

COMMUNICATION ACOUSTICS

AN INTRODUCTION TO SPEECH, AUDIO AND PSYCHOACOUSTICS

Ville Pulkki and Matti Karjalainen

Aalto University, Finland

WILEY

This edition first published 2015
© 2015 John Wiley & Sons, Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

Library of Congress Cataloging-in-Publication Data

Pulkki, Ville.

Communication acoustics : an introduction to speech, audio, and psychoacoustics / Ville Pulkki and Matti Karjalainen.

pages cm

Includes index.

ISBN 978-1-118-86654-2 (hardback)

1. Sound—Recording and reproducing. 2. Hearing. 3. Psychoacoustics. 4. Sound. I. Karjalainen, Matti, 1946-2010.

II. Title.

TK7881.4.P85 2015

620.2—dc23

2014030757

A catalogue record for this book is available from the British Library.

Set in 10/12pt Times LT Std by SPi Publisher Services, Pondicherry, India

Printed in Singapore by C.O.S. Printers Pte Ltd

1 2015

To Vappu, Sampo, Raisa, and HUT people
To Sari, Veikko and Sannakaisa, and Aalto Acoustics people

Eight kHz, LLC
Exhibit 2011
IPR2023-01004
Page 2

16

Speech Technologies

Speech technology covers applications such as the recognition or synthesis of speech, speaker recognition, and optimized coding and enhancement of speech. Major research efforts have been invested in speech technology both in academia and in industry, which have led to several breakthroughs. After several decades of research, coding, synthesis, and recognition of speech are used extensively in several applications. The widespread adoption of speech technologies has been enabled by the advent of sufficiently powerful and relatively inexpensive digital processors.

Many user interfaces based on speech exist. A computer may take commands by recording the speech of the user, thus requiring *speech recognition* abilities, and it may deliver messages to the user by producing intelligible speech, which requires *speech synthesis* technology. This interaction was shown conceptually in Figure I.5 on page 5. In mobile communication, the goal is often to present speech at as low a bit rate as possible, which has led to many methods for *speech coding*, where the special characteristics of speech signals have been taken into account.

This chapter provides a very brief overview of different technologies in speech coding, synthesis, and recognition. The focus is very much on acoustics, signal processing, and audio, and the linguistic and statistical aspects are, in many places, treated superficially. Overall, the aim of this chapter is to give a general description of the main fields in speech technology and to guide interested readers to more comprehensive sources.

Speech enhancement is a wide area, which covers such topics as *error concealment*, *noise reduction*, *bandwidth extension*, and *echo control* in the context of speech communication (Vary and Martin, 2006). The goal of noise reduction and echo control should be clear to the reader. Bandwidth extension refers to techniques used to extend the bandwidth of transmitted narrowband speech to deliver better speech quality to the user. Error concealment refers to applications where the speech decoder is forced to ‘fill in’ missing data from the received speech signal. For example, when some frames of speech are lost in communication, the content of missing frames has to be guessed through some intelligent solutions to maximize speech quality. These techniques are not discussed in detail in this book.

spectral filtering are computed that are assumed to produce an output perceptually similar to the original segment when applied to the source-filter model in synthesis. This is done by first decomposing the speech data into vocoder parameters, and then estimating the Gaussian statistics (mean and variance) of the vocoder parameters and the duration for each state for all context-dependent phonemes. Using multi-stream HMM training (Gales and Young, 2008), the system can model the spectrum, excitation, and duration of each segment in a unified framework.

The parameters utilized in the source and filter depend on their actual implementation. For example, the spectral parameters may be represented as mel frequency cepstral coefficients (Fukada *et al.*, 1992) or as line spectral frequencies (Soong and Juang, 1984), and the fundamental frequency of the voiced excitation is often presented as the logarithm of f_0 . The computation of mel cepstra was discussed in Section 13.1.1. The logarithm of f_0 , in turn, corresponds to the frequency scale used in music (see Section 11.6.2). Commonly, a simple impulse train is used for exciting voiced speech, but the waveform of a voiced glottal excitation, inverse filtered from natural speech, may also be used to improve naturalness (Raitio *et al.*, 2011). The STRAIGHT method is nowadays commonly used as a speech vocoder (Kawahara, 2006), and in other speech processing besides data-based speech synthesis. In STRAIGHT synthesis, the source signal is generated using mixed excitation consisting of impulses and a noise component acting as the aperiodic component of voiced speech. Finally, the pitch-synchronous overlap add (PSOLA) method (Moulines and Charpentier, 1990) is used to reconstruct the excitation signal, which is then applied to excite the filter.

The synthesis part performs an operation that resembles the inverse of speech recognition (Zen *et al.*, 2009). A given word sequence is first converted into a sequence of context-dependent labels. After this, the HMM of an utterance is constructed by concatenating the context-dependent HMMs following the label sequence. Then, smooth sequences of spectral and excitation parameters are generated from the utterance HMMs using the mean and variance values of each state. Finally, a speech signal is synthesized using excitation generation and a speech synthesis filter.

With regard to the quality of speech obtained with statistical parametric synthesis, Zen *et al.* (2009) conclude with the words, 'Although even the proponents of statistical parametric synthesis feel that the best examples of unit-selection synthesis are better than the best examples of statistical parametric synthesis, overall it appears that the quality of statistical parametric synthesis has already reached a level where it can stand in its own right.' The speech quality obtained with statistical parametric synthesis is thus not on the same level with unit-selection synthesis, but better quality is obtained than with knowledge-based speech synthesis.

One benefit of statistical parametric synthesis is that the source-filter-type synthesis of speech signals provides possibilities for changing the characteristics of voice, such as speaking styles and prosodic features, and possibly even provides multilingual support (Zen *et al.*, 2009). With the unit-selection method, such modifications are a bit complicated, as voice conversion techniques have to be utilized (Stylianou *et al.*, 1998), which may degrade the quality of the speech.

However, although understandable speech is obtained with statistical parametric synthesis, its major drawback is still in the quality of speech. The speech is perceived as less authentic than with unit-selection synthesis, and several possible factors affecting this and various suggested refinements are reviewed by Zen *et al.* (2009).

16.3 Speech Recognition

The goal of *speech recognition* (Rabiner and Juang, 1993) is to capture the acoustic human speech signal and process the spoken language into text. It is one of the biggest challenges in speech technology. Evolution and culture have built speech into a fast and robust communication channel over adverse acoustic channels, tolerant to relatively high background noise levels. Speech is not only a rich container of information added to words and non-speech voices, it also carries the identity, age, and gender about the speaker and prosodic features carrying emotions and other meanings. Additionally, the speed of speech may vary substantially, and the vast number of languages, dialects, and speaking styles and disorders make the problem even more challenging.

In principle, speech recognizers should simply mimic brain functions to perform the task as well as humans do. Note that the goal of processing is different in brains and in speech recognizers. Brains turn speech into neural code and speech recognizers to written text. Speech recognition conducted by humans is based strongly on the high-level functions of the brain, such as language, awareness of environment, and assumptions about the content of speech based on previous utterances. Auditory models and models of the brain are clearly not at the level where such functions can be emulated. Thus, the most successful speech recognition techniques are based on generic principles of pattern recognition and data processing.

The complexity of a speech recognition task depends heavily on the definition of the task. The simplest task is to recognize speech from one known speaker uttering temporally separated words from a small, known vocabulary. The current methods perform well in such tasks. The task becomes harder when the vocabulary is made larger or not restricted at all, when words are not separated by silences, when multi-language speech is allowed, and when the content of speech is not limited to any specific topic. Even harder tasks are those involving multiple speakers, and when the level of background noise is high.

The speech recognition process can be divided into different phases. The first phase is to pre-process the speech signal to remove unnecessary redundancy and to describe the signal with *features*. Given the acoustic models of speech sounds, the statistical language models, and the lexicon of words, the next phase is a *pattern recognition* task where the decoder turns the sequence of features directly into the most likely sequence of words in the language.

The most common speech recognition systems are based on HMMs (Rabiner, 1989). The early techniques in speech recognition were dynamic time warping (Vintsyuk, 1968) and neural networks (Kohonen, 1988), but they were largely abandoned in the late 1990s. However, recently, deep neural networks have been proposed for use in speech recognition (Dahl *et al.*, 2012).

The main parts of a typical speech recognizer are shown in Figure 16.6. The speech input is turned into feature vector sequences first. After this, the probability of the vectors or vector sequences belonging to linguistic classes is computed using HMMs. The probability sequences are then decoded into text using a vocabulary and a model of language. Prior to using the recognizer, the HMM has to be trained with natural speech, and the vocabulary and model of language have to be constructed.

Speech recognizers commonly window the input signal at a rate of about 100 Hz and utilize window lengths of about 20 ms. The features computed from the signal segment are typically mel-frequency cepstral coefficients. Differentials between temporally subsequent features are also commonly utilized and have been found to improve the accuracy of recognition.

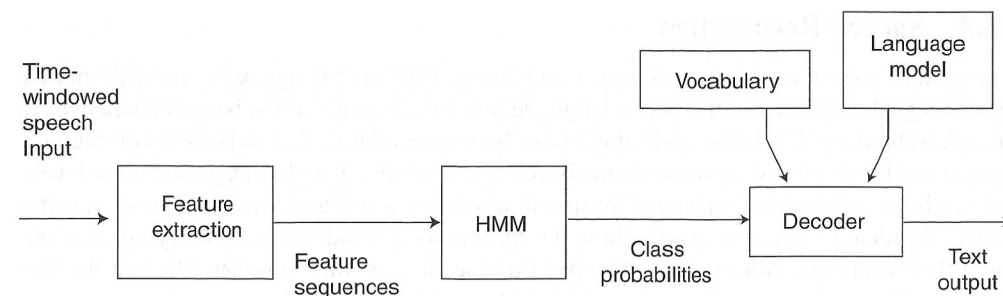


Figure 16.6 A simplified block diagram of speech recognition.

If the recognition task is limited to a small vocabulary and to isolated words, it is possible to construct separate HMMs for all the different words without sharing parameters between the models. The number of states in a word HMM is then determined with some heuristics, such as relating it to the number of phonemes or the average number of observations (Rabiner, 1989). As the size of the vocabulary increases, a word-level approach becomes infeasible due to the large number of states and parameters. Instead, a small set of segments, such as phones or diphones, is modelled with unique HMMs. The word models can then be formed by a simple concatenation of these units. A three-state, left-to-right HMM (Schwartz *et al.*, 1985) is commonly used for the model of a phoneme.

The sequences of probabilities of phonemes are then turned into written text using knowledge about the language they were spoken in. Some information about the language is already embedded in the HMMs of the units of speech, either in the form of word models or as the phone inventory. However, an extensive knowledge of language must be embedded in the recognizer. Most commonly, the grammatical constraints are learned as statistical dependences in the context of words from large text databases (Goodman, 2001). When the vocabulary is small, it may be feasible to make a list of allowed words and define their HMMs. However, when the size of the vocabulary is larger, acoustically similar words are hard to distinguish during recognition. Detecting boundaries of words in continuous speech is also difficult, which adds to the challenge. It becomes necessary to utilize either task-specific or general constraints to overcome these problems.

Speech recognition has already been adopted in many applications. For example, some televisions and handheld devices can already be controlled by voice, and telephone customer services may use speech recognition in automated tasks. Speech recognition is effective in browsing the contents of spoken speech databases (Chelba *et al.*, 2008). One application that is also actively being studied is a telephone service that translates the language of a speaker into another language. Although this service has not yet been achieved, translation in limited cases has been possible. Optimally, the characteristics of the voice of the original speaker should be preserved in the translated speech output. The biggest challenges are, however, in the processing of language, and not in the processing of speech signals.

Summary

Speech coding is perhaps the most mature of the speech technologies. Very good speech quality can be obtained with low bit rates when signal models of speech production mechanisms are utilized. The methods are relatively simple, and no large databases are required.

Speech synthesis and recognition techniques have also been developed fairly extensively. When compared with speech coding, a clear difference is that the systems must have access to large databases of natural language and/or speech in order to obtain good results. Speech synthesis and recognition thus require that the computer has knowledge of the recognized or synthesized language at some level, something that is not needed in speech coding. The performance of the synthesis and recognition techniques is still far from that of human performance, although multiple commercial applications already exist.

Further Reading

The synthesis and recognition of speech is described in more detail by Huang *et al.* (2001), Jurafsky and Martin (2008), and Saon and Chien (2012). A public domain toolkit for research in automatic speech recognition is available, and its documentation also serves as a good explanation of speech recognition (Young *et al.*, 2006). Hidden Markov models are explained in detail by Gales and Young (2008). The enhancement, coding, and error concealment in speech transmission are discussed by Vary and Martin (2006), and more information on audio coding can be found in Chu (2004).

References

- Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (1999) The AT&T next-gen TTS system. *Joint meeting of ASA, EAA, and DAGA*, pp. 18–24.
- Chelba, C., Hazen, T.J., and Saraçlar, M. (2008) Retrieval and browsing of spoken content. *Signal Proc. Mag., IEEE*, 25(3), 39–49.
- Chu, W.C. (2004) *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. John Wiley & Sons.
- Cox, R.V., De Campos Neto, S.F., Lamblin, C., and Sherif, M.H. (2009) ITU-T coders for wideband, superwideband, and fullband speech communication. *Communications Mag., IEEE*, 47(10), 106–109.
- Dahl, G.E., Yu, D., Deng, L., and Acero, A. (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio, Speech, and Language Proc.*, 20(1), 30–42.
- Dudley, H. (1940) The carrier nature of speech. *Bell Sys. Tech. J.* 19(4), 495–515.
- Erro, D., Sainz, I., Navas, E., and Hernaez, I. (2014) Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE J. Selected Topics in Signal Proc.*, 8(2), 184–194.
- ETSI (1992) GSM Full Rate Speech Transcoding. Recommendation ETSI GSM 06.10, European Telecommunication Standards Institute.
- ETSI (2011) AMR speech codec, general description. Standard 3GPP TS 26.071, version 10.0.0, 3rd Generation Partnership Project.
- Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992) An adaptive algorithm for mel-cepstral analysis of speech. *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, volume 1, pp. 137–140 IEEE.
- Gales, M. and Young, S. (2008) The application of hidden Markov models in speech recognition. *Found. Trend. Sig. Proc.*, 1(3), 195–304.
- Goodman, J.T. (2001) A bit of progress in language modeling. *Comp. Speech Lang.*, 15(4), 403–434.
- Huang, X., Acero, A., Hon, H.W., and Foreword By-Reddy, R. (2001) *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall.
- ITU (1988) *Pulse code modulation (PCM) of voice frequencies*. Recommendation ITU-T G.711, International Telecommunication Union, Geneva, Switzerland.
- Jurafsky, D. and Martin, J.H. (2008) *Speech and Language Processing: An Introduction To Natural Language Processing, Computational Linguistics, and Speech*, 2nd edn. Pearson Prentice Hall.
- Karjalainen, M. and Laine, U.K. (1977) Speech synthesis project in Tampere: Results and applications. *Proc. of IV Nordic Meeting on Med. and Biol. Eng.*, pp. 30.1–3.
- Karjalainen, M., Laine, U., and Toivonen, R. (1980) Aids for the handicapped based on "synte 2" speech synthesizer. *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, volume 5, pp. 851–854 IEEE.