of S and $\hat{S}$. $\epsilon_1 = 0.01$

and $\epsilon_2 = 0.02$ are two constant used to avoid zero division. Similarity is measured within the K $7*7$ blocks divided from $\mathbf{S}$ and $\hat{\mathbf{S}}$.

$$\text{SSIM}(\mathbf{S}, \hat{\mathbf{S}}) = \Sigma_{k=1}^K \left(\frac{(2\mu_{\mathbf{S}_k} \mu_{\hat{\mathbf{S}}_k} + \epsilon_1)(2\text{Cov}_{\mathbf{S}_k \hat{\mathbf{S}}_k} + \epsilon_2)}{(\mu_{\mathbf{S}_k}^2 + \mu_{\hat{\mathbf{S}}_k}^2 + \epsilon_1)(\sigma_{\mathbf{S}_k}^2 + \sigma_{\hat{\mathbf{S}}_k}^2 + \epsilon_2)} \right) \quad (28)$$

Scale-invariant signal to noise ratio SiSNR (Le Roux et al., 2019) is widely used in speech restoration literatures to compare the energy of a signal to its background noise. SiSNR is calculated between target waveform $\mathbf{s}$ and waveform estimation $\hat{\mathbf{s}}$:

$$\text{SiSNR}(\mathbf{s}, \hat{\mathbf{s}}) = 10 * \log_{10} \frac{\|\hat{\mathbf{s}}_{target}\|^2}{\|e_{noise}\|^2}, \quad (29)$$

where $\hat{\mathbf{s}}_{target} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\hat{\mathbf{s}}\|^2}$. A higher SiSNR indicates less discrepancy between the estimation and target.

Scale-invariant spectrogram to noise ratio SiSPNR is a spectral metric similar to SiSNR. They have the similar idea except SiSPNR is computed on the magnitude spectrogram. Given the target spectrogram $\mathbf{S}$ and estimation $\hat{\mathbf{S}}$, the computation of SiSPNR can be formulated as

$$\text{SiSPNR}(\mathbf{S}, \hat{\mathbf{S}}) = 10 * \log_{10} \frac{\|\hat{\mathbf{S}}_{target}\|^2}{\|e_{noise}\|^2} \quad (30)$$

where $\hat{\mathbf{S}}_{target} = \frac{\langle \hat{\mathbf{S}}, \mathbf{S} \rangle \mathbf{S}}{\|\hat{\mathbf{S}}\|^2}$. The scale invariant is guranteed by mean normalization of estimated and target spectrogram.

D APPENDIX D

Table 5: Experiments setup. We list the training and testing sets used for each model’s training and evaluation. Check mark is used to denote whether a model adopts the framework of *VoiceFixer* and whether it is trained for SSR or GSR task.

Name	VoiceFixer	SSR	GSR	Training sets	Testing sets
Unprocessed	✓	✓	✓	/	DENOISE; DEREV; SR; DECLI; ALL-GSR;
Oracle-Mel	✓	✓	✓	/	DENOISE; DEREV; SR; DECLI; ALL-GSR;
Vocoder-TFGAN	✓	✓	✓	VCTK-Train; HQ-TTS; AISHELL-3	DENOISE; DEREV; SR; DECLI; ALL-GSR;
Denoise-UNet	✓	✓	✓	VCTK-Train; VD-Noise;	DENOISE; ALL-GSR;
Dereverb-UNet	✓	✓	✓	VCTK-Train; RIR-Train;	DEREV
SR-UNet	✓	✓	✓	VCTK-Train;	SR
Declip-UNet	✓	✓	✓	VCTK-Train;	DECLI
NuWave	✓	✓	✓	VCTK-Train;	SR
SEANet	✓	✓	✓	VCTK-Train;	SR
SSPADE	✓	✓	✓	/	DECLI
GSR-UNet	✓	✓	✓	VCTK-Train; VD-Noise; RIR-Train;	DENOISE; DEREV; SR; DECLI; ALL-GSR;
VF-DNN	✓	✓	✓	VCTK-Train; VD-Noise; RIR-Train;	DENOISE; DEREV; SR; DECLI; ALL-GSR;
VF-BiGRU	✓	✓	✓	VCTK-Train; VD-Noise; RIR-Train;	DENOISE; DEREV; SR; DECLI; ALL-GSR;
VF-UNet-S	✓	✓	✓	VCTK-Train; VD-Noise; RIR-Train;	DENOISE; DEREV; SR; DECLI; ALL-GSR;
VF-UNet	✓	✓	✓	VCTK-Train; VD-Noise; RIR-Train;	DENOISE; DEREV; SR; DECLI; ALL-GSR;

D.1 EVALUATION RESULTS

Table 6: Evaluation results on speech super-resolution test set SR, which includes five sampling rate setup. The metrics is calculated at a target sampling rate of 44.1 kHz

TRAININGSCHEME		ONE-STAGE MODELS				VOICEFIXER MODELS				OTHERS		
SampleRate Up Ratio	Metrics	GSR- UNet	SR-UNet	NuWave	SEANet	VF- DNN	VF- BiGRU	VF- UNet-S	VF- UNet	Unprocessed	Oracle -Mel	Target
2kHz 22.1	LSD	1.34	1.19	1.41	1.33	1.18	1.08	1.08	1.05	3.13	0.89	/
	SiSPNR	11.03	10.89	9.19	9.78	10.67	11.84	11.65	12.10	9.18	13.65	/
	SSIM	0.75	0.77	0.73	0.72	0.75	0.77	0.78	0.78	0.68	0.85	/
4kHz 11.0	LSD	1.27	1.18	1.35	1.24	1.15	1.03	1.04	1.02	2.97	0.89	/
	SiSPNR	11.48	11.10	9.65	10.58	11.07	12.27	11.98	12.41	9.52	13.65	/
	SSIM	0.77	0.78	0.76	0.72	0.75	0.79	0.79	0.79	0.71	0.85	/
8kHz 5.5	LSD	1.21	1.11	1.24	1.20	1.06	0.99	1.01	0.99	2.70	0.89	/
	SiSPNR	12.07	11.82	10.73	11.11	11.94	12.68	12.34	12.74	9.93	13.65	/
	SSIM	0.81	0.82	0.80	0.74	0.78	0.81	0.81	0.81	0.76	0.85	/
16kHz 2.8	MOS	3.37	3.34	3.09	3.37	/	/	/	3.40	3.05	3.53	3.63
	LSD	1.10	0.99	1.18	1.16	1.01	0.94	0.96	0.94	2.32	0.89	/
	SiSPNR	13.02	13.01	11.54	11.90	12.37	13.14	12.70	13.14	10.08	13.65	/
24kHz 1.8	SSIM	0.85	0.88	0.81	0.75	0.82	0.82	0.82	0.82	0.83	0.85	/
	LSD	0.97	0.91	1.12	1.15	0.93	0.91	0.94	0.92	1.91	0.89	/
	SiSPNR	13.96	13.81	11.63	12.58	13.21	13.38	12.86	13.38	10.40	13.65	/
Average Score	SSIM	0.87	0.91	0.81	0.75	0.84	0.83	0.83	0.84	0.89	0.85	/
	MOS	3.56	3.52	3.54	3.65	/	/	/	3.41	3.47	3.44	3.45
	LSD	1.18	1.07	1.26	1.21	1.07	0.99	1.01	0.98	2.61	0.89	/
	SiSPNR	12.31	12.13	10.55	11.19	11.85	12.66	12.31	12.75	9.82	13.65	/
	SSIM	0.81	0.83	0.79	0.74	0.79	0.80	0.81	0.81	0.77	0.85	/

Table 7: Evaluation result on the speech denoising test set DENOISE

Models	SiSNR	PESQ	SiSPNR	MOS
Unprocessed	8.40	1.97	9.78	3.20
Oracle-Mel	-17.52	2.85	12.84	3.64
Target	/	/	/	3.69
SEGAN (Pascual et al., 2017)	/	2.16	/	/
Wave-U-Net (Macartney & Weyde, 2018)	/	2.40	/	/
Weakly Labelled (Kong et al., 2021)	/	2.28	/	/
GSR-UNet	16.42	2.82	12.25	3.64
Denoise-UNet	17.58	2.71	11.82	3.63
VF-DNN	/	1.71	10.93	/
VF-BiGRU	/	2.29	11.72	/
VF-UNet-S	/	2.33	11.19	/
VF-UNet	/	2.43	11.71	3.69

Table 8: Evaluation results on the speech dereverberation test set DEREV

Models	PESQ	SiSPNR	MOS
Unprocessed	1.99	14.58	2.70
Oracle-Mel	2.36	13.65	3.46
Target	/	/	3.51
GSR-UNet	2.35	14.10	3.32
Dereverb-UNet	2.49	14.99	3.25
VF-DNN	1.41	11.70	/
VF-BiGRU	1.69	13.00	/
VF-UNet-S	1.78	12.80	/
VF-UNet	1.86	13.21	3.52

D.2 ANALYSIS STAGE PERFORMANCE

In this section, we report the mel spectrogram restoration score on different test sets. They are used to evaluate the performance of the analysis stage. We calculate the LSD, SiSPNR, and SSIM values.

Table 9: Evaluation results on the speech declipping test set DECLI

Clipping Level Models	0.25				0.10				Average			
	SiSNR	STOI	PESQ	MOS	SiSNR	STOI	PESQ	MOS	SiSNR	STOI	PESQ	MOS
Unprocessed	9.60	0.95	2.38	2.56	4.00	0.89	1.51	2.72	6.80	0.92	1.95	2.64
Oracle-Mel	-19.94	0.81	2.36	3.44	-19.94	0.81	2.36	3.42	-19.94	0.81	2.36	3.43
Target	/	/	/	3.42	/	/	/	3.49	/	/	/	3.46
GSR-UNet	11.01	0.97	3.54	3.38	7.47	0.94	2.89	3.23	9.24	0.95	3.21	3.31
Declip-UNet	12.45	0.99	3.98	3.38	8.43	0.96	3.40	3.38	10.44	0.98	3.69	3.38
SSPADE	17.43	0.98	3.55	3.34	10.31	0.92	2.12	2.63	13.87	0.95	2.84	2.98
VF-DNN	/	0.76	1.72	/	/	0.72	1.48	/	/	0.74	1.60	/
VF-BiGRU	/	0.81	2.09	/	/	0.79	1.82	/	/	0.80	1.95	/
VF-UNet-S	/	0.82	2.13	/	/	0.80	1.85	/	/	0.81	1.99	/
VF-UNet	/	0.82	2.21	3.38	/	0.80	1.93	3.38	/	0.81	2.07	3.38

The *Unprocessed* column is calculated using the target and unprocessed mel spectrogram. And the *Oracle-Mel* column is calculated using the target spectrogram and itself.

Table 10: The Performance of Mel Spectrogram Restroation on DENOISE, DEREV, and ALL-GSR test sets

Models	DENOISE			DEREV			ALL-GSR		
	LSD	SiSPNR	SSIM	LSD	SiSPNR	SSIM	LSD	SiSPNR	SSIM
Unprocessed	1.31	-1.41	0.57	0.84	10.02	0.63	1.65	-3.90	0.47
VF-DNN	0.76	7.61	0.69	0.93	8.86	0.59	0.87	6.26	0.58
VF-BiGRU	0.55	10.98	0.79	0.56	12.91	0.75	0.59	10.49	0.70
VF-UNet-S	0.52	10.29	0.82	0.47	13.61	0.82	0.55	11.08	0.75
VF-UNet	0.46	12.27	0.84	0.46	14.89	0.82	0.53	11.36	0.76

Table 10 shows that on DENOISE, DEREV, and ALL-GSR, all four *VoiceFixer* based models are effective on the restoration of mel spectrogram. Among the four analysis stage models, *UNet* is consistently better than the other three models.

Table 11: The performance of mel spectrogram restroation on the SR test set

SampleRate Upsampling Ratio	Metrics	MODELS					
		VF-DNN	VF-BiGRU	VF-UNet-S	VF-UNet	Unprocessed	Oracle-Mel
2kHz 22.1	LSD	0.80	0.68	0.65	0.60	2.99	0.00
	SiSPNR	8.02	9.62	9.82	11.32	2.54	127.43
	SSIM	0.56	0.63	0.66	0.68	0.40	1.00
4kHz 11.0	LSD	0.68	0.54	0.55	0.50	2.54	0.00
	SiSPNR	9.66	12.23	11.22	12.83	3.16	127.43
	SSIM	0.65	0.72	0.74	0.76	0.51	1.00
8kHz 5.5	LSD	0.51	0.40	0.46	0.42	2.02	0.00
	SiSPNR	12.53	14.85	12.67	14.20	4.26	127.43
	SSIM	0.77	0.82	0.83	0.84	0.64	1.00
16kHz 2.8	LSD	0.43	0.26	0.37	0.33	1.53	0.00
	SiSPNR	13.62	19.00	14.07	16.13	5.64	127.43
	SSIM	0.83	0.91	0.90	0.91	0.77	1.00
24kHz 1.8	LSD	0.29	0.18	0.31	0.27	1.16	0.00
	SiSPNR	17.94	22.16	15.53	18.59	7.40	127.43
	SSIM	0.92	0.95	0.94	0.95	0.86	1.00
Average	LSD	0.54	0.41	0.47	0.43	2.05	0.00
	SiSPNR	12.35	15.57	12.66	14.61	4.60	127.43
	SSIM	0.75	0.80	0.81	0.83	0.64	1.00

Table 11 lists the mel restoration performance on different sampling rates. Although *VF-BiGRU* has fewer parameters than *VF-UNet*, it still achieved the highest score on LSD and SiSPNR averagely. This result indicates that the recurrent structure might be more suitable for the mel spectrogram super-resolution task when the initial sampling rate is high.

D.3 DEMOS

In this section, we provide some restoration demos using our proposed *VoiceFixer*. In Figure 12, we provide eight restoration demos using our *VF-UNet* model. All the audios used in our demo are either collected from the internet or recorded by ourselves. In each example, the left is the unprocessed spectrogram, and the right is the restored one. After restoration, these seriously distorted speech signals can be revert to relatively high quality.

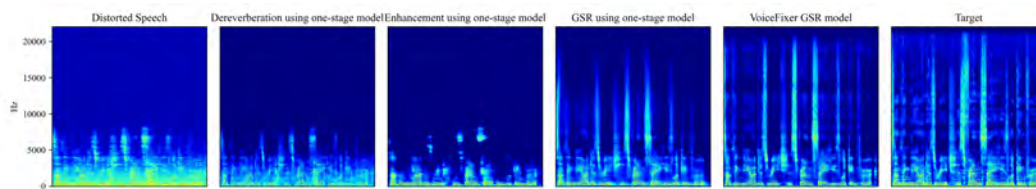


Figure 10: Comparison between different restoration methods. The unprocessed speech is noisy, reverberant, and in low-resolution. The leftmost spectrogram is the unprocessed low-quality speech and the rightmost is the target high-quality spectrogram. In the middle, from left to right, the figures show results processed by one-stage SSR dereverberation model, SSR denoising model, GSR model and *VoiceFixer* based GSR model.

Figure 12b is the speech we recorded using Adobe Audition. We set the sampling rate to 8 kHz and manually add the clipping effect. It also contains some low-frequency noise and reverberation introduced by the recording device and environment. Figure 12a is a speech delivered by *Amelia Earhart*³, 1897-1937, appeared in the Library of Congress, United States. The original version sounds like a mumble. Figure 12f comes from an interview in a TV news program, which includes multiple distortions. Figure 12e is collected from the audio uploaded by a Youtuber⁴. Probably due to the recording device, her speech is deteriorated seriously by noise, and the energy of speech in the low-frequency part is also relatively low. Figure 12c is the restoration of a Chinese famous old movie *railroad guerrilla*⁵, which speech has limited bandwidth. The audio in Figure 12d is selected from a well-known TV series in China, *romance of the three kingdoms*⁶, in which some parts of the spectrogram are masked off due to the previous audio compression. Figure 12g is a recording⁷ selected from a speech delivered by *Sun Yat-sen*, 1866-1925, which is in extremely low-resolution and includes multiple unknown distortions. Figure 12h shows the result of a subway broadcasting we recorded in Shanghai. The low-frequency part of speech almost lost completely, and the reverberation is serious.

To sum up, all these examples prove the effectiveness of the *VoiceFixer* model on GSR. And to our surprise, it can generate well on unseen distortions such as the spectrogram lost in Figure 12c, Figure 12f, and Figure 12d. In addition, Figure 12e shows that *VoiceFixer* is effective for the compensation of low-frequency energy, making speech sound less machinery and distant. Last but not least, despite the abnormal harmonic structure in the low-frequency part in Figure 12g, our proposed model can still repair it into a normal distribution, which proves the advantages of utilizing the prior knowledge of vocoder.

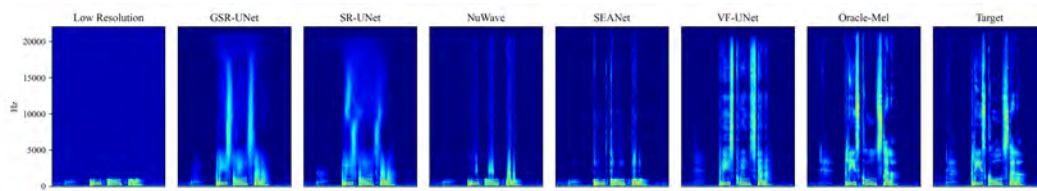
³<https://www.loc.gov/item/afccal000004>

⁴<https://v.ixigua.com/egVW74E/>

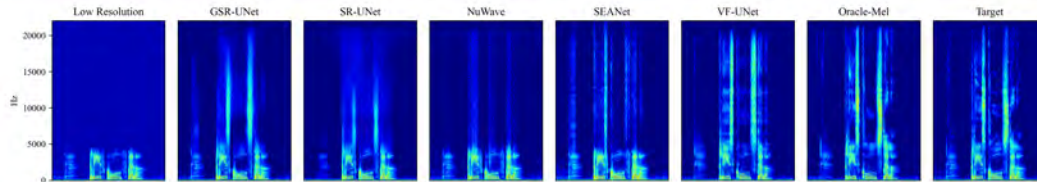
⁵<https://www.youtube.com/watch?v=R8lY1qn2CHA>

⁶<https://www.youtube.com/watch?v=6h7N4Cl1lTw>

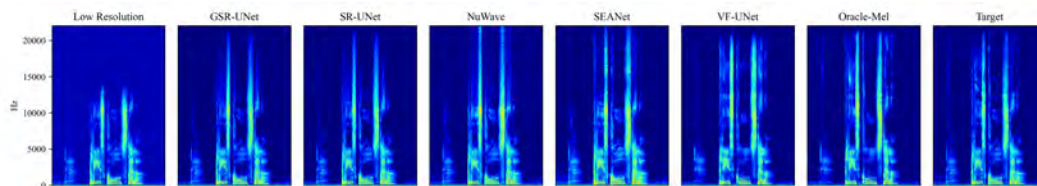
⁷<https://www.bilibili.com/video/BV1WW411V7fR>



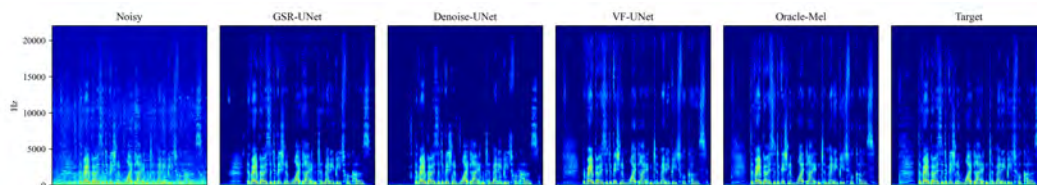
(a) Speech super-resolution results on 2 kHz source samplerate test data.



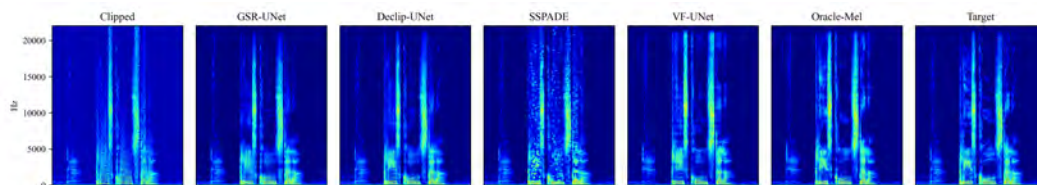
(b) Speech super-resolution results on 8 kHz source samplerate test data.



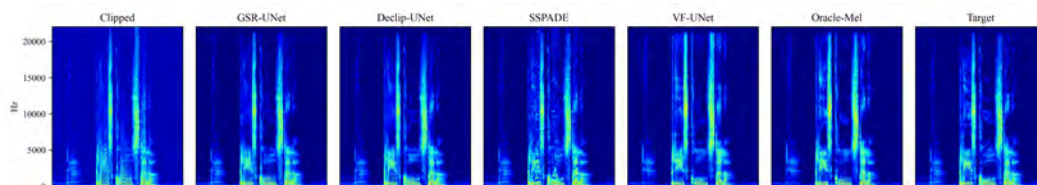
(c) Speech super-resolution results on 24 kHz source samplerate test data.



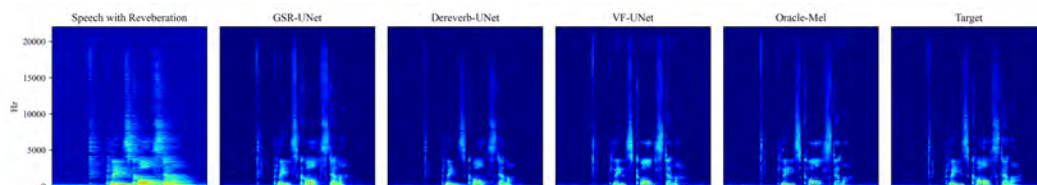
(d) Speech denoising results.



(e) Speech declipping results on speech with 0.1 clipping threshold.

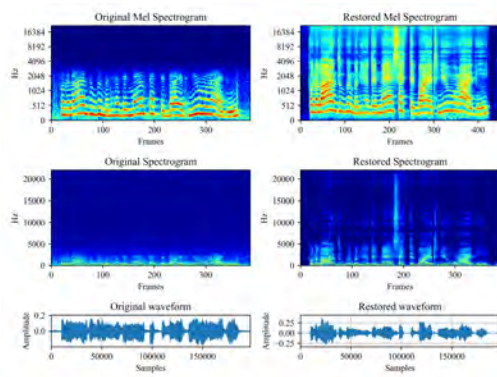


(f) Speech declipping results on speech with 0.25 clipping threshold.

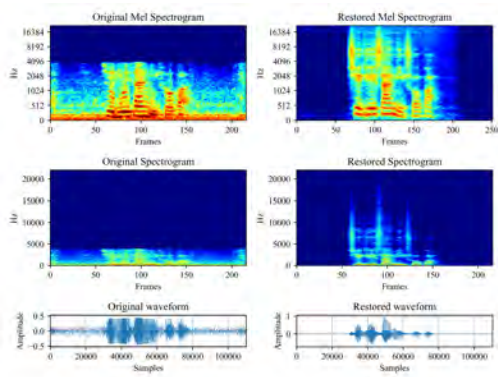


(g) Speech dereverberation results.

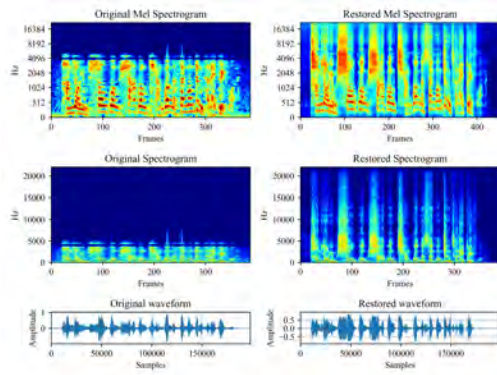
Figure 11: Comparison between different model on four different tasks using simulated data.



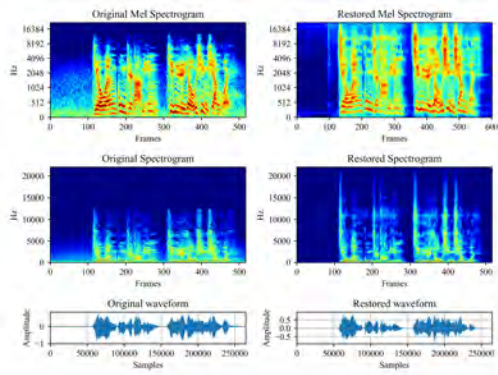
(a) Historical speech



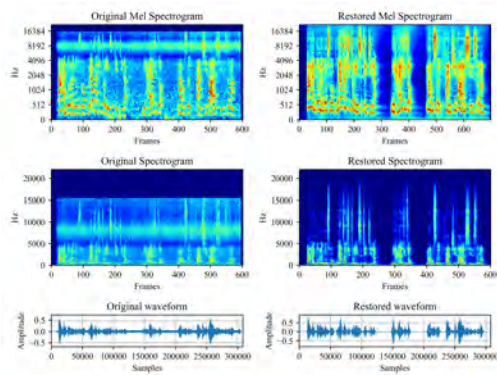
(b) A recording of a person



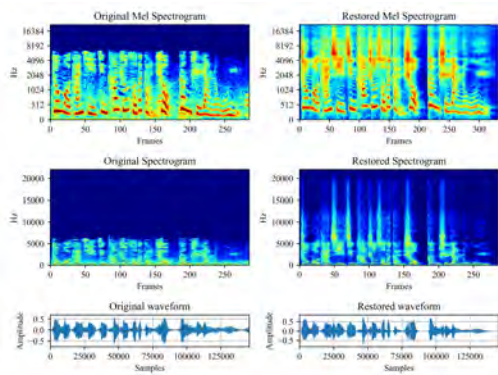
(c) Old movie



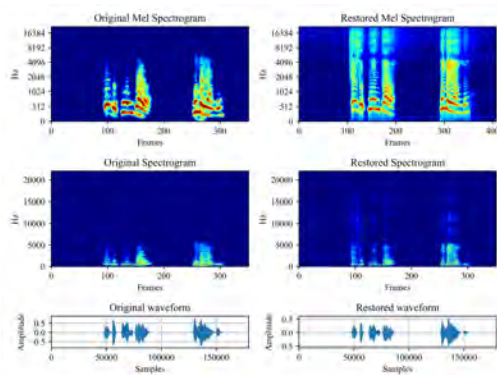
(d) Old TV series



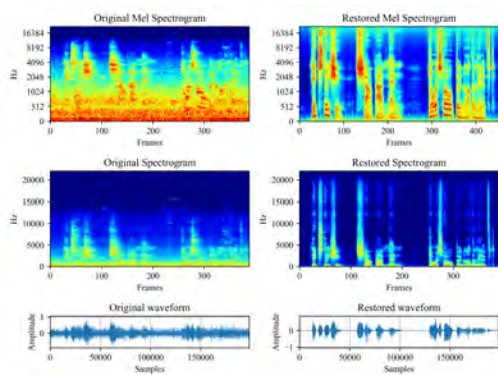
(e) A recording of a Chinese YouTuber



(f) Interview in a TV news program



(g) Historical speech



(h) Subway broadcasting

Figure 12: Restoration on the audios either collected from the internet or recorded by ourselves.

Appendix LIU03



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN UPLOAD

Search

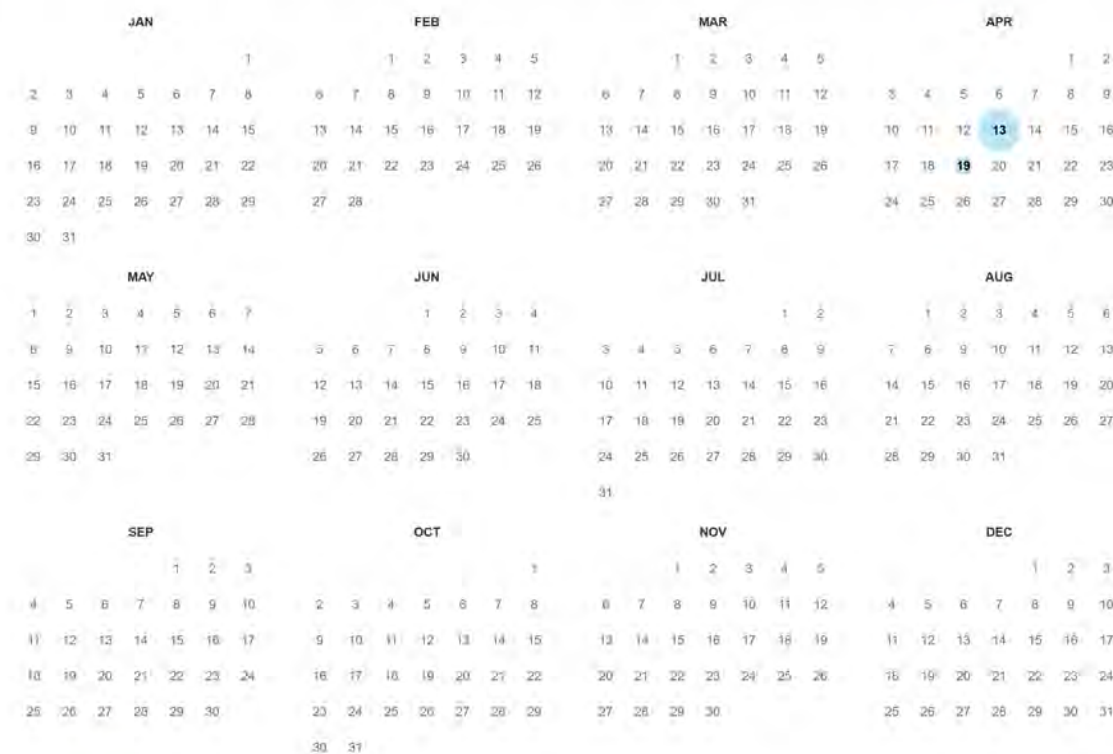
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER



Explore more than 1 trillion web pages saved over time

[Calendar](#) · [Collections](#) · [Changes](#) · [Summary](#) · [Site Map](#) · [URLs](#)

Saved 11 times between April 13, 2022 and December 10, 2025.



Note

This calendar view maps the number of times <https://arxiv.org/abs/2204.05841> 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](#).

INTERNET ARCHIVE

Wayback Machine

https://arxiv.org/abs/2204.05841

Go

MAR 13 2022

APR 13 2022

APR 13 2022

11 captures

13 Apr 2022 - 10 Dec 2025

About this capture

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv > eess > arXiv:2204.05841

Search... All fields Search

Help | Advanced Search

Electrical Engineering and Systems Science > Audio and Speech Processing

[Submitted on 12 Apr 2022]

VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration

Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on a single type of distortion, such as speech denoising or dereverberation. However, speech signals can be degraded by several different distortions simultaneously in the real world. It is thus important to extend speech restoration models to deal with multiple distortions. In this paper, we introduce VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and can expand degraded speech (e.g., noisy speech) with a low bandwidth to 44.1 kHz full-bandwidth high-fidelity speech. We design VoiceFixer based on (1) an analysis stage that predicts intermediate-level features from the degraded speech, and (2) a synthesis stage that generates waveform using a neural vocoder. Both objective and subjective evaluations show that VoiceFixer is effective on severely degraded speech, such as real-world historical speech recordings. Samples of VoiceFixer are available at [this https URL](#).

Comments: Submitted to INTERSPEECH 2022

Subjects: **Audio and Speech Processing (eess.AS)**; Sound (cs.SD); Signal Processing (eess.SP)

Cite as: [arXiv:2204.05841 \[eess.AS\]](#)
(or [arXiv:2204.05841v1 \[eess.AS\]](#) for this version)

Submission history

From: Haohe Liu [[view email](#)]

[v1] Tue, 12 Apr 2022 14:37:39 UTC (5,125 KB)

Download:

- PDF
- Other formats (license)

Current browse context: **eess.AS**

< prev | next >

new | recent | 2204

Change to browse by:

cs

cs.SD

eess

eess.SP

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

Bookmark

Bibliographic Tools

Code & Data

Demos

Related Papers

About arXiv Labs

Bibliographic and Citation Tools

☐ Bibliographic Explorer ([What is the Explorer?](#))

☐ Litmaps ([What is Litmaps?](#))

☐ scite Smart Citations ([What are Smart Citations?](#))

[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))

Appendix LIU04



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN | UPLOAD

Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

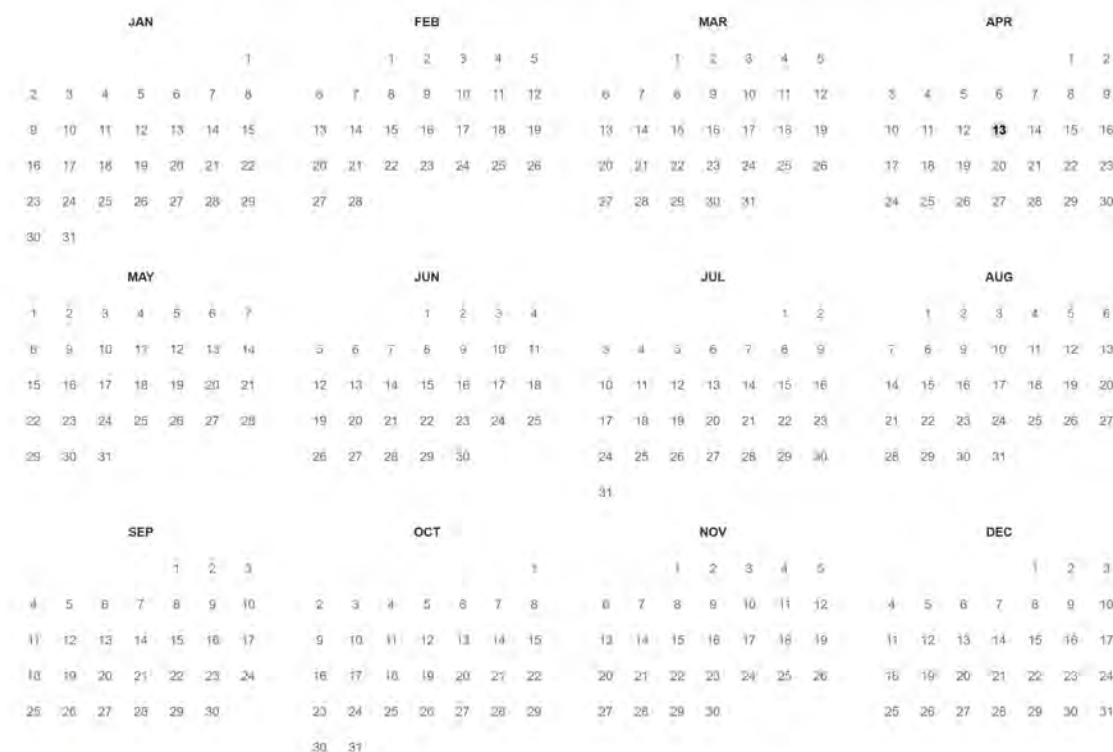


Explore more than 1 trillion web pages saved over time

https://arxiv.org/pdf/2204.05841.pdf

[Calendar](#) · [Collections](#) · [Changes](#) · [Summary](#) · [Site Map](#) · [URLs](#)

Saved 3 times between April 13, 2022 and April 19, 2024.



Note

This calendar view maps the number of times <https://arxiv.org/pdf/2204.05841.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](#).

VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration

Haohe Liu^{1,*}, Xubo Liu^{1,*}, Qiuqiang Kong², Qiao Tian², Yan Zhao²,
DeLiang Wang³, Chuanzeng Huang², Yuxuan Wang²

¹Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK

²Speech, Audio, and Music Intelligence (SAMI) Group, ByteDance, China

³Department of Computer Science and Engineering, The Ohio State University, USA

h101486@surrey.ac.uk, kongqiuqiang@bytedance.com, dwang@cse.ohio-state.edu

Abstract

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on a single type of distortion, such as speech denoising or dereverberation. However, speech signals can be degraded by several different distortions simultaneously in the real world. It is thus important to extend speech restoration models to deal with multiple distortions. In this paper, we introduce VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and can expand degraded speech (e.g., noisy speech) with a low bandwidth to 44.1 kHz full-bandwidth high-fidelity speech. We

However, this approach has limitations such as increasing computational cost and accumulating the artifacts introduced by each speech restoration model. To our knowledge, restoring low-bandwidth speech with multiple distortions has not been studied in the literature.

This paper introduces VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and could expand distorted speech with a low bandwidth between 1 kHz and 22.05 kHz to a full-bandwidth high-fidelity speech signal. We design VoiceFixer based on a two-stage strategy: (1) an analysis stage that performs mel spectrogram estimation; (2)

2 Apr 2022

VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration

Haohe Liu^{1,*}, Xubo Liu^{1,*}, Qiuqiang Kong², Qiao Tian², Yan Zhao²,
DeLiang Wang³, Chuanzeng Huang², Yuxuan Wang²

¹Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK

²Speech, Audio, and Music Intelligence (SAMI) Group, ByteDance, China

³Department of Computer Science and Engineering, The Ohio State University, USA

hl01486@surrey.ac.uk, kongqiuqiang@bytedance.com, dwang@cse.ohio-state.edu

Abstract

Speech restoration aims to remove distortions in speech signals. Prior methods mainly focus on a single type of distortion, such as speech denoising or dereverberation. However, speech signals can be degraded by several different distortions simultaneously in the real world. It is thus important to extend speech restoration models to deal with multiple distortions. In this paper, we introduce VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and can expand degraded speech (e.g., noisy speech) with a low bandwidth to 44.1 kHz full-bandwidth high-fidelity speech. We design VoiceFixer based on (1) an analysis stage that predicts intermediate-level features from the degraded speech, and (2) a synthesis stage that generates waveform using a neural vocoder. Both objective and subjective evaluations show that VoiceFixer is effective on severely degraded speech, such as real-world historical speech recordings. Samples of VoiceFixer are available at <https://haoheliu.github.io/voicefixer>.

Index Terms: speech restoration, speech super-resolution, neural vocoder, speech synthesis, deep learning

1. Introduction

Human speech often suffers from distortions such as background noise, room reverberations, or clipping from low-quality devices. Those distortions degrade the perceptual quality of human listeners. Speech restoration is a task to restore degraded speech to high-quality speech, which is useful in a wide range of applications such as online meeting [1] and hearing aids [2].

Previous speech restoration methods mainly focus on a single type of distortion, such as speech denoising [3], dereverberation [4], super-resolution [5], and declipping [6]. However, in the real world, speech signals can be degraded by several different distortions simultaneously. These mismatches limit the performance of these systems. Several works have explored restoring speech with multiple distortions, such as noise and reverberation [7, 8]. But other distortions such as low-resolution and clipping receive less attention, despite their significant impacts on speech perceptual quality.

Speech fidelity is important to perceptual quality. However, existing methods show limited performance on high-fidelity speech restoration. For example, for a noisy speech with low bandwidth, although the speech denoising method could remove noises, the restored speech would be still in low fidelity. One way to address this issue is to concatenate speech restoration methods (e.g., denoising) with the speech super-resolution method.

However, this approach has limitations such as increasing computational cost and accumulating the artifacts introduced by each speech restoration model. To our knowledge, restoring low-bandwidth speech with multiple distortions has not been studied in the literature.

This paper introduces VoiceFixer, a unified framework for high-fidelity speech restoration. VoiceFixer restores speech from multiple distortions (e.g., noise, reverberation, and clipping) and could expand distorted speech with a low bandwidth between 1 kHz and 22.05 kHz to a full-bandwidth high-fidelity speech signal. We design VoiceFixer based on a two-stage strategy: (1) an analysis stage that performs mel spectrogram estimation; (2) a synthesis stage that generates the speech signal from the estimated mel spectrogram. Compared to the conventional speech restoration methods that operate on spectrogram or waveform, VoiceFixer uses the low dimensional mel spectrogram as the intermediate-level feature, which alleviates the difficulties of restoring multiple distortions simultaneously. In addition, neural vocoders [9] are usually trained on large-scale speech datasets. This provides prior knowledge on synthesizing waveform from low-dimensional mel spectrogram. The contributions of this paper are listed as follows:

- We present VoiceFixer, a unified framework for 44.1 kHz high-fidelity speech restoration. VoiceFixer can restore degraded speech from multiple distortions (e.g., noise, reverberation, clipping, and low-bandwidth).
- Evaluation result shows the effectiveness of VoiceFixer, which achieves a 0.256 higher mean opinion scores (MOS) than the baseline method.
- We release the pre-trained model and source code¹ of VoiceFixer to encourage future research.

The rest of this paper is organized as follows. Section 2 introduces the formulations of speech distortions we addressed. Section 3 describes the architecture of our proposed VoiceFixer. Experiments are presented in Section 4. In Section 5, we summarize this study and discuss our future directions.

2. Problem Formulation

We denote a segment of a speech signal as $s \in \mathbb{R}^L$, where L is the number of samples in the segment. We model the speech distortion process as function $d(\cdot)$. The degraded speech $x \in \mathbb{R}^L$ thus can be written as $x = d(s)$. Speech restoration aims to restore high-quality speech $\hat{s}$ from x by $\hat{s} = f(x)$, where $f(\cdot)$ is the restoration function and can be viewed as an inverse approximation of $d(\cdot)$. The target of the restoration function is to estimate s by restoring $\hat{s}$ from the degraded speech x .

* The first two authors contributed equally to this work.

† Part of this work was done during the internship at ByteDance

¹https://github.com/haoheliu/voicefixer_main

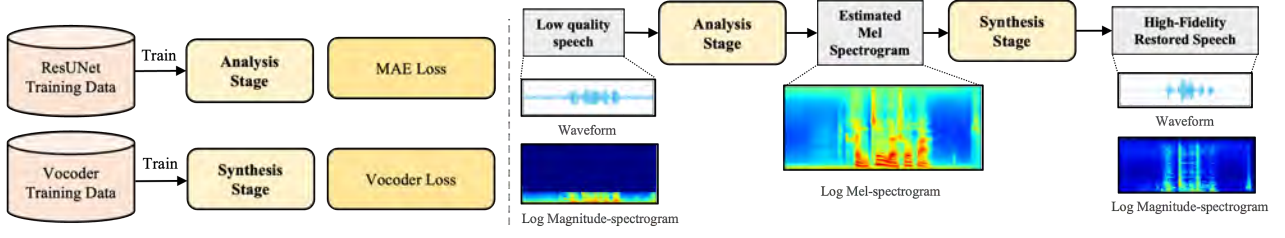


Figure 1: Overview of the proposed VoiceFixer framework. The analysis module and synthesis module are trained separately.

Distortion modeling is an important step to simulate training data when building speech restoration systems. Previous works model distortions in a sequential order [10, 11]. Similarly, we model the distortion $d(\cdot)$ as a composite function:

$$d(x) = d_1 \circ d_2 \circ \dots \circ d_Q(x), d_q \in \mathbb{D}, q = 1, 2, \dots, Q, \quad (1)$$

where $\circ$ stands for function composition and Q is the number of distortions consisted in $d(\cdot)$. $\mathbb{D} = \{d_v(\cdot)\}_{v=1}^V$ is the set of distortion types, where V is the total number of types. Equation 1 describes the procedure of compounding different distortions from $\mathbb{D}$ in a sequential order. The four types of speech distortions we addressed in this work are introduced as follows.

Additive noise is one of the most common distortion and can be modeled by the addition between speech s and noise $n \in \mathbb{R}^L$:

$$d_{\text{noise}}(s) = s + n. \quad (2)$$

Reverberation is caused by the reflections of signal within a space. Reverberation makes speech signals sound distant and blurred. It can be modeled by convolving speech signals with a room impulse response filter (RIR) r :

$$d_{\text{rev}}(s) = s * r, \quad (3)$$

where $*$ stands for convolution operation.

Clipping distortion refers to the clipped amplitude of audio signals when their amplitude exceeds the maximum level. Clipping can be modeled by restricting signal amplitudes within a range $[-\eta, +\eta]$:

$$d_{\text{clip}}(s) = \max(\min(s, \eta), -\eta), \eta \in [0, 1]. \quad (4)$$

In the frequency domain, the clipping effect produces harmonic components in the high-frequency part and degrades speech intelligibility accordingly.

Low-bandwidth distortion refers to the limited bandwidth in the audio recordings caused by low sampling rate or defects in the recording device. We follow the description in [12] to produce low-bandwidth distortions but add more filter types [13]. After designing a low pass filter h , we first convolve it with s to avoid the aliasing phenomenon. Then we perform resampling on the filtered result from the original bandwidth o to a lower bandwidth u :

$$d_{\text{low_bw}}(s) = \text{Resample}(s * h, o, u). \quad (5)$$

3. Approach

The two-stage strategy of VoiceFixer is formulated as follows:

$$f: x \mapsto z, \quad (6)$$

$$g: z \mapsto \hat{s}. \quad (7)$$

Equation 6 denotes the analysis stage of VoiceFixer where a distorted speech x is mapped into an intermediate-level feature z . Equation 7 denotes the synthesis stage of VoiceFixer, which synthesizes z to the restored speech $\hat{s}$. The overview of VoiceFixer framework is depicted in Figure 1.

3.1. Analysis stage

The goal of the analysis stage is to predict the intermediate representation z , which can be used later to recover the speech signal. In our study, we choose the mel spectrogram as the intermediate representation. Mel spectrogram has been widely used in tasks such as speech enhancement [14] and audio synthesis [15, 16]. The frequency dimension of the mel spectrogram is usually much smaller than the magnitude spectrogram calculated using short-time-fourier-transform (STFT), thus working on mel-scale can reduce the dimension of feature space and offer a more tractable restoration process. The objective of the analysis stage is to restore the mel spectrogram of the target signals, which can be written as follows:

$$\hat{S}_{\text{mel}} = f_{\text{mel}}(X_{\text{mel}}; \alpha) \odot (X_{\text{mel}} + \epsilon), \quad (8)$$

where X_{mel} is the mel spectrogram of x . It is calculated by $X_{\text{mel}} = |X|W$, where $|X|$ is the magnitude spectrogram of x and W is a set of mel filter banks. The columns of W are not divided by the numbers of mel bands, because this will make the restoration model difficult to recover the high-frequency part. The mapping function $f_{\text{mel}}(\cdot; \alpha)$ is the mel-restoration mask-estimation model parameterized by α . X_{mel} is added with a minimum value ϵ before multiplying with the output of f_{mel} . ϵ is set to 1×10^{-8} in this work to avoid zero values in X_{mel} .

We use ResUNet [17, 18] to model the analysis stage. ResUNet consists of six encoder and six decoder blocks. There are skip connections between encoder and decoder blocks at the same level. Both encoder and decoder blocks have a similar structure of four residual convolutions. Each residual convolution consists of a batch normalization, a leakyReLU activation, and a two-dimensional convolutional operation. We utilize average pooling and transpose convolution for the upsampling and downsampling in the encoder and decoder blocks. We will refer to ResUNet as UNet in the remaining parts. We optimize the model in the analysis stage using the MAE loss between the estimated and the target mel spectrogram, $\hat{S}_{\text{mel}}$ and S_{mel} :

$$\mathcal{L}_{\text{MAE}} = \left\| \hat{S}_{\text{mel}} - S_{\text{mel}} \right\|_1. \quad (9)$$

3.2. Synthesis stage

We realize the synthesis stage with a neural vocoder, which synthesizes the mel spectrogram into waveform, as denoted in the following Equation 10:

$$\hat{s} = g(X_{\text{mel}}; \beta), \quad (10)$$

where $g(\cdot; \beta)$ stands for the vocoder model parameterized by β . The number of speakers used for the training of vocoder is much larger than that used in the analysis stage, which increases the robustness of VoiceFixer when generalizing to unseen speakers. We employ a pre-trained² time and frequency domain-based generative adversarial network (TFGAN) [19] as a vocoder. TFGAN achieves strong performance on 44.1 kHz speaker-independent speech vocoding, which will be discussed in detail in Section 4.

4. Experiments

We conduct two types of experiments to evaluate the performance of VoiceFixer: (1) High-fidelity speech restoration from simultaneously appearing noise, reverberation, clipping, and low-bandwidth distortions; (2) Single type restoration from speech with only one type of distortion (e.g., denoising). In the following sections, we first describe the experimental data preparation, then present the results of these two experiments. The test sets used in this section are publicly available³.

4.1. Experimental data preparation

Training a speech restoration system relies on pairs of distorted speech and clean speech. In the high-fidelity speech restoration task, we simulate speech with multiple distortions. As introduced in Section 2, we simulate four types of speech distortion: additive noise, reverberation, clipping, and low-bandwidth. Three types of datasets are used for the simulation, including clean speech, noise data, and room impulse response (RIR). Note that clipping and low-resolution distortion only need the clean speech dataset for simulation and do not depend on other datasets. We introduce the three types of datasets as follows.

Clean speech we used is based on VCTK [20], which is a multi-speaker English corpus that consists of 110 speakers with different accents. The version of VCTK we used is 0.92. Following the setups in other studies [21], speakers p280 and p315 are omitted for the technical issues. The remaining part is split into a training set VCTK-Train with 98 speakers and a testing set VCTK-Test with the last 8 speakers. The remaining 2 speakers are omitted as they appear in the test set for denoising.

Noise data we used is based on two datasets. The first one is VCTK-Demand (VD) [22]. VD contains a training part VD-Train and a testing part VD-Test. Both parts contain clean speech and noisy speech data. To obtain the noise data from VD, we minus each noisy data in VD-Train with its corresponding clean part to get the noise dataset VD-Noise for training. The second noise dataset we use is the TUT urban acoustic scenes 2018 dataset [23], which contains 89 hours of high-quality recording from 10 acoustic scenes (e.g., airport). This dataset contains a development part and an evaluation part. We only use the evaluation part (DCASE-Eval) for the simulation of the test set for high-fidelity speech restoration.

Room impulse response is randomly simulated to add reverberation effect on 44.1 kHz speech. Simulation is performed using an open-source tool⁴. All the related parameters are randomized, including the size of the room, the placement of the microphone and the sound source, the RT60 value, and the pickup pattern of the microphone. In total, 43239 filters are simulated, in which we randomly split out 5000 filters as the test set RIR-Test and named the other 38239 filters as RIR-Train.

²<https://github.com/haoheliu/voicefixer>

³<https://zenodo.org/record/5528144>

⁴https://github.com/sunits/rir_simulator_python

4.2. High-Fidelity speech restoration

4.2.1. Data sets and distortion modeling

In this task, we simulate low-quality speech with four distortions in the training and test set, including noise, reverberation, clipping, and low bandwidth. Training data is simulated on the fly based on the speech data in VCTK-Train, the noise in VD-Noise, and RIR in RIR-Train. We set up the parameters of each distortion to be completely random to better cover the real-world cases. The test set we used in high-fidelity speech restoration, HiFi-Res, is constructed based on the clean speech in VCTK-Test, the noise in DCASE-Eval, and RIR in RIR-Test. HiFi-Res consists of 501 three seconds utterances with similar random distortions simulated as the training process. We first generate the distortions following specific order: reverberation, noise, and clipping. Then the degraded speech is low-pass filtered and down-sampled to an arbitrary low sampling rate between 2 kHz to 44.1 kHz. Details of the distortion modeling in this work are made available on GitHub⁵.

4.2.2. Experiment details

All the audio files in our datasets are resampled to 44.1 kHz sampling rate. We calculate STFT using the Hanning window with a window length of 2048 and a hop size of 441. The mel filterbank we used consists of 128 filters. For training, We use Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, an initial learning rate of 3×10^{-4} and a batch size of 24. The first 1000 steps are warmup steps, during which the learning rate grows linearly from 0 to 3×10^{-4} . The learning rate is scheduled for decay by 0.9 every 400 hours of training data. We trained our model using four Nvidia-V100-32GB GPUs for two days.

4.2.3. Baseline systems

We mainly use four baseline systems in the experiment. We implemented an UNet-based system (Baseline-UNet) for the high-fidelity speech restoration task, which structure is similar to the analysis module of VoiceFixer. It performs restoration by estimating STFT of the high-quality speech and reusing the phase of the degraded speech, which is a common approach in previous speech restoration systems [24]. As for the Oracle-Mel system, we directly use the target mel spectrogram as input to the vocoder to simulate the case when the analysis module works ideally. So, Oracle-Mel marks the theoretical upper bound of the VoiceFixer performance. For the Target system, scores are calculated using the ground truth clean speech. Conversely, the Unprocessed system evaluated directly on the distorted speech.

4.2.4. Evaluation metrics

We use both objective and subjective evaluation metrics. The objective metrics including log-spectral distance (LSD) [25], wide-band perceptual evaluation of speech quality (PESQ-wb) [26], and structural similarity (SSIM) [27]. Since neural vocoders generate waveforms directly from mel spectrograms, even with the same perceptual quality, the generated waveforms may not align with the target waveform in the time domain. This misalignment can considerably degrade the objective metrics, as is often the case in generative model [28]. Nevertheless, we report the model performance on these objective metrics for reference.

We use mean opinion scores (MOS) as the subjective evaluation metric and invite eight internal language experts in

⁵https://github.com/haoheliu/voicefixer_main

Table 1: Evaluation result on the high-fidelity speech restoration test set HiFi-Res. Higher PESQ-wb, SSIM, MOS value indicates better performance, while LSD is the opposite. The best value for each metric is shown in bold.

Models	PESQ-wb	LSD	SSIM	MOS
Unprocessed	1.94	2.00	0.64	2.38
Oracle-Mel	2.52	0.91	0.74	3.74
Target	4.64	0.01	1.00	3.95
Baseline-UNet	2.67	1.01	0.79	3.37
VoiceFixer	2.05	1.01	0.71	3.62

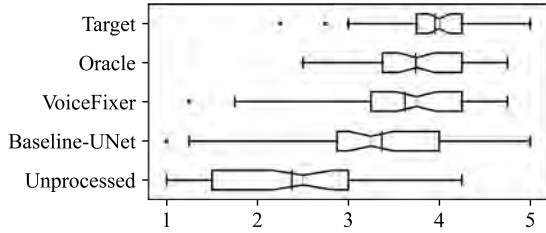


Figure 2: Box plot of the MOS scores on HiFi-Res test set. Red and black vertical lines represent median and mean values.

ByteDance to perform evaluation. Their task is to rate the overall speech quality of an audio clip with a score between 1 (bad) to 5 (excellent). Each system has 38 samples for evaluation. We average the MOS values across all language experts as the final result.

4.2.5. Evaluation results

Table 1 shows the experimental results and Figure 2 depicts the box plot of the MOS scores. The Oracle-Mel system achieves a MOS score of 3.74, which is close to the Target MOS of 3.95, indicating that the vocoder performs well in the synthesis stage. We observe that VoiceFixer obtains 0.256 higher MOS score than that of Baseline-UNet and is only 0.11 lower than the Oracle-Mel, demonstrating its good performance for high-fidelity speech restoration. Although VoiceFixer performs worse on PESQ-wb and SSIM metrics, it has a much better MOS score than Baseline-UNet. This result shows that the improvement in subjective metrics in VoiceFixer is not always consistent with objective evaluations.

4.3. Single type restoration

To further demonstrate the effectiveness of VoiceFixer, we conduct two benchmark speech restoration experiments: speech denoising and speech declipping.

4.3.1. Denoising

For speech denoising, we evaluate the model performance on VD-Test (as described in Section 4.1). VD-Test contains 824 utterances from a female speaker and a male speaker. The test set is simulated at four SNR levels, which are 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. The original data is sampled at 48 kHz. We downsample it to 44.1 kHz to fit our experiments. We adopt three recent methods SEGAN [29], WaveUNet [30], and the model trained with weakly labeled data [31] (referred to WL-Model) as baseline methods.

We use the objective metric PESQ-wb [26] and the subjective metric MOS to evaluate model performance. Experimental

Table 2: Evaluation result on the VD-Test test set. Superscript * indicates the model is only trained on a single restoration task.

Models	PESQ-wb	MOS
Unprocessed	1.97	3.20
Oracle-Mel	2.85	3.64
Target	4.50	3.69
*SEGAN [29]	2.16	/
*WaveUNet [30]	2.40	/
*WL-Model [31]	2.28	/
Baseline-UNet	2.82	3.64
VoiceFixer	2.43	3.69

Table 3: Evaluation result on the declipping test set DECLI.

Clipping Level	0.25		0.1	
Models	STOI	MOS	STOI	MOS
Unprocessed	0.95	2.56	0.89	2.72
Oracle-Mel	0.81	3.44	0.81	3.42
Target	1.00	3.42	1.00	3.49
*SSPADE [32]	0.98	3.34	0.92	2.63
Baseline-UNet	0.97	3.38	0.96	3.38
VoiceFixer	0.82	3.38	0.80	3.38

results are shown in Table 2. The PESQ-wb score of VoiceFixer reaches 2.43, higher than SEGAN, WaveUNet, and WL-Model. The MOS evaluations demonstrate that VoiceFixer outperforms the baseline speech denoising model Baseline-UNet. In addition, we observe that VoiceFixer even outperforms Oracle-Mel and achieves the same level as Target on the MOS scores. This is because the restored results of the VoiceFixer contain more energy in the high-frequency part, which potentially leads to a better perceptual quality for the listener.

4.3.2. Declipping

For the declipping task, we compare VoiceFixer with a state-of-the-art synthesis-based method SSPADE [32]. To evaluate the model performance, we create a test set DECLI based on VCTK-Test (as described in Section 4.1). DECLI is constructed by first normalizing the amplitude of VCTK-Test into $[-1, 1]$, and then simulating clipping on each audio with two clipping levels 0.25 and 0.1. This resulted in two declipping test sets, each containing 2937 clipped and clean speech audios.

We adopt MOS as the subjective metric and STOI [33] as the objective metric. A higher STOI value indicates better performance. Experimental results are shown in Table 3. VoiceFixer outperforms SSPADE on MOS by 0.04 and 1.25 in 0.25 and 0.1 clipping levels, respectively. The higher performance on MOS demonstrates a better perceptual quality restoration offered by VoiceFixer on speech declipping.

5. Conclusion

In this study, we propose VoiceFixer, an effective approach for high-fidelity speech restoration. VoiceFixer consists of an analysis stage modeled by a ResUNet and a synthesis stage using a TF-GAN. The two stages can also be replaced by other deep learning models. The subjective evaluation results show that VoiceFixer achieves superior performance on high-fidelity speech restoration from distortions such as noise, reverberation, clipping, and low bandwidth. In the future, VoiceFixer will be extended to more types of distortions.

6. References

- [1] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” *arXiv:2006.12847*, 2020.
- [2] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, “Speech enhancement with multichannel wiener filter techniques in multi-microphone binaural hearing aids,” *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.
- [3] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.
- [4] J. Zhang, M. D. Plumbley, and W. Wang, “Weighted magnitude-phase loss for speech dereverberation,” in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2021, pp. 5794–5798.
- [5] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks,” *arXiv:1708.00853*, 2017.
- [6] P. Závřiska, P. Rajmic, A. Ozerov, and L. Rencker, “A survey and an extensive evaluation of popular audio declipping methods,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 5–24, 2020.
- [7] Y. Ai, H. Li, X. Wang, J. Yamagishi, and Z. Ling, “Denoising-and-dereverberation hierarchical neural vocoder for robust waveform generation,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 477–484.
- [8] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” *arXiv:2006.05694*, 2020.
- [9] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *arXiv:2010.05646*, 2020.
- [10] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, “Audio-visual speech separation and dereverberation with a two-stage multimodal network,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, 2020.
- [11] X. Shu, Y. Zhu, Y. Chen, L. Chen, H. Liu, C. Huang, and Y. Wang, “Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation,” *arXiv:2107.09298*, 2021.
- [12] H. Wang and D. Wang, “Towards robust speech super-resolution,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021.
- [13] S. Sulun and M. E. Davies, “On filter generalization for music bandwidth extension using deep neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 132–142, 2020.
- [14] S. Maiti and M. I. Mandel, “Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement,” in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2020, pp. 206–210.
- [15] K. Kumar, R. Kumar, T. de Boissiere, L. Gustin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, “Mel-GAN: Generative adversarial networks for conditional waveform synthesis,” *arXiv:1910.06711*, 2019.
- [16] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, “Conditional sound generation using neural discrete time-frequency representation learning,” in *IEEE International Workshop on Machine Learning for Signal Processing*, 2021, pp. 1–6.
- [17] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, “Decoupling magnitude and phase estimation with deep resnet for music source separation,” in *The International Society for Music Information Retrieval*, 2021.
- [18] H. Liu, Q. Kong, and J. Liu, “CWS-PResUNet: Music source separation with channel-wise subband phase-aware resnet,” *arXiv:2112.04685*, 2021.
- [19] Q. Tian, Y. Chen, Z. Zhang, H. Lu, L. Chen, L. Xie, and S. Liu, “TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis,” *arXiv:2011.12206*, 2020.
- [20] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, “CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2019.
- [21] J. Lee and S. Han, “NU-wave: A diffusion probabilistic model for neural audio upsampling,” *arXiv:2104.02321*, 2021.
- [22] C. Valentini-Botinhao *et al.*, “Noisy speech database for training speech enhancement algorithms and TTS models,” 2017.
- [23] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” *arXiv:1807.09840*, 2018.
- [24] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, “Phase-aware single-stage speech denoising and dereverberation with U-Net,” *arXiv:2006.00687*, 2020.
- [25] A. Erell and M. Weintraub, “Estimation using log-spectral-distance criterion for noise-robust speech recognition,” in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 853–856.
- [26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs,” in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 749–752.
- [27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [28] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, “NU-GAN: High resolution neural upsampling with GAN,” *arXiv:2010.11362*, 2020.
- [29] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” in *INTERSPEECH*, 2017.
- [30] C. Macartney and T. Weyde, “Improved speech enhancement with the Wave-U-Net,” *arXiv:1811.11307*, 2018.
- [31] Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, “Speech enhancement with weakly labelled data from AudioSet,” in *INTER-SPEECH*, 2021.
- [32] P. Závřiska, P. Rajmic, O. Mokř, and Z. Průša, “A proper version of synthesis-based sparse audio declipper,” in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 2019, pp. 591–595.
- [33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

Appendix ZHA001



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN | UPLOAD

Search

ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER

INTERNET ARCHIVE

DONATE

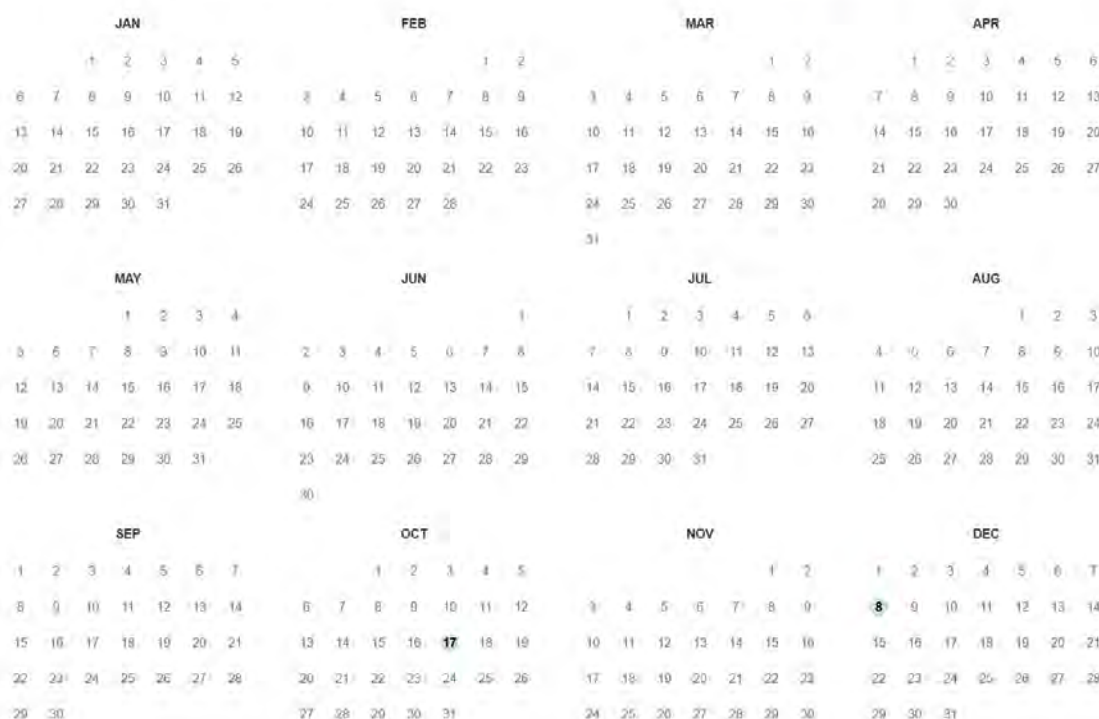
Wayback Machine

Explore more than 1 trillion web pages saved over time

https://psi.engr.tamu.edu/l2-arctic-corpus/

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

Saved 61 times between October 17, 2019 and July 6, 2025.



Note

This calendar view maps the number of times <https://psi.engr.tamu.edu/l2-arctic-corpus/> 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](#).



L2-ARCTIC: a non-native English speech corpus

News

- 6/6/2019: **v4.0** is available. We re-examined some of the annotations and changed most of the "err" tags to more detailed (and informative) annotations — marking them as different deviations from standard English
- 4/8/2019: **v3.0** is available, we added four (4) Vietnamese speakers!
- 9/28/2018: **v2.0** is available, we added 10 new speakers!
- 8/20/2018: uploaded corpus description paper
- 6/3/2018: the corpus description paper was accepted at Interspeech 18! We will make the paper available online soon, and more speakers will be added over the following months
- 3/26/2018, **v1.0** was released

Introduction

Welcome to the homepage of L2-ARCTIC, a speech corpus of non-native English intended for research in voice conversion, accent conversion, and mispronunciation detection. This corpus includes recordings from twenty-four (24) non-native speakers of English whose first languages (L1s) are Hindi, Korean, Mandarin, Spanish, Arabic and Vietnamese, each L1 containing recordings from two male and two female speakers. Each speaker recorded approximately one hour of read speech from CMU's **ARCTIC** prompts, from which we generated orthographic and forced-aligned phonetic transcriptions. In addition, we manually annotated 150 utterances per speaker to identify three types of mispronunciation errors: substitutions, deletions, and additions, making it a valuable resource not only for research in voice conversion and accent conversion but also in computer-assisted pronunciation training.

The corpus is a joint effort of researchers at [Texas A&M University](#) and [Iowa State University](#). In the future, we may also include speakers from other L1s if we find them to be useful to the research community.

Overview

For each speaker, the corpus contains the following data:

- **Speech recordings:** over one hour of prompted recordings of phonetically-balanced short sentences (~1132)
- **Word level transcriptions:** orthographic transcription and forced-aligned word boundaries for each sentence
- **Phoneme level transcriptions:** forced-aligned phoneme transcription for each sentence
- **Manual annotations:** a selected subset of utterances (~150), including 100 sentences produced by all speakers and 50 sentences that include phonemes likely to be difficult according to each speaker's L1, all annotated with corrected word and phone boundaries; phone substitution, deletion, and addition errors are also tagged

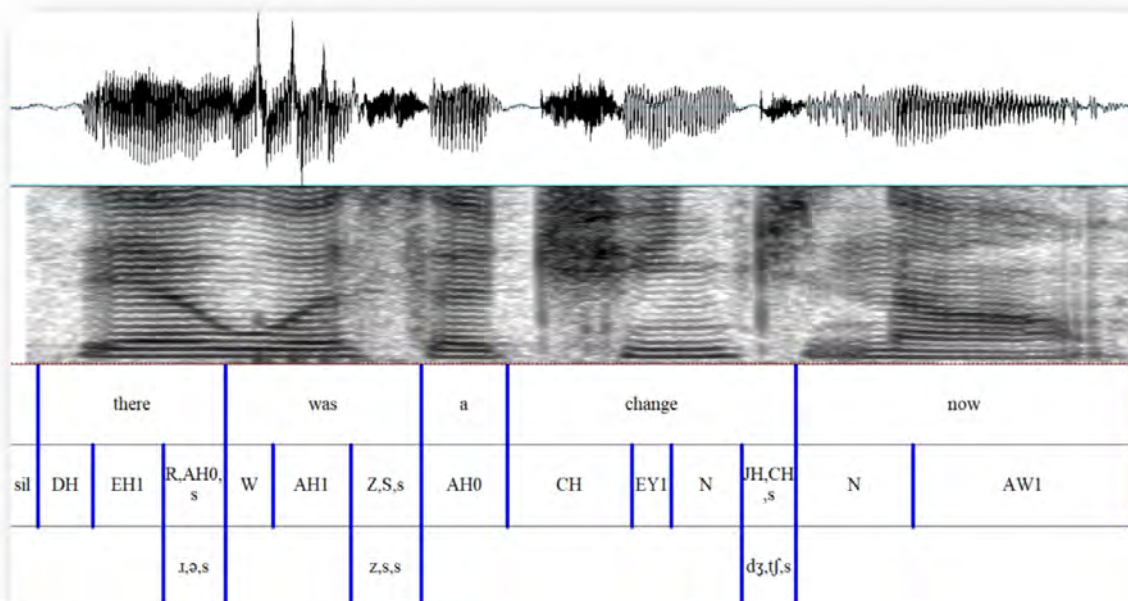
Dataset examples

Audio

Speaker	L1	Gender	Audio
ABA	Arabic	M	
SKA	Arabic	F	
YBAA	Arabic	M	
ZHAA	Arabic	F	
BWC	Mandarin	M	
LXC	Mandarin	F	
NCC	Mandarin	F	
TXHC	Mandarin	M	

ASI	Hindi	M	▶
RRBI	Hindi	M	▶
SVBI	Hindi	F	▶
TNI	Hindi	F	▶
HJK	Korean	F	▶
HKK	Korean	M	▶
YDCK	Korean	F	▶
YKWK	Korean	M	▶
EBVS	Spanish	M	▶
ERMS	Spanish	M	▶
MBMPS	Spanish	F	▶
NJS	Spanish	F	▶
HQTV	Vietnamese	M	▶
PNV	Vietnamese	F	▶
THV	Vietnamese	F	▶
TLV	Vietnamese	M	▶

Annotations



A TextGrid with manual annotations (LXC, arctic_a0018; [wav file](#), [TextGrid file](#)). Top to bottom: speech waveform, spectrogram, words, phonemes, error tags, and comments from the annotator

Access guidelines

Please first review the license terms of L2-ARCTIC. If you agree to those terms, please go to the download section below and fill in your name, email address, and affiliation, then click "Download." An automated email will be sent to the address you provided with the direct download link. Please make sure to provide a correct email address.

Generally, the email will be sent to you immediately. If you haven't received it within ten minutes, please check your spam box and look for an email titled "Access to L2-ARCTIC corpus" or add "gzhao@tamu.edu" to your whitelist and then submit the download form again. If you encountered any difficulty accessing L2-ARCTIC, please feel free to contact Guanlong Zhao (gzhao@tamu.edu) for assistance.

IPR2026-00272

Documentation

Link to [documentation](#) and [publication](#).

License

The corpus is released under the CC BY-NC 4.0 license, a summary of the license can be found [here](#), and the full license can be found [here](#). For any usage that is not covered by the CC BY-NC 4.0 license, please contact Dr. Ricardo Gutierrez-Osuna (rgutier@tamu.edu).

Please cite the following paper if you used L2-ARCTIC for any publications.

```
@inproceedings{zhao2018l2arctic,
```

```
author={Guanlong [Zhao] and Sinem [Sonsaat] and Alif [Silpachai] and Ivana [Lucic] and Evgeny [Chukharev-Hudilainen] and John [Levis] and Ricardo [Gutierrez-Osuna]},
```

```
title={L2-ARCTIC: A Non-native English Speech Corpus},
```

```
year=2018,
```

```
booktitle={Proc. Interspeech},
```

```
pages={2783–2787},
```

```
doi={10.21437/Interspeech.2018-1110},
```

```
url={http://dx.doi.org/10.21437/Interspeech.2018-1110}
```

```
}
```

Download

Your Name (required)

Your Email (required)

Your Affiliation (required)

☐ I agree to the license terms of L2-ARCTIC

DOWNLOAD

Contact

Guanlong Zhao (gzhao@tamu.edu), Department of Computer Science & Engineering @ TAMU

Sinem Sonsaat (sonsaat@iastate.edu), Department of English @ Iowa State University

Alif Silpachai (alif@iastate.edu), Department of English @ Iowa State University

Ivana Lucic (ilucic@iastate.edu), Department of English @ Iowa State University

Evgeny Chukharev-Hudilainen (evgeny@iastate.edu), Department of English @ Iowa State University

John Levis (jlevis@iastate.edu), Department of English @ Iowa State University

Ricardo Gutierrez-Osuna (rgutier@tamu.edu), Department of Computer Science & Engineering @ TAMU

Acknowledgments

Curation of the L2-ARCTIC corpus was supported by NSF awards 1619212 and 1623750. We would like to thank the anonymous participants for recording the corpus. We also would like to thank Ziwei Zhou and Taylor Anne Barriuso for their assistance with the annotations. We appreciate suggestions from Christopher Liberatore, Shaojin Ding, and the reviewers at Interspeech'18.

External links

[CMU ARCTIC speech database](#)

[Speech Accent Archive](#)

[IDEA: The International Dialects of English Archive](#)

[Kaldi-gop](#)

[Dr. John Levis' homepage](#)

[Dr. Sinem Sonsaat's homepage](#)

[Dr. Evgeny Chukharev-Hudilainen's homepage](#)

[Alif Silpachai's homepage](#)

Physical: 506 HR Bright Building, Texas
A&M University
Mailing: PSI Lab, 3112 TAMU, College
Station, TX 77843



Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.

Appendix ZHA002



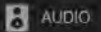
WEB



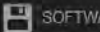
TEXTS



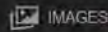
VIDEO



AUDIO



SOFTWARE



IMAGES



SIGN UP



LOG IN



UPLOAD



Search

ABOUT

BLOG

EVENTS

PROJECTS

HELP

DONATE

CONTACT

JOBS

VOLUNTEER

INTERNET ARCHIVE

DONATE

Wayback Machine

Explore more than 1 trillion web pages saved over time

https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1110.pdf

Calendar

Collections

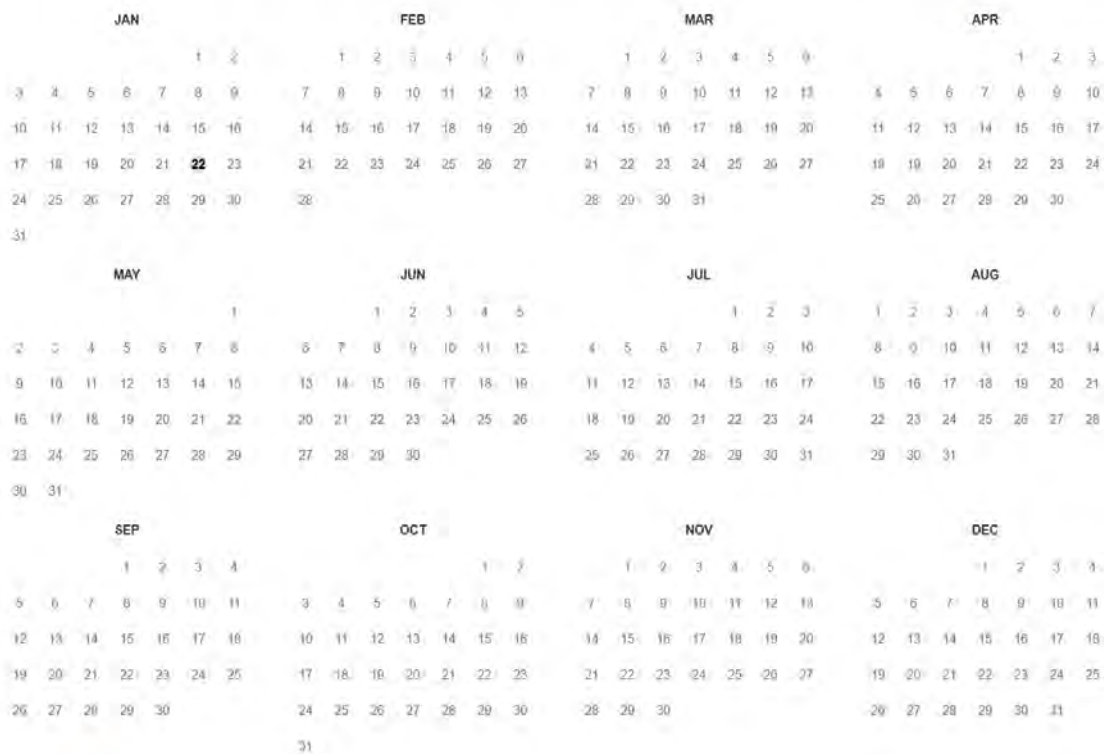
Changes

Summary

Site Map

URLs

Saved 4 times between January 22, 2021 and November 3, 2024.



Note

This calendar view maps the number of times https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1110.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](#).



1



2

Interspeech 2018
2-6 September 2018, Hyderabad



L2-ARCTIC: A Non-Native English Speech Corpus

Guanlong Zhao¹, Sinem Sonsaat², Alif Silpachai², Ivana Lucic²
Evgeny Chukharev-Hudilainen², John Levis² and Ricardo Gutierrez-Osuna¹

¹Department of Computer Science and Engineering, Texas A&M University, United States

²Department of English, Iowa State University, United States

{gzhao, rgutier}@tamu.edu, {sonsaat, alif, ilucic, evgeny, jlevis}@iastate.edu

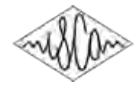
Abstract

In this paper, we introduce L2-ARCTIC, a speech corpus of non-native English that is intended for research in voice conversion, accent conversion, and mispronunciation detection. This *initial* release includes recordings from ten non-native speakers of English whose first languages (L1s) are Hindi, Korean, Mandarin, Spanish, and Arabic, each L1 containing recordings from one male and one female speaker. Each speaker recorded approximately one hour of read speech from the Carnegie Mellon University ARCTIC prompts, from which we generated orthographic and forced-aligned phonetic transcriptions. In addition, we manually annotated 150 utterances per speaker to identify three types of

corpora (e.g., Speech Accent Archive [7] and IDEA [8]) do not fulfill these requirements (refer to section 2 for a detailed discussion.)

To fill this gap, we have built a non-native English speech corpus that contains ten non-native speakers of English in the initial release. The end goal for this corpus is to include 20 speakers from five different native languages: Hindi, Korean, Mandarin, Spanish, and Arabic. For *each* speaker, the corpus contains the following data:

- Speech recordings: over one hour of prompted recordings of phonetically-balanced short sentences
- Word level transcriptions: orthographic transcription and forced-aligned word boundaries for each sentence



L2-ARCTIC: A Non-Native English Speech Corpus

Guanlong Zhao¹, Sinem Sonsaat², Alif Silpachai², Ivana Lucic²
Evgeny Chukharev-Hudilainen², John Levis² and Ricardo Gutierrez-Osuna¹

¹Department of Computer Science and Engineering, Texas A&M University, United States

²Department of English, Iowa State University, United States

{gzhao, rgutier}@tamu.edu, {sonsaat, alif, ilucic, evgeny, jlevis}@iastate.edu

Abstract

In this paper, we introduce L2-ARCTIC, a speech corpus of non-native English that is intended for research in voice conversion, accent conversion, and mispronunciation detection. This *initial* release includes recordings from ten non-native speakers of English whose first languages (L1s) are Hindi, Korean, Mandarin, Spanish, and Arabic, each L1 containing recordings from one male and one female speaker. Each speaker recorded approximately one hour of read speech from the Carnegie Mellon University ARCTIC prompts, from which we generated orthographic and forced-aligned phonetic transcriptions. In addition, we manually annotated 150 utterances per speaker to identify three types of mispronunciation errors: substitutions, deletions, and additions, making it a valuable resource not only for research in voice conversion and accent conversion but also in computer-assisted pronunciation training. The corpus is publicly accessible at <https://psi.engr.tamu.edu/l2-arctic-corpus/>.

Index Terms: speech corpus, voice conversion, accent conversion, mispronunciation detection

1. Introduction

Voice conversion (VC) [1] aims to transform utterances from a source speaker to make them sound as if a target speaker had uttered them. The closely related problem of accent conversion (AC) [2] goes a step further, mixing the source speech's linguistic content and accent with the target speaker's voice quality to create utterances with the target's voice but the content and pronunciation of the source speaker. When teaching a second language (L2), accent conversion can be used to create a "golden speaker," a synthesized voice that has the learner's voice quality but with a native speaker's accent (e.g., prosody, intonation, pronunciation) [3]. Several studies [4, 5] have suggested that having such a "golden speaker" to imitate can be beneficial in pronunciation training. Furthermore, in addition to providing language learners with a suitable voice to mimic, detecting mispronunciations is also a critical component for providing useful feedback to the learners in computer-assisted pronunciation training [6].

To train and evaluate voice and accent conversion systems designed for non-native speakers, one needs high-quality parallel recordings from the source and target speakers. Likewise, to develop and benchmark mispronunciation detection algorithms, detailed phoneme level annotations on pronunciation errors (e.g., phone substitution, additions, and deletions) are required. However, existing non-native English

corpora (e.g., Speech Accent Archive [7] and IDEA [8]) do not fulfill these requirements (refer to section 2 for a detailed discussion.)

To fill this gap, we have built a non-native English speech corpus that contains ten non-native speakers of English in the initial release. The end goal for this corpus is to include 20 speakers from five different native languages: Hindi, Korean, Mandarin, Spanish, and Arabic. For *each* speaker, the corpus contains the following data:

- Speech recordings: over one hour of prompted recordings of phonetically-balanced short sentences
- Word level transcriptions: orthographic transcription and forced-aligned word boundaries for each sentence
- Phoneme level transcriptions: forced-aligned phoneme transcription for each sentence
- Manual annotations: a selected subset of utterances (~150), including 100 sentences produced by all speakers and 50 sentences that include phonemes likely to be difficult according to each speaker's L1, all annotated with corrected word and phone boundaries; phone substitution, deletion, and addition errors are also tagged

The dataset is hosted on an online archive and is freely available to the research community for non-commercial use. To the best of our knowledge, L2-ARCTIC is the first openly available corpus of its kind.

2. The need for a new L2 English corpus

A number of voice conversion studies [9-12] have relied on the Carnegie Mellon University (CMU) ARCTIC speech corpus [13] and, more recently, the Voice Conversion Challenge (VCC) dataset [14]. However, little attention has been paid to voice conversion between non-native speakers of English, in part due to the lack of high-quality speech recordings from those speakers, despite 80% of the English speakers in the world being non-native [15]. For example, CMU ARCTIC only has a few accented English speakers¹, either native speakers of different English dialects or highly proficient non-native speakers, whereas the VCC dataset was recorded solely by professional voice talents who are native English speakers. Therefore, these standard corpora are not suitable for either voice conversion between non-native speakers nor accent conversion tasks.

Among the non-native English corpora, the Speech Accent Archive [7] and IDEA [8] cover a wide range of native languages and speakers. However, each speaker only recorded a short paragraph (Speech Accent Archive) or a short free speech task (IDEA), and most of the recordings have strong

¹ JMK: Canadian accent; AWB: Scottish accent; KSP: Indian accent

background noise, making them ill-suited for voice/accent conversion. The Wildcat [16], LDC2007S08 [17], and NUFAESD [18] datasets have a limited number of recordings for each non-native speaker, and have restricted access – LDC2007S08 requires a fee, while Wildcat and NUFAESD are limited to designated research groups.

As for corpora for mispronunciation detection, the CU-CHLOE [19] and College Learners’ Spoken English Corpus (COLSEC) [20] only contain speech and error tags from Chinese learners of English, and CU-CHLOE is (to our knowledge) not publicly available. The ISLE Speech Corpus [21] contains mispronunciation tags and is open for academic access, but it only focuses on a limited group of English learners (German and Italian). SingaKids-Mandarin [22] has a rich set of speech data, but it only focuses on mispronunciation patterns in Singapore children’s Mandarin speech. In fact, most existing mispronunciation detection systems use their private datasets, which makes it difficult to compare experimental results across different publications [19, 23–25].

To overcome the insufficiencies outlined above, we constructed (and are now releasing) L2-ARCTIC to provide an open corpus for voice conversion between accented speakers, accent conversion, and mispronunciation detection. Zhao et al. [26] have performed a preliminary evaluation on voice/accent conversion tasks using a subset of the speakers in L2-ARCTIC. Using a joint-density GMM with MLPG and global variance compensation [9] (128 mixtures, ~5 min of parallel training data) as the voice conversion system, they obtained 2.5 Mean Opinion Score (MOS) on the converted speech, which was also rated as similar to the target voice. Furthermore, an accent-conversion algorithm based on frame-alignment using posteriorgrams was able to generate speech that was perceived as similar to a non-native target voice but markedly less accented (98% preference compared to non-native speech). This manuscript presents preliminary results on a new task: mispronunciation detection.

3. Corpus curation procedure

This initial release of L2-ARCTIC contains English speech of speakers from five different first languages: Hindi¹ [27], Korean, Mandarin, Spanish, and Arabic. We chose these L1s because each one has a distinct foreign/non-native accent in English and provides unique challenges. **Indian** speakers of English typically have native-like English fluency but use segmental and suprasegmental features in ways that are distinct from American English. Thus, Indian speakers have both advantages in approaching pronunciation changes (e.g., familiarity and comfort with English) and disadvantages (comfort with their English variety makes it particularly difficult to adjust their speech to salient differences with American English.) **Korean** learners of English have a large number of high functional load consonant and vowel difficulties (errors with many minimal pairs). Prosodically, Korean and English employ suprasegmental systems that have little overlap [28, 29]. **Mandarin** (Chinese/Putonghua) learners of English have difficulty with a range of consonant and vowel sounds and in producing correct English stress, intonation, and juncture [30–32]. **Spanish** learners of English may have difficulties distinguishing a number of high functional load contrasts in English [33, 34]. Spanish is also a five-vowel

language, and Spanish learners find the more complex English vowel system especially challenging. Like English, Spanish uses both word stress and nuclear stress for emphasis but, because it does not use the unstressed vowel schwa, realizes stress differently. Finally, **Arabic** also has significantly fewer vowels than English, and while Arabic has word stress, it does not use stress in the same way that English does [35, 36]. In the future, we may also include speakers from other L1s if we find them to be useful to the research community.

3.1. Participants

For this initial release, we recruited two speakers (one male and one female) for each of the L1s, for a total of ten speakers. Speakers were recruited from Iowa State University’s student body; their age range was from 22 to 43 years, with an average of 29 years (std: 6.9.) Demographic information of the speakers is summarized in Table 1. The proficiency level of English was measured using TOEFL iBT scores [37].

Table 1: *Demographic information of the speakers*

Speaker	L1	Gender	TOEFL iBT
HKK	Korean	M	114
YDCK	Korean	F	110
BWC	Mandarin	M	80
LXC	Mandarin	F	86
YBAA	Arabic	M	100
SKA	Arabic	F	79
EBVS	Spanish	M	70
NJS	Spanish	F	110
RRBI	Hindi	M	91
TNI	Hindi	F	99

3.2. Recording the corpus

To create the corpus, we used the 1,132 sentences in the CMU ARCTIC prompts. There were multiple reasons to choose these sentences. First, the ARCTIC prompts are phonetically balanced (100%, 79.6%, and 13.7% coverage for phonemes, diphones, and triphones, respectively), are open source, and can produce around one hour of edited speech. Second, the ARCTIC corpus itself has proven to work well with speech synthesis [38] and voice conversion tasks [9–11, 39]. Finally, the ARCTIC prompts are challenging for non-native English speakers so they can elicit potential pronunciation problems.

The speech was recorded in a quiet room at Iowa State University (ISU). We used a Samson C03U microphone and Earamble studio microphone pop filter for recordings; the microphone was placed 20 cm from the speaker to avoid air puffing. During each recording session, a linguist guided the L2 speaker through the process, asking the speaker to re-record a sentence if the production contained significant disfluency or deviated from the prompt. All speakers were instructed to speak in a natural manner. The speech was sampled at 44.1 kHz and saved as a WAV file.

Once the recording was finished, we removed repetitions and false starts, performed amplitude normalization, and segmented the utterances into individual WAV files. All of the above were done in Audacity [40]. The utterances were carefully trimmed to remove the leading and trailing silence and non-speech sounds such as lip smacks.

¹ Hindi is an Indo-Aryan language that is both an L1 and a language of wider communication. Thus Hindi speakers in the corpus may use Hindi

as an L2, speaking another Indian language as an L1. Educated Indian English is a stable contact variety of English.

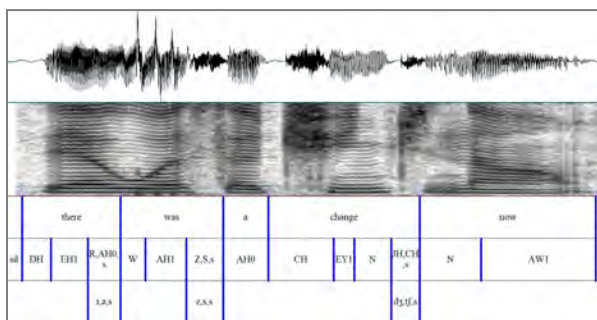


Figure 1: A TextGrid with manual annotations. Top to bottom: speech waveform, spectrogram, words, phonemes and error tags, comments from the annotator

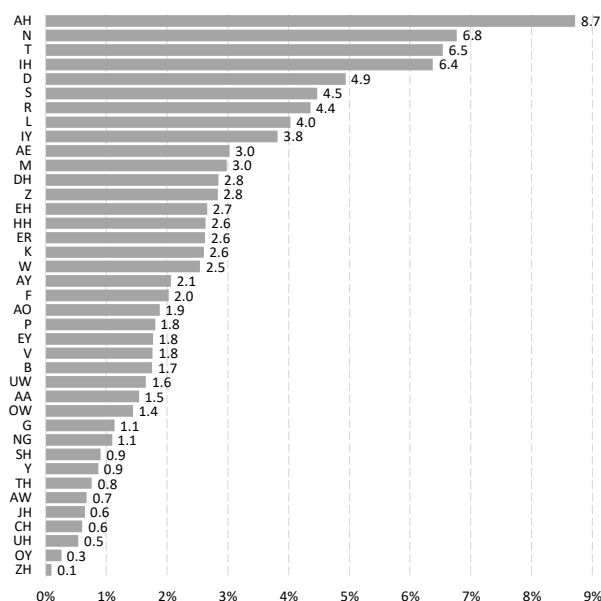


Figure 2: Phoneme distribution of the corpus

3.3. Corpus annotations

Our corpus provides orthographic transcriptions at the word level. We used the Montreal forced-aligner [41] to produce phonetic transcriptions in PRAAT's TextGrid format [42], which contains word and phone boundaries (Figure 1). Further, we performed manual annotations on a selected subset of sentences for each speaker. For all the speakers, we annotated a common set of 100 sentences. In addition, we annotated 50 sentences that included phoneme difficulties that were L1-dependent. In the end, the corpus contains up to 150 curated phonetic transcriptions per speaker¹. Those transcriptions contain manually-adjusted word and phone boundaries, correct phoneme labels, mispronunciation error tags (phone additions, deletions, and substitutions), and comments from the annotators. To facilitate computer processing, we used the ARPabet phoneme set for the phonetic transcriptions as well as the error tags. In the comment part of the transcriptions, however, annotators were allowed to use IPA symbols. To ensure high-quality annotations, we developed automated scripts to check the

annotation consistency and then asked human annotators to fix problems. The annotators (N=3) were PhD students in the Applied Linguistics and Technology program at ISU. They were experienced in transcribing speech samples of native or non-native English speakers.

4. Corpus statistics

In total, the dataset contains 11,026 utterances, with most speakers recording the full ARCTIC set (1,132 utterances.)¹ The total duration of the corpus is 11.2 hours, with an average of 67 minutes (std: 9 minutes) of speech per L2 speaker. On average, each utterance is 3.7 seconds in duration. The pause before and after each utterance is generally no longer than 100 ms. Using the forced alignment results, we estimate a speech to silence ratio of 7:1 across the whole dataset. The dataset contains over 97,000 word segments, giving an average of around nine words per utterance, and over 349,000 phone segments (excluding silence). The phoneme distribution is shown in Figure 2.

Human annotators manually examined 1,499 utterances, annotating 5,199 phone substitutions, 1,048 phone deletions, and 497 phone additions. Figure 3 (a) shows the top-20 most frequent phoneme substitution tags in the corpus. The most dominant substitution errors were “Z→S,” (voicing) “DH→D,” (fricative→stop) “IH→IY,” and “OW→AO” (use of a tense vowel for lax, and vice versa.) Each contains English phoneme distinctions that lead to common substitution errors for varied English learners. Figure 3 (b) shows the phone deletion errors in the annotations. In our sample group, the most frequent phoneme deletions were “D,” “T,” and “R,” almost always in non-initial position. Many non-native speakers of English do not pronounce the American English phoneme “R” in postvocalic position (e.g., in *car* and *farm*.) “T” and “D” often occur as word endings and in consonant clusters both within and across words, where they were often omitted. Figure 3 (c) shows the phone addition errors in the annotations. The ones that stood out were “AH,” “EH,” “R,” “AX (schwa),” “G,” and “IH.” The vowel additions simplify complex syllable structures with consonant clusters and so may serve to make the word more pronounceable. Table 2 provides a breakdown of pronunciation errors by L1s. Although others have used L1 to predict L2 pronunciation errors [33, 34, 43], such predictions are often inaccurate when applied to individual learners. Thus, this list is meant to start a discussion of the types of errors that actually occur in L2-ARCTIC.

Table 2: Most frequent errors by native language; the top-5 error occurrences are listed in descending order

L1	Substitutions	Deletions	Additions
Hindi	DH→D, Z→S, W→V EY→EH, TH→T	R, D, T ER, HH	R, AH, S, Y AA
Korean	DH→D, Z→S, IH→IY OW→AO, EH→AE	D, T, R HH, K	AX, IH, AH, S Y
Mandarin	Z→S, DH→D, IH→IY N→NG, V→F	D, T, R L, N	AH, AX, IH N, R
Spanish	Z→S, IH→IY, DH→D AE→AA, AH→AO	D, T, AH Z, IH	EH, AX, AH IH, IY
Arabic	P→B, OW→AO R→ERR, DH→Z, Z→S	T, R, D AH, IH	G, AH, IH AX, EH

¹ Some speakers did not read all sentences, and a few sentences were removed for some speakers since those recordings did not have the required quality.

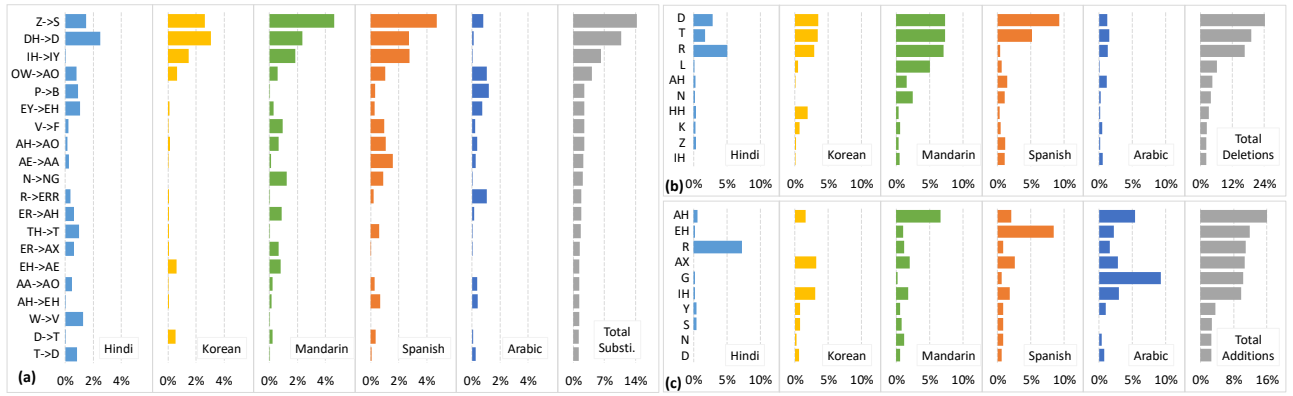


Figure 3: *L1-dependent pronunciation error distributions and the aggregated results. (a) Substitutions (b) Deletions (c) Additions. Errors with low frequencies were omitted; all the values are the percentages with respect to the total number of each error type (i.e., normalized universally); “ERR” means an erroneous pronunciation that is not in the ARPabet phoneme set.*

5. Mispronunciation detection evaluation

This section provides initial results on mispronunciation detection using the 10 speakers that we have currently released. Our implementation is based on the conventional Goodness of Pronunciation (GOP) method as defined in [44]. The acoustic model we used was a triphone model (tri6b) as defined by Kaldi’s Librispeech training script [45]. It is a GMM trained with 960 hours of native English speech [46], and contains 150,000 Gaussian mixtures. Three-state left-to-right HMMs were used for non-silent sounds. The Kaldi implementation does not have a fixed number of Gaussians for each HMM state. The Word Error Rate (WER) of this acoustic model was around 8% on clean speech when combined with a 3-gram language model.

We used the phone-independent thresholding variation of the GOP method to make the classification decisions, i.e., if the GOP score of a phone segment was higher than a threshold P , then it was accepted as a correct pronunciation, otherwise it was rejected as an error. As a preliminary result, we only focused on substitution errors since the GOP is not suited for detecting additions and deletions.

Two hundred and six (206) utterances were withheld to determine the search range of the phoneme-independent detection threshold. The remaining 1,293 utterances were used as the testing set. In the testing data, excluding the additions and deletion tags, there are 41,353 phone samples in total, where 4,415 (10.7%) were tagged as substitution errors. We set the log GOP threshold between -16 and 0 and made the step size 0.1. For each experiment condition, we computed the detection precision rate as N_{TN}/N and the recall rate as N_{TN}/N_{errors} , where N_{TN} is the number of correctly predicted substitution errors, N is the total number of segments predicted as substitution errors, and N_{errors} is the total number of substitution errors in the testing set. The Precision-Recall curve is shown in Figure 4. When we set the threshold to -4.2 (in log scale), the precision equals recall (0.29). From this result, we can see that the dataset is quite challenging, because it contains speech data from different L1 backgrounds and recorded by speakers with a wide range of pronunciation challenges. This GOP implementation is open source and is available online¹.

¹ <https://github.com/guanlongzhao/kaldi-gop>

6. Conclusion

This paper has presented L2-ARCTIC, a new non-native English speech corpus designed for voice conversion, accent conversion, and mispronunciation detection tasks. Each speaker in L2-ARCTIC produced sufficient speech data to capture their voice identity and accent characteristics. Detailed annotations on mispronunciation errors are also included. Thus, it is possible to use this corpus to develop and evaluate mispronunciation detection algorithms. To the best of our knowledge, L2-ARCTIC is the first of its own kind, and we believe it fills gaps in both voice/accent conversion and pronunciation training.

The corpus is released under the CC BY-NC 4.0 license [47] and is available at <https://psi.engr.tamu.edu/l2-arctic-corpus/>. Future work will be focusing on adding ten more speakers to the corpus.

7. Acknowledgments

This work was supported by NSF awards 1619212 and 1623750. We would like to thank the anonymous participants for recording the corpus. We also would like to thank Ziwei Zhou for his assistance with the annotations. We appreciate suggestions from Christopher Liberatorre and Shaojin Ding on early versions of this manuscript.

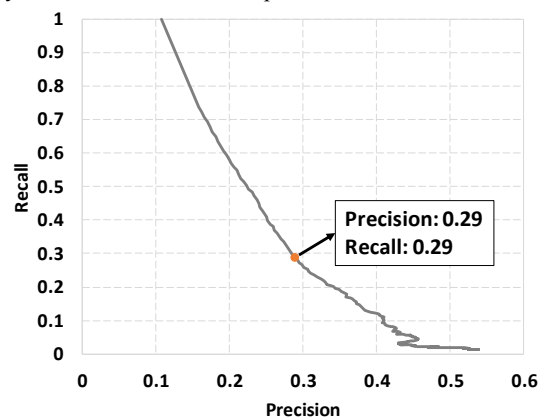


Figure 4: *Precision-Recall curve of a phone-independent GOP system to demo mispronunciation detection on L2-ARCTIC*

8. References

- [1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65-82, 2017.
- [2] S. Aryal and R. Gutierrez-Osuna, "Can Voice Conversion Be Used to Reduce Non-Native Accents?," in *ICASSP*, 2014, pp. 7879-7883.
- [3] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920-932, 2009.
- [4] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors—in search of the golden speaker," *Speech Communication*, vol. 37, no. 3, pp. 161-173, 2002.
- [5] 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 *Australian International Conference on Speech Science & Technology*, 2006, pp. 24-29.
- [6] J. Levis, "Computer technology in teaching and researching pronunciation," *Annual Review of Applied Linguistics*, vol. 27, pp. 184-202, 2007.
- [7] S. Weinberger. Speech accent archive [Online]. Available: <http://accent.gmu.edu>
- [8] P. Meier. IDEA: International Dialects of English Archive [Online]. Available: <http://www.dialectsarchive.com/>
- [9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222-2235, 2007.
- [10] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *ICASSP*, 2015, pp. 4869-4873.
- [11] G. Zhao and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *ICASSP*, 2017, pp. 5525-5529.
- [12] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally Linear Embedding for Exemplar-Based Spectral Conversion," in *Interspeech*, 2016, pp. 1652-1656.
- [13] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 223-224.
- [14] T. Toda *et al.*, "The Voice Conversion Challenge 2016," in *Interspeech*, 2016, pp. 1632-1636.
- [15] J. Jenkins. (2008). *English as a lingua franca*. Available: http://www.jacets.org/2008convention/JACET2008_keynote_jenkins.pdf
- [16] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles," *Language and speech*, vol. 53, no. 4, pp. 510-540, 2010.
- [17] T. Lander. CSLU: Foreign Accented English Release 1.2 LDC2007S08 [Online]. Available: <https://catalog.ldc.upenn.edu/Ldc2007s08>
- [18] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1600-1610, 2003.
- [19] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193-207, 2017.
- [20] H. Yang and N. Wei, "Construction and data analysis of a Chinese learner spoken English corpus," ed: Shanghai Foreign Language Education Press, 2005.
- [21] W. Menzel *et al.*, "The ISLE corpus of non-native spoken English," in *Proceedings of LREC 2000: Language Resources and Evaluation Conference*, vol. 2, 2000, pp. 957-964.
- [22] N. F. Chen, R. Tong, D. Wee, P. X. Lee, B. Ma, and H. Li, "SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin-Chinese," in *Interspeech*, 2016, pp. 1545-1549.
- [23] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154-166, 2015.
- [24] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 3, pp. 564-579, 2015.
- [25] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 787-797, 2015.
- [26] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," in *ICASSP*, 2018.
- [27] P. Pramod, "Indian English Pronunciation," in *The Handbook of English Pronunciation*, M. Reed and J. Levis, Eds.: Wiley Blackwell, 2015, pp. 301-319.
- [28] S.-A. Jun, "Prosody in sentence processing: Korean vs. English," *UCLA Working Papers in Phonetics*, vol. 104, pp. 26-45, 2005.
- [29] M. Ueyama and S.-A. Jun, "Focus realization of Japanese English and Korean English intonation," *UCLA Working Papers in Phonetics*, pp. 110-125, 1996.
- [30] J. Anderson-Hsieh, R. Johnson, and K. Koehler, "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure," *Language learning*, vol. 42, no. 4, pp. 529-555, 1992.
- [31] M. C. Pennington and N. C. Ellis, "Cantonese speakers' memory for English sentences with prosodic cues," *The Modern Language Journal*, vol. 84, no. 3, pp. 372-389, 2000.
- [32] J. Chang, "Chinese speakers," *Learner English*, vol. 2, pp. 310-324, 1987.
- [33] J. Morley, "Teaching American English Pronunciation," ed: JSTOR, 1993.
- [34] B. Smith, *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, 2001.
- [35] M. Benrabah, "Word-stress—a source of unintelligibility in English," *IRAL-International Review of Applied Linguistics in Language Teaching*, vol. 35, no. 3, pp. 157-166, 1997.
- [36] K. De Jong and B. A. Zawaydeh, "Stress, duration, and intonation in Arabic word-level prosody," *Journal of Phonetics*, vol. 27, no. 1, pp. 3-22, 1999.
- [37] Y. Cho and B. Bridgeman, "Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities," *Language Testing*, vol. 29, no. 3, pp. 421-442, 2012.
- [38] H. Zen *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *SSW*, 2007, pp. 294-299.
- [39] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556-566, 2013.
- [40] Audacity®. Available: <http://www.audacityteam.org/>
- [41] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: trainable text-speech alignment using Kaldi," in *Interspeech*, 2017, pp. 498-502.
- [42] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glott international*, vol. 5, 2002.
- [43] M. Munro, "How well can we predict L2 learners' pronunciation difficulties?," *CATESOL Journal*, vol. 30, no. 1, pp. 267-282, 2018.
- [44] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95-108, 2000.
- [45] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition & Understanding*, 2011.
- [46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206-5210.
- [47] *Creative Commons Attribution-NonCommercial 4.0 International Public License*. Available: <https://creativecommons.org/licenses/by-nc/4.0/legalcode>

Appendix ZHAO03



WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

SIGN UP | LOG IN | UPLOAD

Search

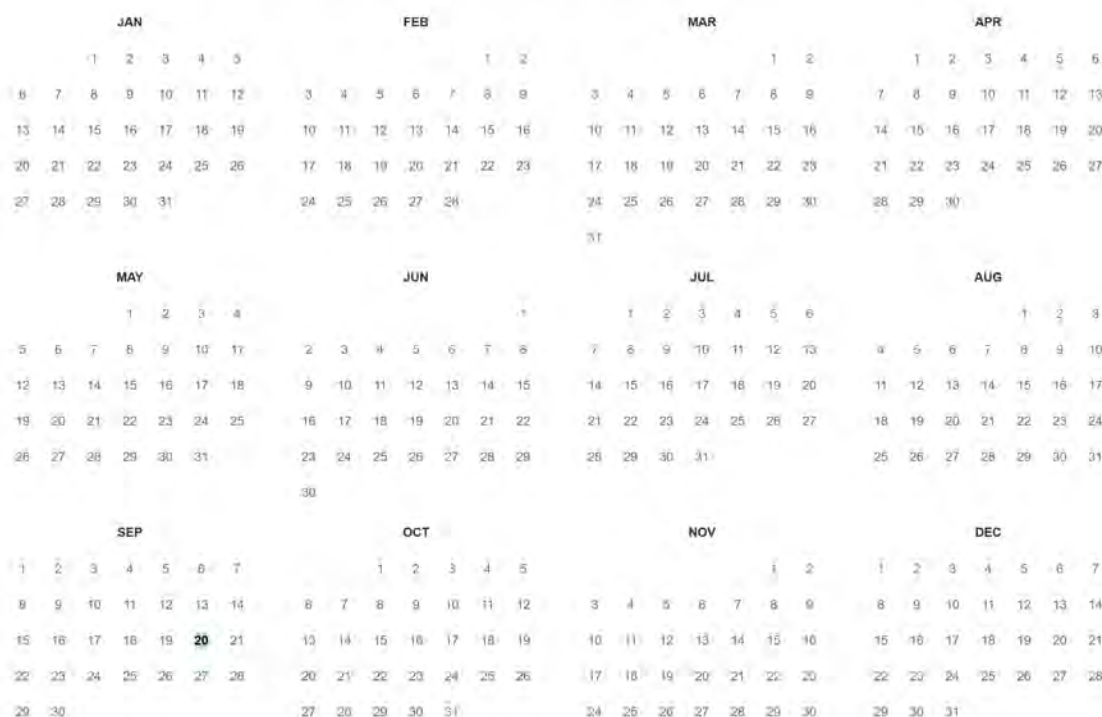
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER



Explore more than 1 trillion web pages saved over time

[Calendar](#) · [Collections](#) · [Changes](#) · [Summary](#) · [Site Map](#) · [URLs](#)

Saved 29 times between September 20, 2019 and May 26, 2025.



Note

This calendar view maps the number of times https://isca-speech.org/archive/Interspeech_2018/ 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 504(c)(3) nonprofit building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive.org.

Your use of the Wayback Machine is subject to the Internet Archive's [Terms of Use](#).

Interspeech 2018

2-6 September 2018, Hyderabad

Chair: B. Yegnanarayana

ISSN: 1998-9772

DOI: 10.21437/Interspeech.2018

ISCA Medal Talk

From Vocoders to Code-Excited Linear Prediction: Learning How We Hear What We Hear

Bishnu S. Atal 

End-to-End Speech Recognition

Semi-Supervised End-to-End Speech Recognition

Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Atsunori Ogawa, Marc Delcroix 

Improved Training of End-to-end Attention Models for Speech Recognition

Albert Zeyer, Kazuki Irie, Ralf Schlüter, Hermann Ney 

End-to-end Speech Recognition Using Lattice-free MMI

Hossein Hadian, Hossein Sameti, Daniel Povey, Sanjeev Khudanpur 

Multi-channel Attention for End-to-End Speech Recognition

Stefan Braun, Daniel Neil, Jithendar Anumula, Enea Ceolini, Shih-Chii Liu 

Quaternion Convolutional Neural Networks for End-to-End Automatic Speech Recognition

Titouan Parcollet, Ying Zhang, Mohamed Morchid, Chiheb Trabelsi, Georges Linares, Renato de Mori, Yoshua Bengio 

Compression of End-to-End Models

Ruoming Pang, Tara Sainath, Rohit Prabhavalkar, Suyog Gupta, Yonghui Wu, Shuyuan Zhang, Chung-Cheng Chiu 

Prosody Modeling and Generation

Learning Interpretable Control Dimensions for Speech Synthesis by Using External Data

Zack Hodari, Oliver Watts, Srikanth Ronanki, Simon King 

Investigating Accuracy of Pitch-accent Annotations in Neural Network-based Speech Synthesis and Denoising Effects

Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, Nobuyuki Nishizawa 

An Exploration of Local Speaking Rate Variations in Mandarin Read Speech

Guan-Ting Liou, Chen-Yu Chiang, Yih-Ru Wang, Sin-Hong Chen 

BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End

Yibin Zheng, Jianhua Tao, Zhengqi Wen, Ya Li 

Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion

Berrak Sisman, Haizhou Li 

Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model

Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang, Yonghe Wang 

Speaker Verification I

Improved Supervised Locality Preserving Projection for I-vector Based Speaker Verification

IPR2026-00272

Krisp Technologies EX1044 Page 87

Lanhua You, Wu Guo, Yan Song, Sheng Zhang 

Double Joint Bayesian Modeling of DNN Local i-Vector for Text Dependent Speaker Verification with Random Digit Strings

Ziqiang Shi, Huibin Lin, Liu Liu, Rujie Liu 

Fast Variational Bayes for Heavy-tailed PLDA Applied to i-vectors and x-vectors

Anna Silnova, Niko Brümmer, Daniel Garcia-Romero, David Snyder, Lukáš Burget 

Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion

Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi 

A Generalization of PLDA for Joint Modeling of Speaker Identity and Multiple Nuisance Conditions

Luciana Ferrer, Mitchell McLaren 

An Investigation of Non-linear i-vectors for Speaker Verification

Nanxin Chen, Jesús Villalba, Najim Dehak 

Spoken Term Detection

CNN Based Query by Example Spoken Term Detection

Dhananjay Ram, Lesly Miculicich, Hervé Bourlard 

Learning Acoustic Word Embeddings with Temporal Context for Query-by-Example Speech Search

Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, Bin Ma, Haizhou Li 

Siamese Recurrent Auto-Encoder Representation for Query-by-Example Spoken Term Detection

Ziwei Zhu, Zhiyong Wu, Runnan Li, Helen Meng, Lianhong Cai 

Fast Derivation of Cross-lingual Document Vectors from Self-attentive Neural Machine Translation Model

Wei Li, Brian Mak 

LSTM Based Attentive Fusion of Spectral and Prosodic Information for Keyword Spotting in Hindi Language

Laxmi Pandey, Karan Nathwani 

Spoken Keyword Detection Using Joint DTW-CNN

Ravi Shankar, Vikram C M, S R Mahadeva Prasanna 

The INTERSPEECH 2018 Computational Paralinguistics Challenge (ComParE): Atypical & Self-Assessed Affect, Crying & Heart Beats 1

The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats

Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B. Pokorny, Eva-Maria Rathner, Katrin D. Bartl-Pokorny, Christa Einspieler, Dajie Zhang, Alice Baird, Shahin Amiriparian, Kun Qian, Zhao Ren, Maximilian Schmitt, Panagiotis Tzirakis, Stefanos Zafeiriou 

An Ensemble of Transfer, Semi-supervised and Supervised Learning Methods for Pathological Heart Sound Classification

Ahmed Imtiaz Humayun, Md. Tauhiduzzaman Khan, Shabnam Ghaffarzadegan, Zhe Feng, Taufiq Hasan 

Monitoring Infant's Emotional Cry in Domestic Environments Using the Capsule Network Architecture

Mehmet Ali Tuğtekin Turan, Engin Erzin 

Neural Network Architecture That Combines Temporal and Summative Features for Infant Cry Classification in the Interspeech 2018 Computational Paralinguistics Challenge

Mark Huckvale 

Evolving Learning for Analysing Mood-Related Infant Vocalisation

Zixing Zhang, Jing Han, Kun Qian, Björn Schuller 

Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?

Johannes Wagner, Dominik Schiller, Andreas Seiderer, Elisabeth André 

Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition

Danqing Luo, Yuejian Zou, Dongyan Huang 

Using Voice Quality Supervectors for Affect Identification

Soo Jin Park, Amber Afshan, Zhi Ming Chua, Abeer Alwan 

An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals

Dengke Tang, Junlin Zeng, Ming Li 

Show and Tell 1

DialogOS: Simple and Extensible Dialogue Modeling

Alexander Koller, Timo Baumann, Arne Köhn 

IPR2026-00272

A Framework for Speech Recognition Benchmarking

Franck Dernoncourt, Trung Bui, Walter Chang 

Flexible Tongue Housed in a Static Model of the Vocal Tract With Jaws, Lips and Teeth

Takayuki Arai 

Voice Analysis Using Acoustic and Throat Microphones for Speech Therapy

Lani Mathew, Gopakumar K. 

A Robust Context-Dependent Speech-to-Speech Phraselator Toolkit for Alexa

Manny Rayner, Nikos Tsourakis, Jan Stanek 

Speech Segments and Voice Quality

Discriminating Nasals and Approximants in English Language Using Zero Time Windowing

RaviShankar Prasad, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, Bayya Yegnanarayana 

Gestural Lenition of Rhotics Captures Variation in Brazilian Portuguese

Phil Howson, Alexei Kochetov 

Identification and Classification of Fricatives in Speech Using Zero Time Windowing Method

RaviShankar Prasad, Bayya Yegnanarayana 

GlobalTIMIT: Acoustic-Phonetic Datasets for the World's Languages

Nattaniun Chanchaochai, Christopher Cieri, Japhet Debrah, Hongwei Ding, Yue Jiang, Sishi Liao, Mark Liberman, Jonathan Wright, Jiahong Yuan, Juhong Zhan, Yuqing Zhan 

Structural Effects on Properties of Consonantal Gestures in Tashlhiyt

Anne Hermes, Doris Mücke, Bastian Auris, Rachid Ridouane 

The Retroflex-dental Contrast in Punjabi Stops and Nasals: A Principal Component Analysis of Ultrasound Images

Alexei Kochetov, Matthew Faytak, Kiranpreet Nara 

Vowels and Diphthongs in Hangzhou Wu Chinese Dialect

Yang Yue, Fang Hu 

Resyllabification in Indian Languages and Its Implications in Text-to-speech Systems

Mahesh M, Jeena J Prakash, Hema Murthy 

Voice Source Contribution to Prominence Perception: Rd Implementation

Andy Murphy, Irena Yanushevskaya, Ailbhe Ni Chasaide, Christer Gobl 

On the Relationship between Glottal Pulse Shape and Its Spectrum: Correlations of Open Quotient, Pulse Skew and Peak Flow with Source Harmonic Amplitudes

Christer Gobl, Andy Murphy, Irena Yanushevskaya, Ailbhe Ni Chasaide 

The Individual and the System: Assessing the Stability of the Output of a Semi-automatic Forensic Voice Comparison System

Vincent Hughes, Philip Harrison, Paul Foulkes, Peter French, Colleen Kavanagh, Eugenia San Segundo Fernández 

Breathy to Tense Voice Discrimination using Zero-Time Windowing Cepstral Coefficients (ZTWCCs)

Sudarsana Reddy Kadiri, Bayya Yegnanarayana 

Analysis of Breathiness in Contextual Vowel of Voiceless Nasals in Mizo

Pamir Gogoi, Sishir Kalita, Parismita Gogoi, Ratree Wayland, Priyankoo Sarmah, S R Mahadeva Prasanna 

Speaker State and Trait

Infant Emotional Outbursts Detection in Infant-parent Spoken Interactions

Yijia Xu, Mark Hasegawa-Johnson, Nancy McElwain 

Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts

Jaemin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, Najim Dehak 

Preference-Learning with Qualitative Agreement for Sentence Level Emotional Annotations

Srinivas Parthasarathy, Carlos Busso 

Transfer Learning for Improving Speech Emotion Classification Accuracy

Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, Julien Epps 

What Do Classifiers Actually Learn? a Case Study on Emotion Recognition Datasets

Patrick Meyer, Eric Buschermöhle, Tim Fingscheldt 

State of Mind: Classification through Self-reported Affect and Word Use in Speech.

Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, Harald Baumeister 

Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition

Ziping Zhao, Yu Zheng, Zixing Zhang, Haishuai Wang, Yiqin Zhao, Chao Li 

End-to-end Deep Neural Network Age Estimation

Pegah Ghahremani, Phani Sankar Nidadavolu, Nanxin Chen, Jesús Villalba, Daniel Povey, Sanjeev Khudanpur, Najim Dehak 

Improving Gender Identification in Movie Audio Using Cross-Domain Data

Rajat Hebbar, Krishna Somandepalli, Shrikanth Narayanan 

On Learning to Identify Genders from Raw Speech Signal Using CNNs

Selen Hande Kabil, Hannah Muckenhirn, Mathew Magimai.-Doss 

Denoising and Raw-waveform Networks for Weakly-Supervised Gender Identification on Noisy Speech

Jilt Sebastian, Manoj Kumar, Pavan Kumar D. S., Mathew Magimai.-Doss, Hema Murthy, Shrikanth Narayanan 

The Effect of Exposure to High Altitude and Heat on Speech Articulatory Coordination

James Williamson, Thomas Quatieri, Adam Lammert, Katherine Mitchell, Katherine Finkelstein, Nicole Ekon, Caitlin Dillon, Robert Kenefick, Kristin Heaton 

Deep Learning for Source Separation and Pitch Tracking

Permutation Invariant Training of Generative Adversarial Network for Monaural Speech Separation

Lianwu Chen, Meng Yu, Yanmin Qian, Dan Su, Dong Yu 

Deep Extractor Network for Target Speaker Recovery from Single Channel Speech Mixtures

Jun Wang, Jie Chen, Dan Su, Lianwu Chen, Meng Yu, Yanmin Qian, Dong Yu 

Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network

Weipeng He, Petr Motlicek, Jean-Marc Odobez 

Detection of Glottal Closure Instants from Speech Signals: A Convolutional Neural Network Based Method

Shuai Yang, Zhiyong Wu, Binbin Shen, Helen Meng 

Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks

Zhong-Qiu Wang, Xuellang Zhang, DeLiang Wang 

Waveform to Single Sinusoid Regression to Estimate the F0 Contour from Noisy Speech Using Recurrent Deep Neural Networks

Akihiro Kato, Tomi Kinnunen 

Reducing Interference with Phase Recovery in DNN-based Monaural Singing Voice Separation

Paul Magron, Konstantinos Drossos, Stylianos Ioannis Mimilakis, Tuomas Virtanen 

Nebula: F0 Estimation and Voicing Detection by Modeling the Statistical Properties of Feature Extractors

Kanru Hua 

Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network

Yi Luo, Nima Mesgarani 

Music Source Activity Detection and Separation Using Deep Attractor Network

Rajath Kumar, Yi Luo, Nima Mesgarani 

Improving Mandarin Tone Recognition Using Convolutional Bidirectional Long Short-Term Memory with Attention

Longfei Yang, Yanlu Xie, Jinsong Zhang 

Acoustic Analysis-Synthesis of Speech Disorders

Vowel Space as a Tool to Evaluate Articulation Problems

Rob van Son, Catherine Middag, Kris Demuyne 

Towards a Better Characterization of Parkinsonian Speech: A Multidimensional Acoustic Study

Veronique Delvaux, Kathy Huet, Myriam Piccaluga, Sophie van Malderen, Bernard Harmegnies 

Self-similarity Matrix Based Intelligibility Assessment of Cleft Lip and Palate Speech

Sishir Kalita, S R Mahadeva Prasanna, Samarendra Dandapat 

Pitch-Adaptive Front-end Feature for Hypernasality Detection

Akhilesh Kumar Dubey, S R Mahadeva Prasanna, S Dandapat 

Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis

Raquel Norel, Mary Pietrowicz, Carla Agurto, Shay Rishoni, Guillermo Cecchi 

Detection of Glottal Activity Errors in Production of Stop Consonants in Children with Cleft Lip and Palate

Vikram C M, S R Mahadeva Prasanna, Ajish K Abraham, Pushpavathi M, Girish K S 

ASR Systems and Technologies

Cold Fusion: Training Seq2Seq Models Together with Language Models

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, Adam Coates 

Investigation on Estimation of Sentence Probability by Combining Forward, Backward and Bi-directional LSTM-RNNs

Kazuki Irie, Zhihong Lei, Lihui Deng, Ralf Schlüter, Hermann Ney 

Subword and Crossword Units for CTC Acoustic Models

Thomas Zenkel, Ramon Sanabria, Florian Metze, Alex Waibel 

Neural Error Corrective Language Models for Automatic Speech Recognition

Tomohiro Tanaka, Ryo Masumura, Hirokazu Masataki, Yushi Aono 

Entity-Aware Language Model as an Unsupervised Reranker

Mohammad Sadegh Rasooli, Sarangarajan Parthasarathy 

Character-level Language Modeling with Gated Hierarchical Recurrent Neural Networks

Iksoo Choi, Jinhwan Park, Wonyong Sung 

Deception, Personality, and Culture Attribute

Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues

Sarah Ita Levitan, Angel Maredia, Julia Hirschberg 

Deep Personality Recognition for Deception Detection

Guozhen An, Sarah Ita Levitan, Julia Hirschberg, Rivka Levitan 

Cross-cultural (A)symmetries in Audio-visual Attitude Perception

Hansjörg Mixdorff, Albert Rilliard, Tan Lee, Matthew K. H. Ma, Angelika Hönemann 

An Active Feature Transformation Method for Attitude Recognition of Video Bloggers

Fasih Haider, Fahim A. Salim, Owen Conlan, Saturnino Luz 

Automatic Assessment of Individual Culture Attribute of Power Distance Using a Social Context-Enhanced Prosodic Network Representation

Fu-Sheng Tsai, Hao-Chun Yang, Wei-Wen Chang, Chi-Chun Lee 

Analysis and Detection of Phonation Modes in Singing Voice using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC)

Sudarsana Reddy Kadiri, Bayya Yegnanarayana 

Automatic Detection and Recognition of Voice and Speech Disorders

A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks

Huiyi Wu, John Soraghan, Anja Lowit, Gaetano Di-Caterina 

Dysarthric Speech Recognition Using Time-delay Neural Network Based Denoising Autoencoder

Chitralekha Bhat, Biswajit Das, Bhavik Vachhani, Sunil Kumar Kopparapu 

A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease

Juan Camilo Vásquez Correa, Tomas Arias, Juan Rafael Orozco-Arroyave, Elmar Nöth 

The Use of Machine Learning and Phonetic Endophenotypes to Discover Genetic Variants Associated with Speech Sound Disorder

Jason Lilley, Erin Crowgey, H Timothy Bunnell 

Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems

Meredith Moore, Hemanth Venkateswara, Sethuraman Panchanathan 

Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition

Bhavik Vachhani, Chitralekha Bhat, Sunil Kumar Kopparapu 

Voice Conversion

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

Shaojin Ding, Guanlong Zhao, Christopher Liberator, Ricardo Gutierrez-Osuna 

Learning Structured Dictionaries for Exemplar-based Voice Conversion

Shaojin Ding, Christopher Liberator, Ricardo Gutierrez-Osuna 

Exemplar-Based Spectral Detail Compensation for Voice Conversion

Yu-Huai Peng, Hsin-Te Hwang, Yichiao Wu, Yu Tsao, Hsin-Min Wang 

Whispered Speech to Neutral Speech Conversion Using Bidirectional LSTMs

G. Nisha Meenakshi, Prasanta Kumar Ghosh 

Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance

Songxiang Liu, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, Helen Meng 

Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations

Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, Lin-shan Lee 

The INTERSPEECH 2018 Computational Paralinguistics Challenge (ComParE): Atypical & Self-Assessed Affect, Crying & Heart Beats 2

Attention-based Sequence Classification for Affect Detection

Cristina Gorrostieta, Richard Brutti, Kye Taylor, Avi Shapiro, Joseph Moran, Ali Azarbayejani, John Kane 

Computational Paralinguistics: Automatic Assessment of Emotions, Mood and Behavioural State from Acoustics of Speech

Zafi Sherhan Syed, Julien Schroeter, Kirill Sidorov, David Marshall 

Investigating Utterance Level Representations for Detecting Intent from Acoustics

SaiKrishna Rallabandi, Bhavya Karki, Carla Viegas, Eric Nyberg, Alan W Black 

LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition

Heysem Kaya, Dmitrii Fedotov, Ali Yeşilkanat, Oxana Verkholiyak, Yang Zhang, Alexey Karpov 

Implementing Fusion Techniques for the Classification of Paralinguistic Information

Bogdan Vlasenko, Jilt Sebastian, Pavan Kumar D. S., Mathew Magimai.-Doss 

General Utterance-Level Feature Extraction for Classifying Crying Sounds, Atypical & Self-Assessed Affect and Heart Beats

Gábor Gosztolya, Tamás Grósz, László Tóth 

Self-Assessed Affect Recognition Using Fusion of Attentional BLSTM and Static Acoustic Features

Bo-Hao Su, Sung-Lin Yeh, Ming-Ya Ko, Huan-Yu Chen, Shun-Chang Zhong, Jeng-Lin Li, Chi-Chun Lee 

Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge

Claude Montacié, Marie-José Caraty 

Show and Tell 2

Intonation tutor by SPIRE (In-SPIRE): An Online Tool for an Automatic Feedback to the Second Language Learners in Learning Intonation

Anand P A, Chiranjeevi Yarra, Kausthubha N K, Prasanta Kumar Ghosh 

Game-based Spoken Dialog Language Learning Applications for Young Students

Keelan Evanini, Veronika Timpe-Laughlin, Eugene Tsuprun, Ian Blood, Jeremy Lee, James Bruno, Vikram Ramanarayanan, Patrick Lange, David Suendermann-Oeft 

The IBM Virtual Voice Creator

Alexander Sorin, Slava Shechtman, Zvi Kons, Ron Hoory, Shay Ben-David, Joe Pavitt, Shai Rozenberg, Carmel Rabinovitz, Tal Drory 

Mobile Application for Learning Languages for the Unlettered

Gayathri G, Mohana N, Radhika Pal, Hema Murthy 

Mandarin-English Code-switching Speech Recognition

Haihua Xu, Van Tung Pham, Zin Tun Kyaw, Zhi Hao Lim, Eng Siong Chng, Haizhou Li 

Spoken Dialogue Systems and Conversational Analysis

Joint Learning of Domain Classification and Out-of-Domain Detection with Dynamic Class Weighting for Satisficing False Acceptance Rates

Joo-Kyung Kim, Young-Bum Kim 

Analyzing Vocal Tract Movements During Speech Accommodation

Sankar Mukherjee, Thierry Legou, Leonardo Lancia, Pauline Hilt, Alice Tomassini, Luciano Fadiga, Alessandro D'Ausilio, Leonardo Badino, Noël Nguyen 

Cross-Lingual Multi-Task Neural Architecture for Spoken Language Understanding

Yujiang Li, Xuemin Zhao, Weiqun Xu, Yonghong Yan 

Statistical Model Compression for Small-Footprint Natural Language Understanding

Grant P. Strimel, Kanthashree Mysore Sathyendra, Stanislav Peshterliev 

Comparison of an End-to-end Trainable Dialogue System with a Modular Statistical Dialogue System

Norbert Braunschweiler, Alexandros Papangelis 

A Discriminative Acoustic-Prosodic Approach for Measuring Local Entrainment

Megan Willi, Stephanie A. Borrie, Tyson S. Barrett, Ming Tu, Visar Berisha 

Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs

Matthew Roddy, Gabriel Skantze, Naomi Harte 

Classification of Correction Turns in Multilingual Dialogue Corpus

Ivan Kraljevski, Diane Hirschfeld 

Contextual Slot Carryover for Disparate Schemas

Chetan Naik, Arpit Gupta, Hancheng Ge, Mathias Lambert, Ruhi Sarikaya 

Capsule Networks for Low Resource Spoken Language Understanding

Vincent Renkens, Hugo van Hamme 

Intent Discovery Through Unsupervised Semantic Text Clustering

Padmasundari, Srinivas Bangalore 

Multimodal Polynomial Fusion for Detecting Driver Distraction

Yulun Du, Alan W Black, Louis-Philippe Morency, Maxine Eskenazi 

Engagement Recognition in Spoken Dialogue via Neural Network by Aggregating Different Annotators' Models

Koji Inoue, Divesh Lala, Katsuya Takanashi, Tatsuya Kawahara 

A First Investigation of the Timing of Turn-taking in Ruuli

Tuarik Buanzur, Margaret Zellers, Saudah Namyalo, Alena Witzlack-Makarevich 

Spoofing Detection

Spoofing Detection Using Adaptive Weighting Framework and Clustering Analysis

Yuanjun Zhao, Roberto Togneri, Victor Sreeram 

Exploration of Compressed ILPR Features for Replay Attack Detection

Sarfraz Jellil, Sishir Kallita, S R Mahadeva Prasanna, Rohit Sinha 

Detection of Replay-Spoofing Attacks Using Frequency Modulation Features

Tharshini Gunendradasan, Buddhi Wickramasinghe, Ngoc Phu Le, Eliathamby Ambikairajah, Julien Epps 

Effectiveness of Speech Demodulation-Based Features for Replay Detection

Madhu Kamble, Hemlata Tak, Hemant Patil 

Novel Variable Length Energy Separation Algorithm Using Instantaneous Amplitude Features for Replay Detection

Madhu Kamble, Hemant Patil 

Feature with Complementarity of Statistics and Principal Information for Spoofing Detection

Jichen Yang, Changhui You, Qianhua He 

Multiple Phase Information Combination for Replay Attacks Detection

Dongbo Li, Longbiao Wang, Jianwu Dang, Meng Liu, Zeyan Oo, Seichi Nakagawa, Haotian Guan, Xiangang Li 

Frequency Domain Linear Prediction Features for Replay Spoofing Attack Detection

Buddhi Wickramasinghe, Saad Irtza, Eliathamby Ambikairajah, Julien Epps 

Auditory Filterbank Learning for Temporal Modulation Features in Replay Spoof Speech Detection

Hardik Sailor, Madhu Kamble, Hemant Patil 

Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric

Kaavya Srisankararaja, Vidhyasaharan Sethu, Eliathamby Ambikairajah 

A Deep Identity Representation for Noise Robust Spoofing Detection

Alejandro Gómez Alanís, Antonio M. Peinado, Jose A. Gonzalez, Angel Gomez 

End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention

Francis Tom, Mohit Jain, Prasenjit Dey 

Decision-level Feature Switching as a Paradigm for Replay Attack Detection

Saranya M S, Hema Murthy 

Modulation Dynamic Features for the Detection of Replay Attacks

Gajan Suthokumar, Vidhyasaharan Sethu, Chamith Wijenayake, Eliathamby Ambikairajah 

Speech Analysis and Representation

On the Usefulness of the Speech Phase Spectrum for Pitch Extraction

Erfan Loweimi, Jon Barker, Thomas Hain 

Time-regularized Linear Prediction for Noise-robust Extraction of the Spectral Envelope of Speech

Manu Airaksinen, Lauri Juvela, Okko Räsänen, Paavo Alku 

Auditory Filterbank Learning Using ConvRBM for Infant Cry Classification

Hardik B. Sailor, Hemant Patil 

Effectiveness of Dynamic Features in INCA and Temporal Context-INCA

Nirmesh Shah, Hemant Patil 

Singing Voice Phoneme Segmentation by Hierarchically Inferring Syllable and Phoneme Onset Positions

Rong Gong, Xavier Serra 

Novel Empirical Mode Decomposition Cepstral Features for Replay Spoof Detection

Prasad Tapkir, Hemant Patil 

Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection

Hemlata Tak, Hemant Patil 

Analysis of sparse representation based feature on speech mode classification

Kumud Tripathi, K. Sreenivasa Rao 

Multicomponent 2-D AM-FM Modeling of Speech Spectrograms

Jitendra Kumar Dhiman, Neeraj Sharma, Chandra Sekhar Seelamantula 

An Optimization Framework for Recovery of Speech from Phase-Encoded Spectrograms

Abhilash Sainathan, Sunil Rudresh, Chandra Sekhar Seelamantula 

Speaker Recognition with Nonlinear Distortion: Clipping Analysis and Impact

Wei Xia, John H.L. Hansen 

Linear Prediction Residual based Short-term Cepstral Features for Replay Attacks Detection

Madhusudan Singh, Debadatta Pati 

Analysis of Variational Mode Functions for Robust Detection of Vowels

Surbhi Sakshi, Avinash Kumar, Gayadhar Pradhan 

Sequence Models for ASR

Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition

Chao Weng, Jia Cui, Guangsen Wang, Jun Wang, Chengzhu Yu, Dan Su, Dong Yu 

Segmental Encoder-Decoder Models for Large Vocabulary Automatic Speech Recognition

Eugen Beck, Mirko Hannemann, Patrick Dötsch, Ralf Schlüter, Hermann Ney 

Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning

ShiLiang Zhang, Ming Lei 

End-to-End Speech Command Recognition with Capsule Network

Jaesung Bae, Dae-Shik Kim 

End-to-End Speech Recognition from the Raw Waveform

Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, Emmanuel Dupoux 

A Multistage Training Framework for Acoustic-to-Word Model

Chengzhu Yu, Chunlei Zhang, Chao Weng, Jia Cui, Dong Yu 

Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese

Shiyu Zhou, Linhao Dong, Shuang Xu, Bo Xu 

Densely Connected Networks for Conversational Speech Recognition

Kyu Han, Akshay Chandrashekar, Jungsuk Kim, Ian Lane 

Multi-Head Decoder for End-to-End Speech Recognition

Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda 

Compressing End-to-end ASR Networks by Tensor-Train Decomposition

Takuma Mori, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura 

Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech

Yu-An Chung, James Glass 

Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin

Linhao Dong, Shiyu Zhou, Wei Chen, Bo Xu 

Source Separation and Spatial Analysis

Joint Noise and Reverberation Adaptive Learning for Robust Speaker DOA Estimation with an Acoustic Vector Sensor

Disong Wang, Yuejian Zou 

Multiple Concurrent Sound Source Tracking Based on Observation-Guided Adaptive Particle Filter

Hong Liu, Haipeng Lan, Bing Yang, Cheng Pang 

Harmonic-Percussive Source Separation of Polyphonic Music by Suppressing Impulsive Noise Events

Gurunath Reddy M, K. Sreenivasa Rao, Partha Pratim Das 

Speaker Activity Detection and Minimum Variance Beamforming for Source Separation

Enea Ceolini, Jithendar Anumula, Adrian Huber, Ilya Kiselev, Shih-Chii Liu 

Sparsity-Constrained Weight Mapping for Head-Related Transfer Functions Individualization from Anthropometric Features

Xiaoke Qi, Jianhua Tao 

IPR2026-00272

Speech Source Separation Using ICA in Constant Q Transform Domain

Dheeraj Sai D.V.L.N, Kishor K.S, Sri Rama Murty Kodukula 

Multi-talker Speech Separation Based on Permutation Invariant Training and Beamforming

Lu Yin, Ziteng Wang, Risheng Xia, Junfeng Li, Yonghong Yan 

Expectation-Maximization Algorithms for Itakura-Saito Nonnegative Matrix Factorization

Paul Magron, Tuomas Virtanen 

Subband Weighting for Binaural Speech Source Localization

Karthik Girija Ramesan, Parth Suresh, Prasanta Kumar Ghosh 

Plenary Talk-1

Universal Tendencies for Cross-Linguistic Prosodic Tendencies: A Review and Some New Proposals

Jacqueline Vaisière 

Acoustic Model Adaptation

Learning to Adapt: A Meta-learning Approach for Speaker Adaptation

Ondřej Klejch, Joachim Fainberg, Peter Bell 

Speaker Adaptation and Adaptive Training for Jointly Optimised Tandem Systems

Yu Wang, Chao Zhang, Mark Gales, Philip Woodland 

Comparison of BLSTM-Layer-Specific Affine Transformations for Speaker Adaptation

Markus Kitzka, Ralf Schlüter, Hermann Ney 

Correlational Networks for Speaker Normalization in Automatic Speech Recognition

Rini A Sharon, Sandeep Reddy Kothinti, Umesh Srinivasan 

Machine Speech Chain with One-shot Speaker Adaptation

Andros Tjandra, Sakriani Sakti, Satoshi Nakamura 

Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition

Khe Chai Sim, Arun Narayanan, Ananya Misra, Anshuman Tripathi, Golan Pundak, Tara Sainath, Parisa Haghani, Bo Li, Michiel Bacchiani 

Statistical Parametric Speech Synthesis

Waveform-Based Speaker Representations for Speech Synthesis

Moquan Wan, Gilles Degottex, Mark J.F. Gales 

Incremental TTS for Japanese Language

Tomoya Yanagita, Sakriani Sakti, Satoshi Nakamura 

Transfer Learning Based Progressive Neural Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis

Ruibao Fu, Jianhua Tao, Yibin Zheng, Zhengqi Wen 

A Unified Framework for the Generation of Glottal Signals in Deep Learning-based Parametric Speech Synthesis Systems

Min-jae Hwang, Eunwoo Song, Jin-Seob Kim, Hong-Goo Kang 

Acoustic Modeling Using Adversarially Trained Variational Recurrent Neural Network for Speech Synthesis

Joun Yeop Lee, Sung Jun Cheon, Byoung Jin Choi, Nam Soo Kim, Eunwoo Song 

On the Application and Compression of Deep Time Delay Neural Network for Embedded Statistical Parametric Speech Synthesis

Yibin Zheng, Jianhua Tao, Zhengqi Wen, Ruibo Fu 

Emotion Modeling

Integrating Recurrence Dynamics for Speech Emotion Recognition

Efthymios Tzinis, Georgios Paraskevopoulos, Christos Baziotis, Alexandros Potamianos 

Towards Temporal Modelling of Categorical Speech Emotion Recognition

Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, Björn Schuller 

Emotion Recognition from Human Speech Using Temporal Information and Deep Learning

John Kim, Rif A. Saurous 

Role of Regularization in the Prediction of Valence from Speech

Kusha Sridhar, Srinivas Parthasarathy, Carlos Busso 

Learning Spontaneity to Improve Emotion Recognition in Speech

Karttikeya Mangalam, Tanaya Guha 

Predicting Categorical Emotions by Jointly Learning Primary and Secondary Emotions through Multitask Learning

Reza Lotfian, Carlos Busso 

Models of Speech Perception

Picture Naming or Word Reading: Does the Modality Affect Speech Motor Adaptation and Its Transfer?

Tiphaine Caudrelle, Pascal Perrier, Jean-Luc Schwartz, Amélie Rochet-Capellan 

Measuring the Band Importance Function for Mandarin Chinese with a Bayesian Adaptive Procedure

Yufan Du, Yi Shen, Hongying Yang, Xihong Wu, Jing Chen 

Wide Learning for Auditory Comprehension

Elnaz Shafaei-Bajestan, R. Harald Baayen 

Analyzing Reaction Time Sequences from Human Participants in Auditory Experiments

Louis ten Bosch, Mirjam Ernestus, Lou Boves 

Prediction of Perceived Speech Quality Using Deep Machine Listening

Jasper Ooster, Rainer Huber, Bernd T. Meyer 

Prediction of Subjective Listening Effort from Acoustic Data with Non-Intrusive Deep Models

Paul Kranzusch, Rainer Huber, Melanie Krüger, Birger Kollmeier, Bernd T. Meyer 

Multimodal Dialogue Systems

A Case Study on the Importance of Belief State Representation for Dialogue Policy Management

Margarita Kotti, Vassilios Diakouloukas, Alexandros Papangelis, Michail Lagoudakis, Yannis Stylianou 

Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers

Kohei Hara, Koji Inoue, Katsuya Takanashi, Tatsuya Kawahara 

Conversational Analysis Using Utterance-level Attention-based Bidirectional Recurrent Neural Networks

Chandrakant Bothe, Sven Magg, Cornelius Weber, Stefan Wermter 

A Comparative Study of Statistical Conversion of Face to Voice Based on Their Subjective Impressions

Yasuhito Ohsugi, Daisuke Saito, Nobuaki Minematsu 

Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching

Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, Huai-Hung Huang 

Coherence Models for Dialogue

Alessandra Cervone, Evgeny Stepanov, Giuseppe Riccardi 

Speech Recognition for Indian Languages

Indian Languages ASR: A Multilingual Phone Recognition Framework with IPA Based Common Phone-set, Predicted Articulatory Features and Feature fusion

Manjunath K E, K. Sreenivasa Rao, Dinesh Babu Jayagopi, V Ramasubramanian 

Rapid Collection of Spontaneous Speech Corpora Using Telephonic Community Forums

Agha Ali Raza, Awais Athar, Shan Randhawa, Zain Tariq, Muhammad Bilal Saleem, Haris Bin Zia, Umar Saif, Roni Rosenfeld 

Effect of TTS Generated Audio on OOV Detection and Word Error Rate in ASR for Low-resource Languages

Savitha Murthy, Dinkar Sitaram, Sunayana Sitaram 

Development of Large Vocabulary Speech Recognition System with Keyword Search for Manipuri

Tanvina Patel, Krishna D N, Noor Fathima, Nisar Shah, Mahima C, Deepak Kumar, Anuroop Iyengar 

Robust Mizo Continuous Speech Recognition

Abhishek Dey, Biswajit Dev Sarma, Wendy Lalminghlui, Lalnunsiamei Ngente, Parismita Gogoi, Priyankoo Sarmah, S R Mahadeva Prasanna, Rohit Sinha, Nirmala S.R. 

Semi-supervised and Active-learning Scenarios: Efficient Acoustic Model Refinement for a Low Resource Indian Language

Maharajan Chellapriyadarshini, Anoop Toffy, Srinivasa Raghavan K. M., V Ramasubramanian 

Automatic Speech Recognition with Articulatory Information and a Unified Dictionary for Hindi, Marathi, Bengali and Oriya

Debadatta Dash, Myungjong Kim, Kristin Teplansky, Jun Wang 

Show and Tell 3

IPR2026-00272

Krisp Technologies EX1044 Page 96

Captaina: Integrated Pronunciation Practice and Data Collection Portal

Aku Rouhe, Reima Karhila, Aija Elg, Minnaleena Toivola, Peter Smit, Anna-Riikka Smolander, Mikko Kurimo 

auMina™ - Enterprise Speech Analytics

Umesh Sachdev, Rajagopal Jayaraman, Zainab Millwala 

HoloCompanion: An MR Friend for EveryOne

Annam Naresh, Rushabh Gandhi, Mallikarjuna Rao Bellamkonda, Mithun Das Gupta 

akeira™ - Virtual Assistant

Umesh Sachdev, Rajagopal Jayaraman, Zainab Millwala 

Brain-Computer Interface using Electroencephalogram Signatures of Eye Blinks

Srihari Maruthachalam, Sidharth Aggarwal, Mari Ganesh Kumar, Mriganka Sur, Hema Murthy 

Speaker Verification II

Voice Comparison and Rhythm: Behavioral Differences between Target and Non-target Comparisons

Moez Ajili, Jean-François Bonastre, Solange Rossato 

Co-whitening of I-vectors for Short and Long Duration Speaker Verification

Longting Xu, Kong Aik Lee, Haizhou Li, Zhen Yang 

Compensation for Domain Mismatch in Text-independent Speaker Recognition

Fahimeh Bahmaninezhad, John H.L. Hansen 

Joint Learning of J-Vector Extractor and Joint Bayesian Model for Text Dependent Speaker Verification

Ziqiang Shi, Liu Liu, Huibin Lin, Rujie Liu 

Latent Factor Analysis of Deep Bottleneck Features for Speaker Verification with Random Digit Strings

Ziqiang Shi, Huibin Lin, Liu Liu, Rujie Liu 

VoxCeleb2: Deep Speaker Recognition

Joon Son Chung, Arsha Nagrani, Andrew Senior 

Supervised I-vector Modeling - Theory and Applications

Shreyas Ramoji, Sriram Ganapathy 

LOCUST - Longitudinal Corpus and Toolset for Speaker Verification

Evgeny Dmitriev, Yulia Kim, Anastasia Matveeva, Claude Montacié, Yannick Boulard, Yadviga Sinyavskaya, Yulia Zhukova, Adam Zarazinski, Egor Akhanov, Ilya Viksnin, Andrei Shlykov, Maria Usova 

Analysis of Language Dependent Front-End for Speaker Recognition

Srikanth Madikeri, Subhadeep Dey, Petr Motlicek 

Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings

Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen Stauffer, Colleen Richey, Aaron Lawson, Martin Graciarina 

Investigation on Bandwidth Extension for Speaker Recognition

Phani Sankar Nidadavolu, Cheng-I Lai, Jesús Villalba, Najim Dehak 

On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs

Hannah Muckenhirn, Mathew Magimai.-Doss, Sebastien Marcel 

On Convolutional LSTM Modeling for Joint Wake-Word Detection and Text Dependent Speaker Verification

Rajath Kumar, Vaishnavi Yeruva, Sriram Ganapathy 

Cosine Metric Learning for Speaker Verification in the I-vector Space

Zhongxin Bai, Xiao-Lei Zhang, Jingdong Chen 

An Unsupervised Neural Prediction Framework for Learning Speaker Embeddings Using Recurrent Neural Networks

Arindam Jati, Panayiotis Georgiou 

Novel Approaches to Enhancement

A New Framework for Supervised Speech Enhancement in the Time Domain

Ashutosh Pandey, DeLiang Wang 

Speech Enhancement Using the Minimum-probability-of-error Criterion

Jishnu Sadasivan, Subhadip Mukherjee, Chandra Sekhar Seelamantula 

Exploring the Relationship between Conic Affinity of NMF Dictionaries and Speech Enhancement Metrics

Pavlos Papadopoulos, Colin Vaz, Shrikanth Narayanan 

Using Shifted Real Spectrum Mask as Training Target for Supervised Speech Separation

Yun Liu, Hui Zhang, Xueliang Zhang 

Enhancement of Noisy Speech Signal by Non-Local Means Estimation of Variational Mode Functions

Nagapuri Srinivas, Gayadhar Pradhan, Syed Shahnawazuddin 

Phase-locked Loop (PLL) Based Phase Estimation in Single Channel Speech Enhancement

Priya Pallavi, Ch V Rama Rao 

Cycle-Consistent Speech Enhancement

Zhong Meng, Jinyu Li, Yifan Gong, Biing-Hwang (Fred) Juang 

Visual Speech Enhancement

Aviv Gabbay, Asaph Shamir, Shmuel Peleg 

Implementation of Digital Hearing Aid as a Smartphone Application

Saketh Sharma, Nitya Tiwari, Prem C. Pandey 

Bone-Conduction Sensor Assisted Noise Estimation for Improved Speech Enhancement

Ching-Hua Lee, Bhaskar D. Rao, Harinath Garudadri 

Artificial Bandwidth Extension with Memory Inclusion Using Semi-supervised Stacked Auto-encoders

Pramod Bachhav, Massimiliano Todisco, Nicholas Evans 

Large Vocabulary Concatenative Resynthesis

Soumi Maiti, Joey Ching, Michael Mandel 

Concatenative Resynthesis with Improved Training Signals for Speech Enhancement

Ali Raza Syed, Viet Anh Trinh, Michael Mandel 

Syllabification, Rhythm, and Voice Activity Detection

Comparison of Syllabification Algorithms and Training Strategies for Robust Word Count Estimation across Different Languages and Recording Conditions

Okko Räsänen, Seshadri Shreyas, Marisa Casillas 

A Comparison of Input Types to a Deep Neural Network-based Forced Aligner

Matthew C. Kelley, Benjamin V. Tucker 

Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection

Youngmoon Jung, Younggwon Kim, Yeunju Choi, Hoirin Kim 

Information Bottleneck Based Percussion Instrument Diarization System for Taniavartanam Segments of Carnatic Music Concerts

Nauman Dawalatabad, Jom Kuriakose, Chandra Sekhar Chellu, Hema Murthy 

Robust Voice Activity Detection Using Frequency Domain Long-Term Differential Entropy

Debayan Ghosh, Muralishankar R, Sanjeev Gurugopinath 

Device-directed Utterance Detection

Sri Harish Mallidi, Roland Maas, Kyle Goehner, Ariya Rastrow, Spyros Matsoukas, Björn Hoffmeister 

Acoustic-Prosodic Features of Tabla Bol Recitation and Correspondence with the Tabla Imitation

Rohit M A, Preeti Rao 

Who Said That? a Comparative Study of Non-negative Matrix Factorization Techniques

Teun Krikke, Frank Broz, David Lane 

AVA-Speech: A Densely Labeled Dataset of Speech Activity in Movies

Sourish Chaudhuri, Joseph Roth, Daniel P. W. Ellis, Andrew Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin Wilson, Zhonghua Xi 

Audiovisual Speech Activity Detection with Advanced Long Short-Term Memory

Fei Tao, Carlos Busso 

Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI

Prमित Saha, Praneeth Srungarapu, Sidney Fels 

Selected Topics in Neural Speech Processing

Structured Word Embedding for Low Memory Neural Network Language Model

Kaiyu Shi, Kai Yu 

Role Play Dialogue Aware Language Models Based on Conditional Hierarchical Recurrent Encoder-Decoder

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Hirokazu Masataki, Yushi Aono 

Efficient Keyword Spotting Using Time Delay Neural Networks

Samuel Myer, Vikrant Singh Tomar 

Automatic DNN Node Pruning Using Mixture Distribution-based Group Regularization

Tsukasa Yoshida, Takafumi Moriya, Kazuho Watanabe, Yusuke Shinohara, Yoshikazu Yamaguchi, Yushi Aono 

Conditional-Computation-Based Recurrent Neural Networks for Computationally Efficient Acoustic Modelling

Raffaele Tavarone, Leonardo Badino 

Leveraging Translations for Speech Transcription in Low-resource Settings

Antonios Anastasopoulos, David Chiang 

Sequence-to-sequence Neural Network Model with 2D Attention for Learning Japanese Pitch Accents

Antoine Bruguier, Heiga Zen, Arkady Arkhangorodsky 

Task Specific Sentence Embeddings for ASR Error Detection

Sahar Ghannay, Yannick Estève, Nathalie Camelin 

Low-Latency Neural Speech Translation

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, Alex Waibel 

Low-Resource Speech-to-Text Translation

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, Sharon Goldwater 

VoiceGuard: Secure and Private Speech Processing

Ferdinand Brasser, Tommaso Frassetto, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, Christian Weinert 

Perspective Talk-1

Deep Learning based Situated Goal-oriented Dialogue Systems

Dilek Hakkani-Tür 

Dereverberation

Single-channel Speech Dereverberation via Generative Adversarial Training

Chenxing Li, Tieqiang Wang, Shuang Xu, Bo Xu 

Single-Channel Dereverberation Using Direct MMSE Optimization and Bidirectional LSTM Networks

Wolfgang Mack, Sourmitro Chakrabarty, Fabian-Robert Stöter, Sebastian Braun, Bernd Edler, Emanuel Habets 

Single-channel Late Reverberation Power Spectral Density Estimation Using Denoising Autoencoders

Ina Kodrasi, Hervé Bourlard 

A Non-convolutive NMF Model for Speech Dereverberation

Nikhil M, Rajbabu Velmurugan, Preeti Rao 

Cross-Corpora Convolutional Deep Neural Network Dereverberation Preprocessing for Speaker Verification and Speech Enhancement

Peter Guzewich, Stephen Zahorian, Xiao Chen, Hao Zhang 

Dereverberation and Beamforming in Robust Far-Field Speaker Recognition

Ladislav Mošner, Oldřich Plchot, Pavel Matějka, Ondřej Novotný, Jan Černocký 

Audio Events and Acoustic Scenes

Comparing the Max and Noisy-Or Pooling Functions in Multiple Instance Learning for Weakly Supervised Sequence Learning Tasks

Yun Wang, Juncheng Li, Florian Metze 

A Simple Model for Detection of Rare Sound Events

Weiran Wang, Chieh-Chi Kao, Chao Wang 

Temporal Transformer Networks for Acoustic Scene Classification

Teng Zhang, Kailai Zhang, Ji Wu 

Temporal Attentive Pooling for Acoustic Event Detection

Xugang Lu, Peng Shen, Sheng Li, Yu Tsao, Hisashi Kawai 

R-CRNN: Region-based Convolutional Recurrent Neural Network for Audio Event Detection

Chieh-Chi Kao, Weiran Wang, Ming Sun, Chao Wang 

Detecting Media Sound Presence in Acoustic Scenes

Constantinos Papayiannis, Justice Amoh, Viktor Rozgic, Shiva Sundaram, Chao Wang 

Speaker Diarization

S4D: Speaker Diarization Toolkit in Python

Pierre-Alexandre Broux, Florent Desnous, Anthony Larcher, Simon Petitrenaud, Jean Carrive, Sylvain Meignier 

Multimodal Speaker Segmentation and Diarization Using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks

Tae Jin Park, Panayiotis Georgiou 

Combined Speaker Clustering and Role Recognition in Conversational Speech

Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, Shrikanth Narayanan 

The ACLEW DiViME: An Easy-to-use Diarization Tool

Adrien Le Franc, Eric Riebling, Julien Karadayi, Yun Wang, Camila Scaff, Florian Metze, Alejandrina Cristia 

Automatic Detection of Multi-speaker Fragments with High Time Resolution

Evdokia Kazimirova, Andrey Belyaev 

Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization

Ruiqing Yin, Hervé Bredin, Claude Barras 

Phonation

Pitch or Phonation: on the Glottalization in Tone Productions in the Ruokeng Hui Chinese Dialect

Minghui Zhang, Fang Hu 

Speaker-specific Structure in German Voiceless Stop Voice Onset Times

Marc Antony Hullebus, Stephen Tobin, Adamantios Gafos 

Creak in the Respiratory Cycle

Kätlin Aare, Pärtel Lippus, Marcin Włodarczak, Mattias Heldner 

Acoustic Analysis of Whispersy Voice Disguise in Mandarin Chinese

Cuiling Zhang, Bin Li, Si Chen, Yike Yang 

The Zurich Corpus of Vowel and Voice Quality, Version 1.0

Dieter Maurer, Christian d'Heureuse, Heidy Suter, Volker Dellwo, Daniel Friedrichs, Thayabaran Kathiresan 

Weighting of Coda Voicing Cues: Glottalisation and Vowel Duration

Joshua Penney, Felicity Cox, Anita Szakay 

Cognition and Brain Studies

Revealing Spatiotemporal Brain Dynamics of Speech Production Based on EEG and Eye Movement

Bin Zhao, Jinfeng Huang, Gaoyan Zhang, Jianwu Dang, Minbo Chen, YingjianFu, Longbiao Wang 

Neural Response Development During Distributional Learning

Natalie Boll-Avetisyan, Jessie S. Nixon, Tomas O. Lentz, Lique Liu, Sandrien van Ommen, Çağrı Çöltekin, Jacolien van Rijn 

Learning Two Tone Languages Enhances the Brainstem Encoding of Lexical Tones

Akshay Raj Maggu, Wenqing Zong, Vina Law, Patrick C. M. Wong 

Perceptual Sensitivity to Spectral Change in Australian English Close Front Vowels: An Electroencephalographic Investigation

Daniel Williams, Paola Escudero, Adamantios Gafos 

Effective Acoustic Cue Learning Is Not Just Statistical, It Is Discriminative

Jessie S. Nixon 

Analyzing EEG Signals in Auditory Speech Comprehension Using Temporal Response Functions and Generalized Additive Models

Kimberley Mulder, Louis ten Bosch, Lou Boves 

Deep Neural Networks: How Can We Interpret What They Learned?

Information Encoding by Deep Neural Networks: What Can We Learn?

Louis ten Bosch, Lou Boves 

Scalable Factorized Hierarchical Variational Autoencoder Training

Wei-Ning Hsu, James Glass 

State Gradients for RNN Memory Analysis

Lyan Verwimp, Hugo van Hamme, Vincent Renkens, Patrick Wambacq 

Exploring How Phone Classification Neural Networks Learn Phonetic Information by Visualising and Interpreting Bottleneck Features

Linxue Bai, Philip Weber, Peter Jančovič, Martin Russell 

Memory Time Span in LSTMs for Multi-Speaker Source Separation

Jeroen Zegers, Hugo van Hamme 

Visualizing Phoneme Category Adaptation in Deep Neural Networks

Odette Scharenborg, Sebastian Tiesmeyer, Mark Hasegawa-Johnson, Najim Dehak 

Early Vocabulary Development Through Picture-based Software Solutions

Kasthuri G, Prabha Ramanathan, Hema Murthy, Namita Jacob, Anil Prabhakar 

Automatic Detection of Expressiveness in Oral Reading

Kamini Sabu, Kanhaiya Kumar, Preeti Rao 

PannoMulloKathan: Voice Enabled Mobile App for Agricultural Commodity Price Dissemination in Bengali Language

Madhab Pal, Rajib Roy, Soma Khan, Milton S. Bepari, Joyanta Basu 

Visualizing Punctuation Restoration in Speech Transcripts with Prosograph

Alp Oktem, Mireia Farrús, Antonio Bonafonte 

CACTAS - Collaborative Audio Categorization and Transcription for ASR Systems

Mithul Mathivanan, Kinnera Saranu, Abhishek Pandey, Jithendra Vepa 

Speech and Singing Production

FACTS: A Hierarchical Task-based Control Model of Speech Incorporating Sensory Feedback

Benjamin Parrell, Vikram Ramanarayanan, Srikantan Nagarajan, John Houde 

Sensorimotor Response to Tongue Displacement Imagery by Talkers with Parkinson's Disease

William Katz, Patrick Reidy, Divya Prabhakaran 

Automatic Pronunciation Evaluation of Singing

Chitrakha Gupta, Haizhou Li, Ye Wang 

Classification of Nonverbal Human Produced Audio Events: A Pilot Study

Rachel E. Bouserhal, Philippe Chabot, Milton Sarria-Paja, Patrick Cardinal, Jérémie Voix 

UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science

Lorenzo Spreafico, Michael Pucher, Anna Matosova 

Articulatory Consequences of Vocal Effort Elicitation Method

Elisabet Eir Cortes, Marcin Włodarczyk, Juraj Šimko 

Age-related Effects on Sensorimotor Control of Speech Production

Anne Hermes, Jane Mertens, Doris Mücke 

An Ultrasound Study of Gemination in Coronal Stops in Eastern Oromo

Maida Percival, Alexei Kochetov, Yoonjung Kang 

Processing Transition Regions of Glottal Stop Substituted /S/ for Intelligibility Enhancement of Cleft Palate Speech

Protima Nomo Sudro, Sishir Kalita, S R Mahadeva Prasanna 

Reconstructing Neutral Speech from Tracheoesophageal Speech

Abinay Reddy N, Achuth Rao MV, G. Nisha Meenakshi, Prasanta Kumar Ghosh 

Automatic Evaluation of Soft Articulatory Contact for Stuttering Treatment

Keiko Ochi, Koichi Mori, Naomi Sakai 

Korean Singing Voice Synthesis Based on an LSTM Recurrent Neural Network

Juntae Kim, Heejin Choi, Jinuk Park, Minsoo Hahn, Sangjin Kim, Jong-jin Kim 

The Trajectory of Voice Onset Time with Vocal Aging

Chen Xuanda, Xiong Ziyu, Hu Jian 

Robust Speech Recognition

The Fifth 'ChiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines

Jon Barker, Shinji Watanabe, Emmanuel Vincent, Jan Trmal 

Voices Obscured in Complex Environmental Settings (VOICES) Corpus

Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeffrey Hetherly, Cory Stephenson, Karl Ni 

Building State-of-the-art Distant Speech Recognition Using the ChiME-4 Challenge with a Setup of Speech Enhancement Baseline

Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, Shinji Watanabe 

Unsupervised Adaptation with Interpretable Disentangled Representations for Distant Conversational Speech Recognition

Wei-Ning Hsu, Hao Tang, James Glass 

Investigating Generative Adversarial Networks Based Speech Dereverberation for Robust Speech Recognition

Ke Wang, Junbo Zhang, Sining Sun, Yujun Wang, Fei Xiang, Lei Xie 

Monaural Multi-Talker Speech Recognition with Attention Mechanism and Gated Convolutional Networks

IPR2026-00272

Xuankai Chang, Yanmin Qian, Dong Yu 

Weighting Time-Frequency Representation of Speech Using Auditory Saliency for Automatic Speech Recognition

Cong-Thanh Do, Yannis Stylianou 

Acoustic Modeling from Frequency Domain Representations of Speech

Pegah Ghahremani, Hossein Hadian, Hang Lv, Daniel Povey, Sanjeev Khudanpur 

Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition

Ishwar Chandra Yadav, Avinash Kumar, Syed Shahnawazuddin, Gayadhar Pradhan 

Bidirectional Long-Short Term Memory Network-based Estimation of Reliable Spectral Component Locations

Aaron Nicolson, Kuldip K. Paliwal 

Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network

Lili Guo, Longbiao Wang, Jianwu Dang, Linjuan Zhang, Haotian Guan, Xiangang Li 

Bubble Cooperative Networks for Identifying Important Speech Cues

Viet Anh Trinh, Brian McFee, Michael I Mandel 

Applications in Education and Learning

Real-Time Scoring of an Oral Reading Assessment on Mobile Devices

Jian Cheng 

A Deep Learning Approach to Assessing Non-native Pronunciation of English Using Phone Distances

Konstantinos Kyriakopoulos, Kate Knill, Mark Gales 

Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment

Yujia Xiao, Frank Soong, Wenping Hu 

Investigating the Role of L1 in Automatic Pronunciation Evaluation of L2 Speech

Ming Tu, Anna Grabek, Julie Liss, Visar Berisha 

Impact of ASR Performance on Free Speaking Language Assessment

Kate Knill, Mark Gales, Konstantinos Kyriakopoulos, Andrey Malinin, Anton Ragni, Yu Wang, Andrew Caines 

Automatic Miscue Detection Using RNN Based Models with Data Augmentation

Yoon Seok Hong, Kyung Seo Ki, Gahgene Gweon 

A Study of Objective Measurement of Comprehensibility through Native Speakers' Shadowing of Learners' Utterances

Yusuke Inoue, Suguru Kabashima, Daisuke Saito, Nobuaki Minematsu, Kumi Kanamura, Yutaka Yamauchi 

Factorized Deep Neural Network Adaptation for Automatic Scoring of L2 Speech in English Speaking Tests

Dean Luo, Chunxiao Zhang, Linzhong Xia, Lixin Wang 

On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children

Gary Yeung, Abeer Alwan 

Improved Acoustic Modelling for Automatic Literacy Assessment of Children

Mauro Nicolao, Michiel Sanders, Thomas Hain 

Integrating Speech Science and Technology for Clinical Applications

Anomaly Detection Approach for Pronunciation Verification of Disordered Speech Using Speech Attribute Features

Mostafa Shahin, Beena Ahmed, Jim X. Ji, Kirrie Ballard 

Effectiveness of Voice Quality Features in Detecting Depression

Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, Abeer Alwan 

Fusing Text-dependent Word-level i-Vector Models to Screen 'at Risk' Child Speech

Prasanna Kothalkar, Johanna Rudolph, Christine Dollaghan, Jennifer McGlothlin, Thomas Campbell, John H.L. Hansen 

Testing Paradigms for Assistive Hearing Devices in Diverse Acoustic Environments

Ram Charan Chandra Shekar, Hussnain Ali, John H.L. Hansen 

Detection of Dementia from Responses to Atypical Questions Asked by Embodied Conversational Agents

Tsuyoki Ujiro, Hiroki Tanaka, Hiroyoshi Adachi, Hiroaki Kazui, Manabu Ikeda, Takashi Kudo, Satoshi Nakamura 

Acoustic Features Associated with Sustained Vowel and Continuous Speech Productions by Chinese Children with Functional Articulation Disorders

Wang Zhang, Xiangquan Gui, Tianqi Wang, Manwa Ng, Feng Yang, Lan Wang, Nan Yan 

Estimation of Hypernasality Scores from Cleft Lip and Palate Speech

Vikram C M, Ayush Tripathi, Sishir Kalita, S R Mahadeva Prasanna 

Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data

Tifani Warnita, Nakamasa Inoue, Koichi Shinoda 

IPR2026-00272

Automatic Detection of Orofacial Impairment in Stroke

Andrea Bandini, Jordan Green, Brian Richburg, Yana Yunusova 

Detecting Depression with Audio/Text Sequence Modeling of Interviews

Tuka Al Hanai, Mohammad Ghassemi, James Glass 

Speaker Characterization and Analysis

Discourse Marker Detection for Hesitation Events on Mandarin Conversation

Yu-Wun Wang, Hen-Hsen Huang, Kuan-Yu Chen, Hsin-Hsi Chen 

Acoustic and Perceptual Characteristics of Mandarin Speech in Homosexual and Heterosexual Male Speakers

Puyang Geng, Wentao Gu, Hiroya Fujisaki 

Automatic Question Detection from Acoustic and Phonetic Features Using Feature-wise Pre-training

Atsushi Ando, Reine Asakawa, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, Yushi Aono 

Improving Response Time of Active Speaker Detection Using Visual Prosody Information Prior to Articulation

Fasih Haider, Saturnino Luz, Carl Vogel, Nick Campbell 

Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions

Bekir Berker Türker, Engin Erzin, Yücel Yemez, Metin Sezgin 

Analyzing Effect of Physical Expression on English Proficiency for Multimodal Computer-Assisted Language Learning

Haoran Wu, Yuya Chiba, Takashi Nose, Akinori Ito 

Analysis of the Effect of Speech-Laugh on Speaker Recognition System

Sri Harsha Dumpala, Ashish Panda, Sunil Kumar Kopparapu 

Vocal Biomarkers for Cognitive Performance Estimation in a Working Memory Task

Jennifer Sloboda, Adam Lammert, James Williamson, Christopher Smalt, Daryush D. Mehta, COL Ian Curry, Kristin Heaton, Jeffrey Palmer, Thomas Quatieri 

Lexical and Acoustic Deep Learning Model for Personality Recognition

Guozhen An, Rivka Levitan 

Perspective Talk-2

Open Problems in Speech Recognition

Bhuvana Ramabhadran 

Plenary Talk-2

Evolution of Neural Network Architectures for Speech Recognition

Hervé Bourlard 

Novel Neural Network Architectures for Acoustic Modelling

Layer Trajectory LSTM

Jinyu Li, Changliang Liu, Yifan Gong 

Semi-tied Units for Efficient Gating in LSTM and Highway Networks

Chao Zhang, Philip Woodland 

Gaussian Process Neural Networks for Speech Recognition

Max W. Y. Lam, Shoukang Hu, Xurong Xie, Shansong Liu, Jianwei Yu, Rongfeng Su, Xunying Liu, Helen Meng 

Acoustic Modeling with Densely Connected Residual Network for Multichannel Speech Recognition

Jian Tang, Yan Song, Lirong Dai, Ian McLoughlin 

Gated Recurrent Unit Based Acoustic Modeling with Future Context

Jie Li, Xiaorui Wang, Yuanyuan Zhao, Yan Li 

Output-Gate Projected Gated Recurrent Unit for Speech Recognition

Gaofeng Cheng, Daniel Povey, Lu Huang, Ji Xu, Sanjeev Khudanpur, Yonghong Yan 

Language Identification

Performance Analysis of the 2017 NIST Language Recognition Evaluation

IPR2026-00272

Seyed Omid Sadjadi, Timothee Kheyrkha, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, Jaime Hernandez-Cordero 

Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal

Lukas Mateju, Petr Cerva, Jindrich Zdansky, Radek Safarik 

Adding New Classes without Access to the Original Training Data with Applications to Language Identification

Hagai Taitelbaum, Ehud Ben-Reuven, Jacob Goldberger 

Feature Representation of Short Utterances Based on Knowledge Distillation for Spoken Language Identification

Peng Shen, Xugang Lu, Sheng Li, Hisashi Kawai 

Sub-band Envelope Features Using Frequency Domain Prediction for Short Duration Language Identification

Sarith Fernando, Vidhyasaharan Sethu, Eliathamby Ambikairajah 

Effectiveness of Single-Channel BLSTM Enhancement for Language Identification

Peter Sibbern Frederiksen, Jesús Villalba, Shinji Watanabe, Zheng-Hua Tan, Najim Dehak 

Production of Prosody

Articulation Rate as a Speaker Discriminant in British English

Erica Gold 

Truncation and Compression in Southern German and Australian English

Jenny Yu, Katharina Zahner 

Prominence-based Evaluation of L2 Prosody

Heini Kallio, Antti Suni, Päivi Virkkunen, Juraž Šimko 

Length Contrast and Covarying Features: Whistled Speech as a Case Study

Rachid Ridouane, Giuseppina Turco, Julien Meyer 

Information Structure, Affect and Prenuclear Prominence in American English

Eleanor Chodroff, Jennifer Cole 

Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments

John S. Novak, III, Robert V. Kenyon 

Speech Intelligibility and Quality

Binaural Speech Intelligibility Estimation Using Deep Neural Networks

Kazuhiro Kondo, Kazuya Taira, Yosuke Kobayashi 

Multi-resolution Gammachirp Envelope Distortion Index for Intelligibility Prediction of Noisy Speech

Katsuhiko Yamamoto, Toshio Irino, Narumi Ohashi, Shoko Araki, Keisuke Kinoshita, Tomohiro Nakatani 

Speech Intelligibility Enhancement Based on a Non-causal Wavenet-like Model

Muhammed Shifas PV, Vassilis Tsiaras, Yannis Stylianou 

Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM

Szu-wei Fu, Yu Tsao, Hsin-Te Hwang, Hsin-Min Wang 

Global SNR Estimation of Speech Signals Using Entropy and Uncertainty Estimates from Dropout Networks

Rohith Aralikatti, Dillip Kumar Margam, Tanay Sharma, Abhinav Thanda, Shankar Venkatesan 

Detecting Packet-Loss Concealment Using Formant Features and Decision Tree Learning

Gabriel Mittag, Sebastian Möller 

Integrating Speech Science and Technology for Clinical Applications

UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions

Aciel Eshky, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James M Scobbie, Alan Wrench 

Detecting Signs of Dementia Using Word Vector Representations

Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, Heidi Christensen 

Classification of Huntington Disease Using Acoustic and Lexical Features

Matthew Perez, Wenyu Jin, Duc Le, Noelle Carlozzi, Praveen Dayalu, Angela Roberts, Emily Mower Provost 

The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild

Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, Emily Mower Provost 

Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions

Nikolaos Flemotomos, Victor Martinez, James Gibson, David Atkins, Torrey Creed, Shrikanth Narayanan 

Speech Technologies for Code-Switching in Multilingual Communities

A Study of Lexical and Prosodic Cues to Segmentation in a Hindi-English Code-switched Discourse

Preeti Rao, Mugdha Pandya, Kamini Sabu, Kanhaiya Kumar, Nandini Bondale 

Building a Unified Code-Switching ASR System for South African Languages

Emre Yilmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, Thomas Niesler 

Study of Semi-supervised Approaches to Improving English-Mandarin Code-Switching Speech Recognition

Pengcheng Guo, Haihua Xu, Lei Xie, Eng Siong Chng 

Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech

Emre Yilmaz, Henk van den Heuvel, David van Leeuwen 

The Role of Cognate Words, POS Tags and Entrainment in Code-Switching

Victor Soto, Nishmar Cestero, Julia Hirschberg 

Homophone Identification and Merging for Code-switched Speech Recognition

Brij Mohan Lal Srivastava, Sunayana Sitaram 

Code-switching in Indic Speech Synthesizers

Anju Leela Thomas, Anusha Prakash, Arun Baby, Hema Murthy 

A Novel Approach for Effective Recognition of the Code-Switched Data on Monolingual Language Model

Sreeram Ganji, Rohit Sinha 

Show and Tell 5

Hierarchical Accent Determination and Application in a Large Scale ASR System

Ramya Viswanathan, Periyasamy Paramasivam, Jithendra Vepa 

Toward Scalable Dialog Technology for Conversational Language Learning: Case Study of the TOEFL® MOOC

Vikram Ramanarayanan, David Pautler, Patrick Lange, Eugene Tsuprun, Rutuja Ubale, Keelan Evanini, David Suendermann-Oeft 

Machine Learning Powered Data Platform for High-Quality Speech and NLP Workflows

João Freitas, Jorge Ribeiro, Daan Baldwijn, Sara Oliveira, Daniela Braga 

Fully Automatic Speaker Separation System, with Automatic Enrolling of Recurrent Speakers

Raphael Cohen, Orgad Keller, Jason Levy, Russell Levy, Micha Breakstone, Amit Ashkenazi 

Online Speech Translation System for Tamil

Madhavaraj Ayyavu, Shiva Kumar H R, Ramakrishnan A G 

Voice Conversion and Speech Synthesis

Unsupervised Vocal Tract Length Warped Posterior Features for Non-Parallel Voice Conversion

Nirmesh Shah, Maulik C. Madhavi, Hemant Patil 

Voice Conversion with Conditional SampleRNN

Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, Dan Darcy 

A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder

Berrak Sisman, Mingyang Zhang, Haizhou Li 

WaveNet Vocoder with Limited Training Data for Voice Conversion

Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, Li-Rong Dai 

Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder

Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, Tomoki Toda 

High-quality Voice Conversion Using Spectrogram-Based WaveNet Vocoder

Kuan Chen, Bo Chen, Jiahao Lai, Kai Yu 

Spanish Statistical Parametric Speech Synthesis Using a Neural Vocoder

Antonio Bonafonte, Santiago Pascual, Georgina Dorca 

Experiments with Training Corpora for Statistical Text-to-speech Systems.

Monika Podsiadko, Victor Ungureanu 

Multi-task WaveNet: A Multi-task Generative Model for Statistical Parametric Speech Synthesis without Fundamental Frequency Conditions

Yu Gu, Yongguo Kang 

Speaker-independent Raw Waveform Model for Glottal Excitation

Lauri Juvela, Vassilis Tsirias, Bajjibabu Bollepalli, Manu Airaksinen, Junichi Yamagishi, Paavo Alku 

A New Glottal Neural Vocoder for Speech Synthesis

Yang Cui, Xi Wang, Lei He, Frank K. Soong 

Exemplar-based Speech Waveform Generation

Oliver Watts, Cassia Valentini-Botinhao, Felipe Espic, Simon King 

Frequency Domain Variants of Velvet Noise and Their Application to Speech Processing and Synthesis

Hideki Kawahara, Ken-Ichi Sakakibara, Masanori Morise, Hideki Banno, Tomoki Toda, Toshio Irino 

Extracting Information from Audio

Joint Learning of Interactive Spoken Content Retrieval and Trainable User Simulator

Pei-Hung Chung, Kuan Tung, Ching-Lun Tai, Hung-yi Lee 

Attention-based End-to-End Models for Small-Footprint Keyword Spotting

Changhao Shan, Junbo Zhang, Yujun Wang, Lei Xie 

Prediction of Aesthetic Elements in Karnatic Music: A Machine Learning Approach

Ragesh Rajan M, Ashwin Vijayakumar, Deepu Vijayaseenan 

Topic and Keyword Identification for Low-resourced Speech Using Cross-Language Transfer Learning

Wenda Chen, Mark Hasegawa-Johnson, Nancy F. Chen 

Automatic Speech Recognition and Topic Identification from Speech for Almost-Zero-Resource Languages

Matthew Wiesner, Chunxi Liu, Lucas Ondel, Craig Harman, Vimal Manohar, Jan Trmal, Zhongqiang Huang, Najim Dehak, Sanjeev Khudanpur 

Play Duration Based User-Entity Affinity Modeling in Spoken Dialog System

Bo Xiao, Nicholas Monath, Shankar Ananthakrishnan, Abishek Ravi 

Empirical Analysis of Score Fusion Application to Combined Neural Networks for Open Vocabulary Spoken Term Detection

Shi-wook Lee, Kazuyo Tanaka, Yoshiaki Itoh 

Phonological Posterior Hashing for Query by Example Spoken Term Detection

Afsaneh Asaei, Dhananjay Ram, Hervé Bourlard 

Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, Sebastian Stüker 

Semi-supervised Learning for Information Extraction from Dialogue

Anjuli Kannan, Kai Chen, Diana Jaunzeikare, Alvin Rajkomar 

Slot Filling with Delexicalized Sentence Generation

Youhyun Shin, Kang Min Yoo, Sang-goo Lee 

Music Genre Recognition Using Deep Neural Networks and Transfer Learning

Deepanway Ghosal, Maheshkumar H. Kolekar 

Efficient Voice Trigger Detection for Low Resource Hardware

Siddharth Sigtia, Rob Haynes, Hywel Richards, Erik Marchi, John Bridle 

Signal Analysis for the Natural, Biological and Social Sciences

A Novel Normalization Method for Autocorrelation Function for Pitch Detection and for Speech Activity Detection

Qiguang Lin, Yiwen Shao 

Estimation of the Vocal Tract Length of Vowel Sounds Based on the Frequency of the Significant Spectral Valley

TV Ananthapadmanabha, Ramakrishnan A G 

Deep Learning Techniques for Koala Activity Detection

Ivan Himawan, Michael Towsey, Bradley Law, Paul Roe 

Glottal Closure Instant Detection from Speech Signal Using Voting Classifier and Recursive Feature Elimination

Jindřich Matoušek, Daniel Tihelka 

Assessing Speaker Engagement in 2-Person Debates: Overlap Detection in United States Presidential Debates

Midia Yousefi, Navid Shokouhi, John H.L. Hansen 

All-Conv Net for Bird Activity Detection: Significance of Learned Pooling

Arjun Pankajakshan, Anshul Thakur, Daksh Thapar, Padmanabhan Rajan, Aditya Nigam 

Deep Convex Representations: Feature Representations for Bioacoustics Classification

Anshul Thakur, Vinayak Abrol, Pulkit Sharma, Padmanabhan Rajan 

Detection of Glottal Excitation Epochs in Speech Signal Using Hilbert Envelope

IPR2026-00272

Hirak Dasgupta, Prem C. Pandey, K S Nataraj 

Analyzing Thai Tone Distribution through Functional Data Analysis

Hong Zhang 

Articulatory Feature Classification Using Convolutional Neural Networks

Danny Merkx, Odette Scharenborg 

A New Frequency Coverage Metric and a New Subband Encoding Model, with an Application in Pitch Estimation

Shoufeng Lin 

Improved Epoch Extraction from Telephonic Speech Using Chebfun and Zero Frequency Filtering

Ganga Gowri B, Soman K.P, Govind D 

Speech Prosody

An Empirical Analysis of the Correlation of Syntax and Prosody

Arne Köhn, Timo Baumann, Oskar Dörfler 

Analysing the Focus of a Hierarchical Attention Network: the Importance of Enjambments When Classifying Post-modern Poetry

Timo Baumann, Hussein Hussein, Burkhard Meyer-Sickendiek 

Language-Dependent Melody Embeddings

Daniil Kocharov, Alla Menshikova 

Stress Distribution of Given Information in Chinese Reading Texts

Yuan Jia, Xiaoxiao Ma 

Acoustic-prosodic Entrainment in Structural Metadata Events

Vera Cabarrão, Fernando Batista, Helena Moniz, Isabel Trancoso, Ana Isabel Mata 

Formant Measures of Vowels Adjacent to Alveolar and Retroflex Consonants in Arrernte: Stressed and Unstressed Position

Marija Tabain, Richard Beare, Andrew Butcher 

Automatic Assessment of L2 English Word Prosody Using Weighted Distances of F0 and Intensity Contours

Quy-Thao Truong, Tsuneo Kato, Seichi Yamamoto 

Homogeneity vs Heterogeneity in Indian English: Investigating Influences of L1 on f0 Range

Olga Maxwell, Elinor Payne, Rosey Billington 

Emotional Prosody Perception in Mandarin-speaking Congenital Amusics

Yixin Zhang, Tianzhu Geng, Jinsong Zhang 

Cultural Differences in Pattern Matching: Multisensory Recognition of Socio-affective Prosody

Takaaki Shochi, Jean-Luc Rouas, Marine Guerry, Donna Erickson 

Perspective Talk-3

Speech Processing in the Human Brain Meets Deep Learning

Nima Mesgarani 

Recurrent Neural Models for ASR

ESPnet: End-to-End Speech Processing Toolkit

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, Tsubasa Ochiai 

A GPU-based WFST Decoder with Exact Lattice Generation

Zhehuai Chen, Justin Luitjens, Hainan Xu, Yiming Wang, Daniel Povey, Sanjeev Khudanpur 

Automatic Speech Recognition System Development in the "Wild"

Anton Ragni, Mark Gales 

Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant

Leonid Velikovich, Ian Williams, Justin Scheiner, Petar Aleksic, Pedro Moreno, Michael Riley 

Contextual Speech Recognition in End-to-end Neural Network Systems Using Beam Search

Ian Williams, Anjuli Kannan, Petar Aleksic, David Rybach, Tara Sainath 

Forward-Backward Attention Decoder

Masato Mimura, Shinsuke Sakai, Tatsuya Kawahara 

Learning Discriminative Features for Speaker Identification and Verification

Sarthak Yadav, Atul Rai 

Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition

Sergey Novoselov, Vadim Shchemelinin, Andrey Shulipa, Alexandr Kozlov, Ivan Kremnev 

Speaker Embedding Extraction with Phonetic Information

Yi Liu, Liang He, Jia Liu, Michael T. Johnson 

Attentive Statistics Pooling for Deep Speaker Embedding

Koji Okabe, Takafumi Koshinaka, Koichi Shinoda 

Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization

Nam Le, Jean-Marc Odobez 

Deep Discriminative Embeddings for Duration Robust Speaker Verification

Na Li, Deyi Tuo, Dan Su, Zhifeng Li, Dong Yu 

Speech Perception in Adverse Conditions

Impact of Different Speech Types on Listening Effort

Olympia Simantiraki, Martin Cooke, Simon King 

Who Are You Listening to? Towards a Dynamic Measure of Auditory Attention to Speech-on-speech.

Moira-Phoebé Huet, Christophe Michéyl, Etienne Gaudrain, Etienne Parizet 

Investigating the Role of Familiar Face and Voice Cues in Speech Processing in Noise

Jeesun Kim, Sonya Karisma, Vincent Aubanel, Chris Davis 

The Conversation Continues: the Effect of Lyrics and Music Complexity of Background Music on Spoken-Word Recognition

Odette Scharenborg, Martha Larson 

Loud and Shouted Speech Perception at Variable Distances in a Forest

Julien Meyer, Fanny Meunier, Laure Dentel, Noelia Do Carmo Blanco, Frédéric Sèbe 

Phoneme Resistance and Phoneme Confusion in Noise: Impact of Dyslexia

Noelia Do Carmo Blanco, Julien Meyer, Michel Hoen, Fanny Meunier 

Measuring Pitch and Articulation

Conditional End-to-End Audio Transforms

Albert Haque, Michelle Guo, Prateek Verma 

Detection of Glottal Closure Instants in Degraded Speech Using Single Frequency Filtering Analysis

Gunnam Aneja, Sudarsana Reddy Kadiri, Bayya Yegnanarayana 

Tone Recognition Using Lifters and CTC

Loren Lugosch, Vikrant Singh Tomar 

Epoch Extraction from Pathological Children Speech Using Single Pole Filtering Approach

Vikram C M, S R Mahadeva Prasanna 

Automated Classification of Vowel-Gesture Parameters Using External Broadband Excitation

Balamurali B T, Jer-Ming Chen 

Estimation of Fundamental Frequency from Singing Voice Using Harmonics of Impulse-like Excitation Source

Sudarsana Reddy Kadiri, Bayya Yegnanarayana 

Speech and Language Analytics for Mental Health

Investigating the Effect of Audio Duration on Dementia Detection Using Acoustic Features

Jochen Weiner, Miguel Angrick, Srinivasan Umesh, Tanja Schultz 

An Interlocutor-Modulated Attentional LSTM for Differentiating between Subgroups of Autism Spectrum Disorder

Yun-Shao Lin, Susan Shur-Fen Gau, Chi-Chun Lee 

Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks

Shahin Amiriparian, Alice Baird, Sahib Julka, Alyssa Alcorn, Sandra Otti, Sunčica Petrović, Eloise Ainger, Nicholas Cummins, Björn Schuller 

Modeling Interpersonal Influence of Verbal Behavior in Couples Therapy Dyadic Interactions

Sandeep Nallan Chakravarthula, Brian Baucom, Panayiotis Georgiou 

Computational Modeling of Conversational Humor in Psychotherapy

IPR2026-00272

Multimodal I-vectors to Detect and Evaluate Parkinson's Disease

Nicanor García, Juan Camilo Vásquez Correa, Juan Rafael Orozco-Arroyave, Elmar Nöth 

Spoken CALL Shared Task, Second Edition

Overview of the 2018 Spoken CALL Shared Task

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Manny Rayner, Martin Russell, Helmer Strik, Xizi Wei 

The CSU-K Rule-Based System for the 2nd Edition Spoken CALL Shared Task

Dominik Jülg, Mario Kunstek, Cem Philipp Freimoser, Kay Berkling, Mengjie Qian 

Liulishuo's System for the Spoken CALL Shared Task 2018

Huy Nguyen, Lei Chen, Ramon Prieto, Chuan Wang, Yang Liu 

An Optimization Based Approach for Solving Spoken CALL Shared Task

Mohammad Ateeq, Abualsoud Hanani, Aziz Qaroush 

The University of Birmingham 2018 Spoken CALL Shared Task Systems

Mengjie Qian, Xizi Wei, Peter Jančovič, Martin Russell 

Improvements to an Automated Content Scoring System for Spoken CALL Responses: the ETS Submission to the Second Spoken CALL Shared Task

Keelan Evanini, Matthew Mulholland, Rutuja Ubale, Yao Qian, Robert Pugh, Vikram Ramanarayanan, Aoife Cahill 

Show and Tell 6

Extracting Speaker's Gender, Accent, Age and Emotional State from Speech

Nagendra Goel, Mousmita Sarma, Tejendra Kushwah, Dharmesh Agarwal, Zikra Iqbal, Surbhi Chauhan 

Determining Speaker Location from Speech in a Practical Environment

BHVS Narayanamurthy, JV Satyanarayana, Bayya Yegnanarayana 

An Automatic Speech Transcription System for Manipuri Language

Tanvina Patel, Krishna D N, Noor Fathima, Nisar Shah, Mahima C, Deepak Kumar, Anuroop Iyengar 

SPIRE-SST: An Automatic Web-based Self-learning Tool for Syllable Stress Tutoring (SST) to the Second Language Learners

Chiranjeevi Yarra, Anand P A, Kausthubha N K, Prasanta Kumar Ghosh 

Glottal Vibrato Graph: A Device and Method for Recording, Analysis and Visualization of Glottal Activity

Kishalay Chakraborty, Senjam Shantirani Devi, Sanjeevan Devnath, S R Mahadeva Prasanna, Priyankoo Sarmah 

Adjusting to Speaker, Accent, and Domain

Multi-Modal Data Augmentation for End-to-end ASR

Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, Shinji Watanabe 

Multi-task Learning with Augmentation Strategy for Acoustic-to-word Attention-based Encoder-decoder Speech Recognition

Takafumi Moriya, Sei Ueno, Yusuke Shinohara, Marc Delcroix, Yoshikazu Yamaguchi, Yushi Aono 

Training Augmentation with Adversarial Examples for Robust Speech Recognition

Sining Sun, Ching-Feng Yeh, Mari Ostendorf, Mei-Yuh Hwang, Lei Xie 

Data Augmentation Improves Recognition of Foreign Accented Speech

Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, Gakuto Kurata 

Speaker Adaptive Training and Mixup Regularization for Neural Network Acoustic Models in Automatic Speech Recognition

Natalia Tomashenko, Yuri Khokhlov, Yannick Estève 

Neural Language Codes for Multilingual Acoustic Models

Markus Müller, Sebastian Stüker, Alex Waibel 

Encoder Transfer for Attention-based Acoustic-to-word Speech Recognition

Sei Ueno, Takafumi Moriya, Masato Mimura, Shinsuke Sakai, Yusuke Shinohara, Yoshikazu Yamaguchi, Yushi Aono, Tatsuya Kawahara 

Empirical Evaluation of Speaker Adaptation on DNN Based Acoustic Model

Ke Wang, Junbo Zhang, Yujun Wang, Lei Xie 

Improving DNNs Trained with Non-Native Transcriptions Using Knowledge Distillation and Target Interpolation

Amit Das, Mark Hasegawa-Johnson 

Improving Cross-Lingual Knowledge Transferability Using Multilingual TDNN-BLSTM with Language-Dependent Pre-Final Layer

Siyuan Feng, Tan Lee 

Auxiliary Feature Based Adaptation of End-to-end ASR Systems

Marc Delcroix, Shinji Watanabe, Atsunori Ogawa, Shigeki Karita, Tomohiro Nakatani 

Leveraging Native Language Information for Improved Accented Speech Recognition

Shahram Ghorbani, John H.L. Hansen 

Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning

Abhinav Jain, Minall Upreti, Preethi Jyothi 

Fast Language Adaptation Using Phonological Information

Sibo Tong, Philip N. Garner, Hervé Bourlard 

Speech Synthesis Paradigms and Methods

Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient's Speech Using Spectral Differential Modification in Voice Conversion

Hiroki Murakami, Sunao Hara, Masanobu Abe, Masaaki Sato, Shogo Minagi 

Audio-visual Voice Conversion Using Deep Canonical Correlation Analysis for Deep Bottleneck Features

Satoshi Tamura, Kento Horio, Hajime Endo, Satoru Hayamizu, Tomoki Toda 

An Investigation of Convolution Attention Based Models for Multilingual Speech Synthesis of Indian Languages

Pallavi Baljekar, SaiKrishna Rallabandi, Alan W Black 

The Effect of Real-Time Constraints on Automatic Speech Animation

Danny Websdale, Sarah Taylor, Ben Milner 

Joint Learning of Facial Expression and Head Pose from Speech

David Greenwood, Iain Matthews, Stephen Laycock 

Acoustic-dependent Phonemic Transcription for Text-to-speech Synthesis

Kévin Vythelingum, Yannick Estève, Olivier Rosec 

Multimodal Speech Synthesis Architecture for Unsupervised Speaker Adaptation

Hieu-Thi Luong, Junichi Yamagishi 

Articulatory-to-speech Conversion Using Bi-directional Long Short-term Memory

Fumiaki Taguchi, Tokihiko Kaburagi 

Implementation of Respiration in Articulatory Synthesis Using a Pressure-Volume Lung Model

Keisuke Tanihara, Shogo Yonekura, Yasuo Kuniyoshi 

Learning and Modeling Unit Embeddings for Improving HMM-based Unit Selection Speech Synthesis

Xiao Zhou, Zhen-Hua Ling, Zhi-Ping Zhou, Li-Rong Dai 

Deep Metric Learning for the Target Cost in Unit-Selection Speech Synthesizer

Ruibo Fu, Jianhua Tao, Yibin Zheng, Zhengqi Wen 

DNN-based Speech Synthesis for Small Data Sets Considering Bidirectional Speech-Text Conversion

Kentaro Sone, Toru Nakashika 

A Weighted Superposition of Functional Contours Model for Modelling Contextual Prominence of Elementary Prosodic Contours

Branislav Gerazov, Gérard Bailly, Yi Xu 

LSTBM: A Novel Sequence Representation of Speech Spectra Using Restricted Boltzmann Machine with Long Short-Term Memory

Toru Nakashika 

Second Language Acquisition and Code-switching

Should Code-switching Models Be Asymmetric?

Barbara E. Bullock, Gualberto Guzmán, Jacqueline Serigos, Almeida Jacqueline Toribio 

Cross-language Perception of Mandarin Lexical Tones by Mongolian-speaking Bilinguals in the Inner Mongolia Autonomous Region, China

Kimiko Tsukada, Yu Rong 

Automatically Measuring L2 Speech Fluency without the Need of ASR: A Proof-of-concept Study with Japanese Learners of French

Lionel Fontan, Maxime Le Coz, Sylvain Detey 

Analysis of L2 Learners' Progress of Distinguishing Mandarin Tone 2 and Tone 3

Yue Sun, Win Thuzar Kyaw, Jinsong Zhang, Yoshinori Sagisaka 

Unsupervised Discovery of Non-native Phonetic Patterns in L2 English Speech for Mispronunciation Detection and Diagnosis

Xu Li, Shaoguang Mao, Xixin Wu, Kun Li, Xunying Liu, Helen Meng 

Wuxi Speakers' Production and Perception of Coda Nasals in Mandarin

Lei Wang, Jie Cui, Ying Chen 

The Diphthongs of Formal Nigerian English: A Preliminary Acoustic Analysis

Natalia Dyrenko, Robert Fuchs 

Characterizing Rhythm Differences between Strong and Weak Accented L2 Speech

Chris Davis, Jeesun Kim 

Analysis of Phone Errors Attributable to Phonological Effects Associated With Language Acquisition Through Bottleneck Feature Visualisations

Eva Fringi, Martin Russell 

Category Similarity in Multilingual Pronunciation Training

Jacques Koreman 

Talker Diarization in the Wild: the Case of Child-centered Daylong Audio-recordings

Alejandrina Cristia, Shobhana Ganesh, Marisa Casillas, Sriram Ganapathy 

Automated Classification of Children's Linguistic versus Non-Linguistic Vocalisations

Zixing Zhang, Alejandrina Cristia, Anne Warlaumont, Björn Schuller 

Pitch Characteristics of L2 English Speech by Chinese Speakers: A Large-scale Study

Jiahong Yuan, Qiusi Dong, Fei Wu, Huan Luan, Xiaofei Yang, Hui Lin, Yang Liu 

Topics in Speech Recognition

Dual Language Models for Code Switched Speech Recognition

Saurabh Garg, Tanmay Parekh, Preethi Jyothi 

Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech

Astik Biswas, Febe de Wet, Ewald van der Westhuizen, Emre Yilmaz, Thomas Niesler 

Fast ASR-free and Almost Zero-resource Keyword Spotting Using DTW and CNNs for Humanitarian Monitoring

Raghav Menon, Herman Kamper, John Quinn, Thomas Niesler 

Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection

Meng Yu, Xuan Ji, Yi Gao, Lianwu Chen, Jie Chen, Jimeng Zheng, Dan Su, Dong Yu 

Improved ASR for Under-resourced Languages through Multi-task Learning with Acoustic Landmarks

Di He, Boon Pang Lim, Xuesong Yang, Mark Hasegawa-Johnson, Deming Chen 

Cross-language Phoneme Mapping for Low-resource Languages: An Exploration of Benefits and Trade-offs

Nick K Chibuye, Todd Rosenstock, Brian DeRenzi 

User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning

Máté Ákos Tündik, György Szaszák, Gábor Gosztolya, András Beke 

Punctuation Prediction Model for Conversational Speech

Piotr Żelasko, Piotr Szymański, Jan Mizgajski, Adrian Szymczak, Yishay Carmiel, Najim Dehak 

BUT OpenSAT 2017 Speech Recognition System

Martin Karafiát, Murali Karthick Baskar, Igor Szóke, Vladimír Malenovský, Karel Veselý, František Grézil, Lukáš Burget, Jan Černocký 

Visual Recognition of Continuous Cued Speech Using a Tandem CNN-HMM Approach

Li Liu, Thomas Hueber, Gang Feng, Denis Beaudet 

Building Large-vocabulary Speaker-independent Lipreading Systems

Kwanchiva Thangthai, Richard Harvey 

CRIM's System for the MGB-3 English Multi-Genre Broadcast Media Transcription

Vishwa Gupta, Gilles Boullianne 

Sampling Strategies in Siamese Networks for Unsupervised Speech Representation Learning

Rachid Riad, Corentin Dancette, Julien Karadayi, Neil Zeghidour, Thomas Schatz, Emmanuel Dupoux 

Compact Feedforward Sequential Memory Networks for Small-footprint Keyword Spotting

Mengzhe Chen, ShiLiang Zhang, Ming Lei, Yong Liu, Haitao Yao, Jie Gao 

Zero-resource Speech Recognition

Multilingual Bottleneck Features for Subword Modeling in Zero-resource Languages

Enno Hermann, Sharon Goldwater 

Exploiting Speaker and Phonetic Diversity of Mismatched Language Resources for Unsupervised Subword Modeling

Siyuan Feng, Tan Lee 

Unsupervised Word Segmentation from Speech with Attention

Pierre Godard, Marcelly Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, Laurent Besacier 

Learning Word Embeddings: Unsupervised Methods for Fixed-size Representations of Variable-length Speech Segments

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, Emmanuel Dupoux 

IPR2026-00272

Full Bayesian Hidden Markov Model Variational Autoencoder for Acoustic Unit Discovery

Thomas Glärner, Patrick Hanebrink, Janek Ebbers, Reinhold Haeb-Umbach 

Unspeech: Unsupervised Speech Context Embeddings

Benjamin Milde, Chris Biemann 

Spatial and Phase Cues for Source Separation and Speech Recognition

Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models

Yuan Gong, Christian Poellabauer 

Keyword Based Speaker Localization: Localizing a Target Speaker in a Multi-speaker Environment

Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr 

End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction

Zhong-Qiu Wang, Jonathan Le Roux, DeLiang Wang, John Hershey 

PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation

Naoya Takahashi, Purvi Agrawal, Nabarun Goswami, Yuki Mitsufuji 

Integrating Spectral and Spatial Features for Multi-Channel Speaker Separation

Zhong-Qiu Wang, DeLiang Wang 

DNN Driven Speaker Independent Audio-Visual Mask Estimation for Speech Separation

Mandar Gogate, Ahsan Adeel, Ricard Marxer, Jon Barker, Amir Hussain 

Dialectal Variation

Exploring Temporal Reduction in Dialectal Spanish: A Large-scale Study of Lenition of Voiced Stops and Coda-s

Ioana Vasilescu, Nidia Hernandez, Bianca Vieru, Lori Lamel 

Dialect-geographical Acoustic-Tonetics: Five Disyllabic Tone Sandhi Patterns in Cognate Words from the Wu Dialects of Zhèjiāng Province

Phil Rose 

Regional Variation of /r/ in Swiss German Dialects

Adrian Leemann, Stephan Schmid, Dieter Studer-Joho, Marie-José Kolly 

Variation in the FACE Vowel across West Yorkshire: Implications for Forensic Speaker Comparisons

Kate Earnshaw, Erica Gold 

The 'West Yorkshire Regional English Database': Investigations into the Generalizability of Reference Populations for Forensic Speaker Comparison Casework

Erica Gold, Sula Ross, Kate Earnshaw 

Studying Vowel Variation in French-Algerian Arabic Code-switched Speech

Jane Wottawa, Amazouz Djegdjiga, Martine Adda-Decker, Lori Lamel 

Spoken Corpora and Annotation

Fearless Steps: Apollo-11 Corpus Advancements for Speech Technologies from Earth to the Moon

John H.L. Hansen, Abhijeet Sangwan, Aditya Joglekar, Ahmet E. Bulut, Lakshmish Kaushik, Chengzhu Yu 

A Knowledge Driven Structural Segmentation Approach for Play-Talk Classification During Autism Assessment

Manoj Kumar, Pooja Chebolu, So Hyun Kim, Kassandra Martinez, Catherine Lord, Shrikanth Narayanan 

An Open Source Emotional Speech Corpus for Human Robot Interaction Applications

Jesin James, Li Tian, Catherine Inez Watson 

Speech Database and Protocol Validation Using Waveform Entropy

Itshak Lapidot, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Jean-François Bonastre 

A French-Spanish Multimodal Speech Communication Corpus Incorporating Acoustic Data, Facial, Hands and Arms Gestures Information

Lucas D. Terissi, Gonzalo Sad, Mauricio Cerda, Slim Ouni, Rodrigo Galvez, Juan C. Gómez, Bernard Girau, Nancy Hitschfeld-Kahler 

L2-ARCTIC: A Non-native English Speech Corpus

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, Ricardo Gutierrez-Osuna 

The First DIHARD Speech Diarization Challenge

ZCU-NTIS Speaker Diarization System for the DIHARD 2018 Challenge

Zbyněk Zajíč, Marie Kunešová, Jan Zelinka, Marek Hruží 

IPR2026-00272

Speaker Diarization with Enhancing Speech for the First DIHARD Challenge

Lei Sun, Jun Du, Chao Jiang, Xueyang Zhang, Shan He, Bing Yin, Chin-Hui Lee 

BUT System for DIHARD Speech Diarization Challenge 2018

Mireia Diez, Federico Landini, Lukáš Burget, Johan Rohdin, Anna Silnova, Kateřina Žmolíková, Ondřej Novotný, Karel Veselý, Ondřej Glembek, Oldřich Plchot, Ladislav Mošner, Pavel Matějka 

Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge.

Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel, Eduardo Lleida 

Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villaiba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, Sanjeev Khudanpur 

The EURECOM Submission to the First DIHARD Challenge

Jose Patino, Héctor Delgado, Nicholas Evans 

Joint Discriminative Embedding Learning, Speech Activity and Overlap Detection for the DIHARD Speaker Diarization Challenge

Valter Akira Miasato Filho, Diego Augusto Silva, Luis Gustavo Depra Cuzzo 

Text Analysis, Multilingual Issues and Evaluation in Speech Synthesis

Multilingual Grapheme-to-Phoneme Conversion with Global Character Vectors

Jinfu Ni, Yoshinori Shiga, Hisashi Kawai 

A Hybrid Approach to Grapheme to Phoneme Conversion in Assamese

Somnath Roy, Shakuntala Mahanta 

Investigation of Using Disentangled and Interpretable Representations for One-shot Cross-lingual Voice Conversion

Seyed Hamidreza Mohammadi, Taehwan Kim 

Using Pupillometry to Measure the Cognitive Load of Synthetic Speech

Avashna Govender, Simon King 

Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm

Avashna Govender, Simon King 

Attentive Sequence-to-Sequence Learning for Diacritic Restoration of Yorùbá Language Text

Iroko Orife 

Gated Convolutional Neural Network for Sentence Matching

Peixin Chen, Wu Guo, Zhi Chen, Jian Sun, Lanhua You 

On Training and Evaluation of Grapheme-to-Phoneme Mappings with Limited Data

Dravyansh Sharma 

The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech

Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, Björn Schuller 

Word Emphasis Prediction for Expressive Text to Speech

Yosi Mass, Slava Shechtman, Moran Mordechai, Ron Hoory, Oren Sar Shalom, Guy Lev, David Konopnicki 

A Comparison of Speaker-based and Utterance-based Data Selection for Text-to-Speech Synthesis

Kai-Zhan Lee, Erica Cooper, Julia Hirschberg 

Data Requirements, Selection and Augmentation for DNN-based Speech Synthesis from Crowdsourced Data

Markus Toman, Geoffrey S. Meltzner, Rupal Patel 

Neural Network Training Strategies for ASR

Lightly Supervised vs. Semi-supervised Training of Acoustic Model on Luxembourgish for Low-resource Automatic Speech Recognition

Karel Veselý, Carlos Segura, Igor Szöke, Jordi Luque, Jan Černocký 

Investigation on the Combination of Batch Normalization and Dropout in BLSTM-based Acoustic Modeling for ASR

Li Wenjie, Gaofeng Cheng, Fengpei Ge, Pengyuan Zhang, Yonghong Yan 

Inference-Invariant Transformation of Batch Normalization for Domain Adaptation of Acoustic Models

Masayuki Suzuki, Tohru Nagano, Gakuto Kurata, Samuel Thomas 

Active Learning for LF-MMI Trained Neural Networks in ASR

Yanhua Long, Hong Ye, Yijie Li, Jiaen Liang 

An Investigation of Mixup Training Strategies for Acoustic Models in ASR

Ivan Medennikov, Yuri Khokhlov, Aleksei Romanenko, Dmitry Popov, Natalia Tomashenko, Ivan Sorokin, Alexander Zatvornitskiy 

Comparison of Unsupervised Modulation Filter Learning Methods for ASR

Purvi Agrawal, Sriram Ganapathy 

Improved Training for Online End-to-end Speech Recognition Systems

Suyoun Kim, Michael Seltzer, Jinyu Li, Rui Zhao 

Combining Natural Gradient with Hessian Free Methods for Sequence Training

Adnan Haider, Philip Woodland 

Lattice-free State-level Minimum Bayes Risk Training of Acoustic Models

Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu 

A Study of Enhancement, Augmentation and Autoencoder Methods for Domain Adaptation in Distant Speech Recognition

Hao Tang, Wei-Ning Hsu, François Grondin, James Glass 

Multilingual Deep Neural Network Training Using Cyclical Learning Rate

Andreas Søeborg Kirkedal, Yeon-Jun Kim 

Application of ASR in Medical Practice

Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus

Jianwei Yu, Xurong Xie, Shansong Liu, Shoukang Hu, Max W. Y. Lam, Xixin Wu, Ka Ho Wong, Xunying Liu, Helen Meng 

Automatic Evaluation of Speech Intelligibility Based on I-vectors in the Context of Head and Neck Cancers

Imed Laaridh, Corinne Fredouille, Alain Ghio, Muriel Lalain, Virginie Woisard 

Dysarthric Speech Recognition Using Convolutional LSTM Neural Network

Myungjong Kim, Beiming Cao, Kwanghoon An, Jun Wang 

Perceptual and Automatic Evaluations of the Intelligibility of Speech Degraded by Noise Induced Hearing Loss Simulation

Imed Laaridh, Julien Tardieu, Cynthia Magnen, Pascal Gaillard, Jérôme Farinas, Julien Pinquier 

Articulatory Features for ASR of Pathological Speech

Emre Yilmaz, Vikramjit Mitra, Chris Bartels, Horacio Franco 

Mining Multimodal Repositories for Speech Affecting Diseases

Joana Correia, Bhiksha Raj, Isabel Trancoso, Francisco Teixeira 

Long Distance Voice Channel Diagnosis Using Deep Neural Networks

Zhen Qin, Tom Ko, Guangjian Tian 

Speech Recognition for Medical Conversations

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, Xuedong Zhang 

Source and Supra-segmentals

Prosodic Focus Acquisition in French Early Cochlear Implanted Children

Chadi Farah, Stephane Roman, Mariapaola D'Imperio 

The Role of Temporal Variation in Narrative Organization

Nassima Fezza 

Interaction Mechanisms between Glottal Source and Vocal Tract in Pitch Glides

Tiina Murtola, Jarmo Malinen 

Relating Articulatory Motions in Different Speaking Rates

Astha Singh, G. Nisha Meenakshi, Prasanta Kumar Ghosh 

Estimation of the Asymmetry Parameter of the Glottal Flow Waveform Using the Electroglottographic Signal

João Cabral 

Classification of Disorders in Vocal Folds Using Electroglottographic Signal

Tanumay Mandal, K. Sreenivasa Rao, Sanjay Kumar Gupta 

Automatic Glottis Localization and Segmentation in Stroboscopic Videos Using Deep Neural Network

Achuth Rao MV, Rahul Krishnamurthy, Pebbili Gopikishore, Veeramani Priyadharshini, Prasanta Kumar Ghosh 

Respiratory and Respiratory Muscular Control in JL1's and JL2's Text Reading Utilizing 4-RSTs and a Soft Respiratory Mask with a Two-Way Bulb

Toshiko Isei-Jaakkola, Keiko Ochi, Keikichi Hirose 

A Preliminary Study on Tonal Coarticulation in Continuous Speech

Lixia Hao, Wei Zhang, Yanlu Xie, Jinsong Zhang 

Plenary Talk-3

Speech and Language Processing for Learning and Wellbeing

Helen Meng 

IPR2026-00272

Distant ASR

Far-Field Speech Recognition Using Multivariate Autoregressive Models

Sriram Ganapathy, Madhumita Harish 

Efficient Implementation of the Room Simulator for Training Deep Neural Network Acoustic Models

Chanwoo Kim, Ehsan Variani, Arun Narayanan, Michiel Bacchiani 

Stream Attention for Distributed Multi-Microphone Speech Recognition

Xiaofei Wang, Ruizhi Li, Hynek Hermansky 

Recognizing Overlapped Speech in Meetings: A Multichannel Separation Approach Using Neural Networks

Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, Fil Allela 

Integrating Neural Network Based Beamforming and Weighted Prediction Error Dereverberation

Lukas Drude, Christoph Boeddeker, Jahn Heymann, Reinhold Haeb-Umbach, Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani 

A Probability Weighted Beamformer for Noise Robust ASR

Suliang Bu, Yunxin Zhao, Meiyuh Hwang, Sining Sun 

Expressive Speech Synthesis

Effects of Dimensional Input on Paralinguistic Information Perceived from Synthesized Dialogue Speech with Neural Network

Masaki Yokoyama, Tomohiro Nagata, Hiroki Mori 

Neural MultiVoice Models for Expressing Novel Personalities in Dialog

Shereen Oraby, Lena Reed, Sharath T.S., Shubhangi Tandon, Marilyn Walker 

Expressive Speech Synthesis Using Sentiment Embeddings

Igor Jauk, Jaime Lorenzo-Trueba, Junichi Yamagishi, Antonio Bonafonte 

Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder

Kei Akuzawa, Yusuke Iwasawa, Yutaka Matsuo 

Rapid Style Adaptation Using Residual Error Embedding for Expressive Speech Synthesis

Xixin Wu, Yuewen Cao, Mu Wang, Songxiang Liu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, Helen Meng 

EMPHASIS: An Emotional Phoneme-based Acoustic Model for Speech Synthesis System

Hao Li, Yongguo Kang, Zhenyu Wang 

Representation Learning for Emotion

Bags in Bag: Generating Context-Aware Bags for Tracking Emotions from Speech

Jing Han, Zixing Zhang, Maximilian Schmitt, Zhao Ren, Fabien Ringeval, Björn Schuller 

An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition

Pengcheng Li, Yan Song, Ian McLoughlin, Wu Guo, Lirong Dai 

Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks

Zixiaofan Yang, Julia Hirschberg 

Emotion Identification from Raw Speech Signals Using DNNs

Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, Najim Dehak 

Encoding Individual Acoustic Features Using Dyad-Augmented Deep Variational Representations for Dialog-level Emotion Recognition

Jeng-Lin Li, Chi-Chun Lee 

Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study

Siddique Latif, Rajib Rana, Junaid Qadir, Julien Epps 

Articulatory Information, Modeling and Inversion

Phoneme-to-Articulatory Mapping Using Bidirectional Gated RNN

Théo Biasutto-Lervat, Slim Ouni 

Tongue Segmentation with Geometrically Constrained Snake Model

Zhihua Su, Jianguo Wei, Qiang Fang, Jianrong Wang, Kiyoshi Honda 

Low Resource Acoustic-to-articulatory Inversion Using Bi-directional Long Short Term Memory

Aravind Illa, Prasanta Kumar Ghosh 

Automatic Visual Augmentation for Concatenation Based Synthesized Articulatory Videos from Real-time MRI Data for Spoken Language Training

Chandana S, Chiranjeevi Yarra, Ritu Aggarwal, Sanjeev Kumar Mittal, Kausthubha N K, Raseena K T, Astha Singh, Prasanta Kumar Ghosh 

Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video Using Semantic Segmentation with Fully Convolutional Networks

Valliappan CA, Renuka Mannem, Prasanta Kumar Ghosh 

Noise Robust Acoustic to Articulatory Speech Inversion

Nadee Seneviratne, Ganesh Sivaraman, Vikramjit Mitra, Carol Espy-Wilson 

Novel Paradigms for Direct Synthesis Based on Speech-Related Biosignals

Designing a Pneumatic Bionic Voice Prosthesis - A Statistical Approach for Source Excitation Generation

Farzaneh Ahmadi, Tomoki Toda 

A Neural Model to Predict Parameters for a Generalized Command Response Model of Intonation

Bastian Schnell, Philip N. Garner 

Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information

Beiming Cao, Myungjong Kim, Jun R. Wang, Jan van Santen, Ted Mau, Jun Wang 

Effectiveness of Generative Adversarial Network for Non-Audible Murmur-to-Whisper Speech Conversion

Neil Shah, Nirmesh Shah, Hemant Patil 

Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion

Lorenz Diener, Tanja Schultz 

Domain-Adversarial Training for Session Independent EMG-based Speech Recognition

Michael Wand, Tanja Schultz, Jürgen Schmidhuber 

Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces

László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, Tamás Gábor Csapó 

Low Resource Speech Recognition Challenge for Indian Languages

Transcription Correction for Indian Languages Using Acoustic Signatures

Jeena JPrakash, Golda Brunet Rajan, Hema Murthy 

BUT System for Low Resource Indian Language ASR

Bhargav Pulugundla, Murali Karthick Baskar, Santosh Kesiraju, Ekaterina Egorova, Martin Karafiát, Lukáš Burget, Jan Černocký 

DA-ICT/IIITV System for Low Resource Speech Recognition Challenge 2018

Hardik B. Sailor, Maddala Venkata Siva Krishna, Diksha Chhabra, Ankur T. Patil, Madhu Kamble, Hemant Patil 

An Exploration towards Joint Acoustic Modeling for Indian Languages: IIIT-H Submission for Low Resource Speech Recognition Challenge for Indian Languages, INTERSPEECH 2018

Hari Krishna, Krishna Gurugubelli, Vishnu Vidyadhara Raju V, Anil Kumar Vuppala 

TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages

Noor Fathima, Tanvina Patel, Mahima C, Anuroop Iyengar 

Articulatory and Stacked Bottleneck Features for Low Resource Speech Recognition

Vishwas M. Shetty, Rini A Sharon, Basil Abraham, Tejaswi Seeram, Anusha Prakash, Nithya Ravi, S. Umesh 

ISI ASR System for the Low Resource Speech Recognition Challenge for Indian Languages

Jayadev Billa 

Show and Tell 7

An Automated Assistant for Medical Scribes

Gregory Finley, Erik Edwards, Amanda Robinson, Najmeh Sadoughi, James Fone, Mark Miller, David Suendermann-Oeft, Michael Brenndoerfer, Nico Axtmann 

AGROASSAM: A Web Based Assamese Speech Recognition Application for Retrieving Agricultural Commodity Price and Weather Information

Abhishek Dey, Abhash Deka, Siddika Imani, Barsha Deka, Rohit Sinha, S R Mahadeva Prasanna, Priyankoo Sarmah, K Samudravijaya, Nirmala S.R. 

Voice-powered Solutions with Cloud AI

Dan Aharon 

Speech Synthesis in the Wild

Ganesh Sivaraman, Parav Nagarsheth, Elie Khoury 

Deep Noise Tracking Network: A Hybrid Signal Processing/Deep Learning Approach to Speech Enhancement

Shuai Nie, Shan Liang, Bin Liu, Yaping Zhang, Wenju Liu, Jianhua Tao 

A Deep Neural Network Based Harmonic Noise Model for Speech Enhancement

Zhiheng Ouyang, Hongjiang Yu, Wei-Ping Zhu, Benoit Champagne 

A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement

Ke Tan, DeLiang Wang 

All-Neural Multi-Channel Speech Enhancement

Zhong-Qiu Wang, DeLiang Wang 

Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios

Hao Zhang, DeLiang Wang 

The Conversation: Deep Audio-Visual Speech Enhancement

Triantafyllos Afouras, Joon Son Chung, Andrew Senior 

Student-Teacher Learning for BLSTM Mask-based Speech Enhancement

Aswin Shanmugam Subramanian, Szu-Jui Chen, Shinji Watanabe 

Speech Enhancement Using Deep Mixture of Experts Based on Hard Expectation Maximization

Pavan Karjol, Prasanta Kumar Ghosh 

Adversarial Feature-Mapping for Speech Enhancement

Zhong Meng, Jinyu Li, Yifan Gong, Biing-Hwang (Fred) Juang 

Biophysically-inspired Features Improve the Generalizability of Neural Network-based Speech Enhancement Systems

Deepak Baby, Sarah Verhulst 

Error Modeling via Asymmetric Laplace Distribution for Deep Neural Network Based Single-Channel Speech Enhancement

Li Chai, Jun Du, Chin-Hui Lee 

A Priori SNR Estimation Based on a Recurrent Neural Network for Robust Speech Enhancement

Yangyang Xia, Richard Stern 

Acoustic Scenes and Rare Events

Multiple Instance Deep Learning for Weakly Supervised Small-Footprint Audio Event Detection

Shao-Yen Tseng, Juncheng Li, Yun Wang, Florian Metze, Joseph Szurley, Samarjit Das 

Unsupervised Temporal Feature Learning Based on Sparse Coding Embedded BoAW for Acoustic Event Recognition

Liwen Zhang, Jiqing Han, Shiwen Deng 

Data Independent Sequence Augmentation Method for Acoustic Scene Classification

Zhang Teng, Kailai Zhang, Ji Wu 

A Compact and Discriminative Feature Based on Auditory Summary Statistics for Acoustic Scene Classification

Hongwei Song, Jiqing Han, Shiwen Deng 

ASE: Acoustic Scene Embedding Using Deep Archetypal Analysis and GMM

Pulkit Sharma, Vinayak Abrol, Anshul Thakur 

Deep Convolutional Neural Network with Scalogram for Audio Scene Modeling

Hangting Chen, Pengyuan Zhang, Haichuan Bai, Qingsheng Yuan, Xiuguo Bao, Yonghong Yan 

Time Aggregation Operators for Multi-label Audio Event Detection

Pankaj Joshi, Digvijaysingh Gautam, Ganesh Ramakrishnan, Preethi Jyothi 

Early Detection of Continuous and Partial Audio Events Using CNN

Ian McLoughlin, Yan Song, Lam Dang Pham, Ramaswamy Palaniappan, Huy Phan, Yue Lang 

Robust Acoustic Event Classification Using Bag-of-Visual-Words

Manjunath Mulimani, Shashidhar G Koolagudi 

Wavelet Transform Based Mel-scaled Features for Acoustic Scene Classification

Shelfali Waldekar, Goutam Saha 

Multi-modal Attention Mechanisms in LSTM and Its Application to Acoustic Scene Classification

Teng Zhang, Kailai Zhang, Ji Wu 

Language Modeling

Contextual Language Model Adaptation for Conversational Agents

Anirudh Raju, Behnam Hedayatnia, Linda Liu, Ankur Gandhe, Chandra Khatri, Angeliki Metallinou, Anu Venkatesh, Ariya Rastrow 

Active Memory Networks for Language Modeling

IPR2026-00272

Oscar Chen, Anton Ragni, Mark Gales, Xie Chen 

Unsupervised and Efficient Vocabulary Expansion for Recurrent Neural Network Language Models in ASR

Yerbolat Khassanov, Eng Siong Chng 

Improving Language Modeling with an Adversarial Critic for Automatic Speech Recognition

Yike Zhang, Pengyuan Zhang, Yonghong Yan 

Training Recurrent Neural Network through Moment Matching for NLP Applications

Yue Deng, Yilin Shen, KaWai Chen, Hongxia Jin 

Investigation on LSTM Recurrent N-gram Language Models for Speech Recognition

Zoltán Tüske, Ralf Schlüter, Hermann Ney 

Online Incremental Learning for Speaker-Adaptive Language Models

Chih Chi Hu, Bing Liu, John Shen, Ian Lane 

Efficient Language Model Adaptation with Noise Contrastive Estimation and Kullback-Leibler Regularization

Jesús Andrés-Ferrer, Nathan Bodendstab, Paul Vozila 

Recurrent Neural Network Language Model Adaptation for Conversational Speech Recognition

Ke Li, Hainan Xu, Yiming Wang, Daniel Povey, Sanjeev Khudanpur 

What to Expect from Expected Kneser-Ney Smoothing

Michael Levit, Sarangarajan Parthasarathy, Shuangyu Chang 

i-Vectors in Language Modeling: An Efficient Way of Domain Adaptation for Feed-Forward Models

Karel Beneš, Santosh Kesiraju, Lukáš Burget 

Speech Pathology, Depression, and Medical Applications

How Did You like 2017? Detection of Language Markers of Depression and Narcissism in Personal Narratives

Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, Harald Baumeister 

Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions

Zhaocheng Huang, Julien Epps, Dale Joachim, Michael Chen 

Multi-Lingual Depression-Level Assessment from Conversational Speech Using Acoustic and Text Features

Yasin Özkanca, Cenk Demiroglu, Aslı Besirli, Selime Celik 

Dysarthric Speech Classification Using Glottal Features Computed from Non-words, Words and Sentences

Narendra N P, Paavo Alku 

Identifying Schizophrenia Based on Temporal Parameters in Spontaneous Speech

Gábor Gosztolya, Anita Bagi, Szilvia Szalóki, István Szendi, Ildikó Hoffmann 

Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla, Zhuohao Chen, Nikolaos Flemotomos, James Gibson, Dogan Can, David Atkins, Shrikanth Narayanan 

Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning

Ying Qin, Tan Lee, Siyuan Feng, Anthony Pak Hin Kong 

Towards an Unsupervised Entrainment Distance in Conversational Speech Using Deep Neural Networks

Md Nasir, Brian Baucum, Shrikanth Narayanan, Panayiotis Georgiou 

Patient Privacy in Paralinguistic Tasks

Francisco Teixeira, Alberto Abad, Isabel Trancoso 

A Lightly Supervised Approach to Detect Stuttering in Children's Speech

Sadeen Alharbi, Madina Hasan, Anthony J H Simons, Shelagh Brumfitt, Phil Green 

Learning Conditional Acoustic Latent Representation with Gender and Age Attributes for Automatic Pain Level Recognition

Jeng-Lin Li, Yi-Ming Weng, Chip-Jin Ng, Chi-Chun Lee 

Perspective Talk-4

Speaker and Language Recognition -- From Laboratory Technologies to the Wild

Sriram Ganapathy 

Spoken Language Understanding

A Deep Reinforcement Learning Based Multimodal Coaching Model (DCM) for Slot Filling in Spoken Language Understanding(SLU)

Yu Wang, Abhishek Patel, Yilin Shen, Hongxia Jin 

IPR2026-00272

Krisp Technologies EX1044 Page 118

Is ATIS Too Shallow to Go Deeper for Benchmarking Spoken Language Understanding Models?

Frédéric B  chet, Christian Raymond 

Robust Spoken Language Understanding via Paraphrasing

Avik Ray, Yilin Shen, Hongxia Jin 

Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, Hung-yi Lee 

User Information Augmented Semantic Frame Parsing Using Progressive Neural Networks

Yilin Shen, Xiangyu Zeng, Yu Wang, Hongxia Jin 

An Efficient Approach to Encoding Context for Spoken Language Understanding

Raghav Gupta, Abhinav Rastogi, Dilek Hakkani-T  r 

Source Separation from Monaural Input

Deep Speech Denoising with Vector Space Projections

Jeffrey Hetherly, Paul Gamble, Maria Alejandra Barrios, Cory Stephenson, Karl Ni 

A Shifted Delta Coefficient Objective for Monaural Speech Separation Using Multi-task Learning

Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li 

A Two-Stage Approach to Noisy Cochannel Speech Separation with Gated Residual Networks

Ke Tan, DeLiang Wang 

Monoaural Audio Source Separation Using Variational Autoencoders

Laxmi Pandey, Anurendra Kumar, Vinay Namboodiri 

Towards Automated Single Channel Source Separation Using Neural Networks

Arpita Gang, Pravesh Blyani, Akshay Soni 

Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech-Background Separation

Hakan Erdogan, Takuya Yoshioka 

Multimodal Systems

Annotator Trustability-based Cooperative Learning Solutions for Intelligent Audio Analysis

Simone Hantke, Christoph Stemp, Bj  rn Schuller 

Semi-supervised Cross-domain Visual Feature Learning for Audio-Visual Broadcast Speech Transcription

Rongfeng Su, Xunying Liu, Lan Wang 

Deep Lip Reading: A Comparison of Models and an Online Application

Triantafyllos Afouras, Joon Son Chung, Andrew Senior 

Iterative Learning of Speech Recognition Models for Air Traffic Control

Ajay Srinivasamurthy, Petr Moticek, Mittul Singh, Youssef Oualil, Matthias Kleinert, Heiko Ehr, Hartmut Helmke 

Speaker Adaptive Audio-Visual Fusion for the Open-Vocabulary Section of AVICAR

Leda Sari, Mark Hasegawa-Johnson, Kumaran S, Georg Stemmer, Krishnakumar N Nair 

Multimodal Name Recognition in Live TV Subtitling

Marek H  r  z, Ale   Pra    k, Michal Bu    a 

Coding

Dithered Quantization for Frequency-Domain Speech and Audio Coding

Tom B  ckstr  m, Johannes Fischer, Sneha Das 

Postfiltering with Complex Spectral Correlations for Speech and Audio Coding

Sneha Das, Tom B  ckstr  m 

Postfiltering Using Log-Magnitude Spectrum for Speech and Audio Coding

Sneha Das, Tom B  ckstr  m 

Temporal Noise Shaping with Companding

Arijit Biswas, Per Hedelin, Lars Villemoes, Vinay Melkote 

Multi-frame Quantization of LSF Parameters Using a Deep Autoencoder and Pyramid Vector Quantizer

Yaxing Li, Eshete Derb Emiru, Shengwu Xiong, Anna Zhu, Pengfei Duan, Yichang Li 

Multi-frame Coding of LSF Parameters Using Block-Constrained Trellis Coded Vector Quantization

Yaxing Li, Shan Xu, Shengwu Xiong, Anna Zhu, Pengfei Duan, Yueming Ding 

Speaker Verification Using Neural Network Methods II

Training Utterance-level Embedding Networks for Speaker Identification and Verification

Heewoong Park, Sukhyun Cho, Kyubyong Park, Namju Kim, Jonghun Park 

Analysis of Complementary Information Sources in the Speaker Embeddings Framework

Maresh Kumar Nandwana, Mitchell McLaren, Diego Castan, Julien van Hout, Aaron Lawson 

Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification

Yingke Zhu, Tom Ko, David Snyder, Brian Mak, Daniel Povey 

An Improved Deep Embedding Learning Method for Short Duration Speaker Verification

Zhifu Gao, Yan Song, Ian McLoughlin, Wu Guo, Lirong Dai 

Avoiding Speaker Overfitting in End-to-End DNNs Using Raw Waveform for Text-Independent Speaker Verification

Jee-weon Jung, Hee-soo Heo, IL-ho Yang, Hye-jin Shim, Ha-jin Yu 

Deeply Fused Speaker Embeddings for Text-Independent Speaker Verification

Gautam Bhattacharya, Md Jahangir Alam, Vishwa Gupta, Patrick Kenny 

Employing Phonetic Information in DNN Speaker Embeddings to Improve Speaker Recognition Performance

Md Hafizur Rahman, Ivan Himawan, Mitchell McLaren, Clinton Fookes, Sridha Sridharan 

End-to-end Text-dependent Speaker Verification Using Novel Distance Measures

Subhadeep Dey, Srikanth Madikeri, Petr Motlicek 

Robust Speaker Clustering using Mixtures of von Mises-Fisher Distributions for Naturalistic Audio Streams

Harishchandra Dubey, Abhijeet Sangwan, John H.L. Hansen 

Triplet Network with Attention for Speaker Diarization

Huan Song, Megan Willi, Jayaraman J. Thiagarajan, Visar Berisha, Andreas Spanias 

I-vector Transformation Using Conditional Generative Adversarial Networks for Short Utterance Speaker Verification

Jiacen Zhang, Nakamasa Inoue, Koichi Shinoda 

Analysis of Length Normalization in End-to-End Speaker Verification System

Weicheng Cai, Jinkun Chen, Ming Li 

Angular Softmax for Short-Duration Text-independent Speaker Verification

Zili Huang, Shuai Wang, Kai Yu 

An End-to-End Text-Independent Speaker Identification System on Short Utterances

Ruifang Ji, Xinyuan Cai, Xu Bo 

MTGAN: Speaker Verification through Multitasking Triplet Generative Adversarial Networks

Wenhao Ding, Liang He 

Emotion Recognition and Analysis

Categorical vs Dimensional Perception of Italian Emotional Speech

Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Alice Baird, Björn Schuller 

A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech

Xingfeng Li, Masato Akagi 

Cross-lingual Speech Emotion Recognition through Factor Analysis

Brecht Desplanques, Kris Demuyne 

Modeling Self-Reported and Observed Affect from Speech

Jian Cheng, Jared Bernstein, Elizabeth Rosenfeld, Peter W. Foltz, Alex S. Cohen, Terje B. Holmlund, Brita Elvevåg 

Stochastic Shake-Shake Regularization for Affective Learning from Speech

Che-Wei Huang, Shrikanth Narayanan 

Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition

Anderson R. Avila, Md Jahangir Alam, Douglas O'Shaughnessy, Tiago Falk 

Demonstrating and Modelling Systematic Time-varying Annotator Disagreement in Continuous Emotion Annotation

Mia Atcheson, Vidhyasaharan Sethu, Julien Epps 

Speech Emotion Recognition from Variable-Length Inputs with Triplet Loss Function

Jian Huang, Ya Li, Jianhua Tao, Zhen Lian 

Imbalance Learning-based Framework for Fear Recognition in the MediaEval Emotional Impact of Movies Task

Xiaotong Zhang, Xingliang Cheng, Mingxing Xu, Thomas Fang Zheng 

Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms

IPR2026-00272

Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, Lianhong Cai 

Speech Emotion Recognition Using Spectrogram & Phoneme Embedding

Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, Jithendra Vepa 

On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks

Saurabh Sahu, Rahul Gupta, Carol Espy-Wilson 

Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes

Srinivas Parthasarathy, Carlos Busso 

Acoustic Modelling

Knowledge Distillation for Sequence Model

Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, Kai Yu 

Improving CTC-based Acoustic Model with Very Deep Residual Time-delay Neural Networks

Sheng Li, Xugang Lu, Ryoichi Takashima, Peng Shen, Tatsuya Kawahara, Hisashi Kawai 

Filter Sampling and Combination CNN (FSC-CNN): A Compact CNN Model for Small-footprint ASR Acoustic Modeling Using Raw Waveforms

Jinxi Guo, Ning Xu, Xin Chen, Yang Shi, Kaiyuan Xu, Abeer Alwan 

Twin Regularization for Online Speech Recognition

Mirco Ravanelli, Dmitriy Serdyuk, Yoshua Bengio 

Self-Attentional Acoustic Models

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, Alex Waibel 

Hierarchical Recurrent Neural Networks for Acoustic Modeling

Jinhwan Park, Iksoo Choi, Yoonho Boo, Wonyong Sung 

Dictionary Augmented Sequence-to-Sequence Neural Network for Grapheme to Phoneme Prediction

Antoine Bruguier, Anton Bakhtin, Dravyansh Sharma 

Leveraging Second-Order Log-Linear Model for Improved Deep Learning Based ASR Performance

Ankit Raj, Shakti P Rath, Jithendra Vepa 

Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, Sanjeev Khudanpur 

Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings

Da-Rong Liu, Kuan-Yu Chen, Hung-yi Lee, Lin-shan Lee 

Phone Recognition Using a Non-Linear Manifold with Broad Phone Class Dependent DNNs

Mengjie Qian, Linxue Bai, Peter Jančovič, Martin Russell 

A Multi-Discriminator CycleGAN for Unsupervised Non-Parallel Speech Domain Adaptation

Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, Richard Socher 

Speech and Speaker Perception

Interactions between Vowels and Nasal Codas in Mandarin Speakers' Perception of Nasal Finals

Chong Cao, Wei Wei, Wei Wang, Yanlu Xie, Jinsong Zhang 

Weighting Pitch Contour and Loudness Contour in Mandarin Tone Perception in Cochlear Implant Listeners

Qinglin Meng, Nengheng Zheng, Ambika Prasad Mishra, Jacinta Dan Luo, Jan W. H. Schnupp 

Implementing DIANA to Model Isolated Auditory Word Recognition in English

Filip Nenadić, Louis ten Bosch, Benjamin V. Tucker 

Effects of Homophone Density on Spoken Word Recognition in Mandarin Chinese

Bhamini Sharma 

Visual Timing Information in Audiovisual Speech Perception: Evidence from Lexical Tone Contour

Hui Xie, Biao Zeng, Rui Wang 

COSMO SylPhon: A Bayesian Perceptuo-motor Model to Assess Phonological Learning

Marie-Lou Barnaud, Julien Diard, Pierre Bessière, Jean-Luc Schwartz 

Experience-dependent Influence of Music and Language on Lexical Pitch Learning Is Not Additive

Akshay Raj Maggu, Patrick C. M. Wong, Hanjun Liu, Francis C. K. Wong 

Influences of Fundamental Oscillation on Speaker Identification in Vocalic Utterances by Humans and Computers

Volker Dellwo, Thayabaran Kathiresan, Elisa Pellegrino, Lei He, Sandra Schwab, Dieter Maurer 