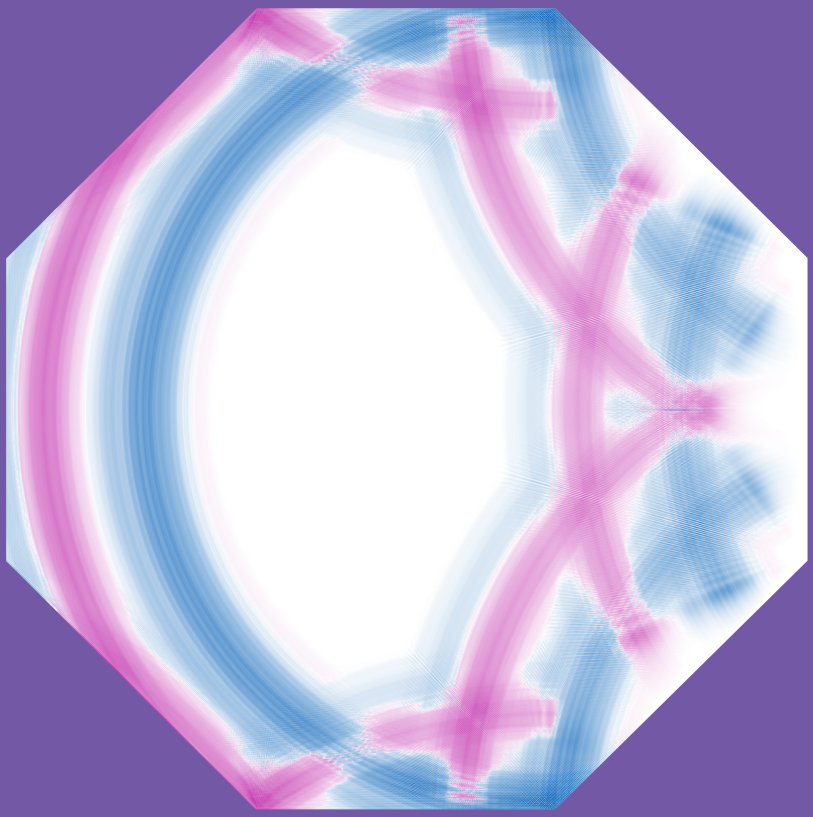
Department of Signal Processing and Acoustics

Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech

Heikki Kallasjoki



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 31/2016

Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech

Heikki Kallasjoki

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on 15 April 2016 at 12.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics

Supervising professor
Prof. Mikko Kurimo

Thesis advisor
Doc. Kalle Palomäki

Preliminary examiners
Doc. Tomi Kinnunen, University of Eastern Finland, Finland
Dr. Emmanuel Vincent, Inria, France

Opponent
Prof. Dorothea Kolossa, Ruhr-Universität Bochum, Germany

Aalto University publication series
DOCTORAL DISSERTATIONS 31/2016

ISBN 978-952-60-6665-3 (printed)
ISBN 978-952-60-6666-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6666-0

Unigrafia Oy
Helsinki 2016

Finland

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Heikki Kallasjoki

Name of the doctoral dissertation
Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 31/2016

Field of research Speech and Language Technology

Manuscript submitted 28 September 2015 **Date of the defence** 15 April 2016

Permission to publish granted (date) 27 November 2015 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

The task of automatic speech recognition has received considerable research attention and many systems have seen large-scale commercial deployment. However, lack of robustness is still a barrier to their use in novel applications. While human listeners are adept in understanding spoken language in diverse environments, the signal distortion caused by noise and reflected sounds severely degrades the accuracy of conventional systems. This thesis studies methods of reducing the effects of such distortions, improving the performance of speech recognition in challenging conditions.

The emphasis of this thesis is on algorithms that enhance the sequence of input features observed by a speech recognition system, with the aim of making them more invariant to noise and reverberation. Research on several ways of addressing the problem is included. Weighted linear prediction is considered as a method to incorporate robustness in spectral modeling used for speech feature extraction. To counteract additive noise, improvements are proposed to algorithms based on the missing data framework and the use of non-negative matrix factorization as a tool for separating sound sources. Speech corrupted by reverberation is addressed by extending the source separation model to account for convolutional distortion. Further, a method of transforming the corrupted features based on matching their distribution to that of uncorrupted speech is presented. The positive impact of the proposed approaches on speech recognition performance is confirmed and quantified by experimental evaluation on large vocabulary continuous speech recognition tasks.

Complementing the work, methods to extract and utilize information about the varying uncertainty of the enhanced features are investigated. While no system is capable of perfectly removing all traces of noise from the speech features, it is often possible to estimate the local accuracy of the processed speech. This information can be used in the decoding stage of a speech recognition system, to de-emphasize the regions of the input where the uncertainty is high, and the input features are more likely to be incorrect. This thesis proposes and evaluates heuristic uncertainty metrics compatible with the missing data and non-negative matrix factorization feature enhancement systems.

Keywords automatic speech recognition, noise robust ASR, non-negative matrix factorization, observation uncertainty, speech dereverberation

ISBN (printed) 978-952-60-6665-3 **ISBN (pdf)** 978-952-60-6666-0

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2016

Pages 214 **urn** http://urn.fi/URN:ISBN:978-952-60-6666-0

Tekijä
Heikki Kallasjoki

Väitöskirjan nimi
Piirteiden korjaus ja epävarmuuden arviointi melua ja kohinaa sisältävän puheen tunnistuksessa

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 31/2016

Tutkimusala Puhe- ja kieliteknologia

Käsikirjoituksen pvm 28.09.2015 **Väitöspäivä** 15.04.2016

Julkaisuluvan myöntämispäivä 27.11.2015 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Puheen automaattista muuttamista tekstiksi on tutkittu laajasti, ja sillä on jo monia mittavia kaupallisia sovelluskohteita. Puutteet järjestelmien kyvyssä sietää häiriöitä vaikeuttavat kuitenkin edelleen niiden käyttöä monissa uudenlaisissa käyttötarkoituksissa. Vaikka ihmiset pystyvät ongelmitta ymmärtämään puhetta erilaisissa ympäristöissä, melun ja pinnoista heijastuvien äänien aiheuttamat häiriöt puhesignaalissa heikentävät merkittävästi tavanomaisten järjestelmien tarkkuutta. Tämä väitöskirja tutkii menetelmiä, joiden tarkoitus on vähentää tällaisten häiriöiden vaikutusta, ja auttaa siten parantamaan puheentunnistuksen laatua haastavissa olosuhteissa.

Väitöskirjan pääaiheena ovat algoritmit, joilla voidaan muokata puhesignaalin käyttämiä piirteitä siten, että melu ja kaiku vaikuttaa niihin vähemmän. Väitöskirjan tutkimus kohdistuu useisiin tapoihin ratkaista tämä ongelma. Painotettua lineaariprediktiota tutkitaan menetelmänä huomioida häiriönsietävyys puheentunnistuksen piirteiden käyttämissä spektrimalleissa. Additiivisen melun vaimentamiseksi väitöskirjassa esitetään parannuksia algoritmeihin, jotka perustuvat puuttuvan tiedon käsittelyyn sekä ei-negatiivisen matriisihajotelman käyttöön äänilähteiden erottelussa. Kaiuntaa sisältävän puheen ongelmaa käsitellään laajentamalla äänilähteiden erottelussa käytettyä matriisihajotelmaa siten, että se sisältää myös konvoluutioon pohjautuvan häiriön mallin. Lisäksi esitellään menetelmä kaiunnan vaikutuksen vähentämiseksi muokkaamalla piirteiden jakaumaa vastaamaan paremmin kaiuttoman puheen jakaumaa. Ehdotettujen lähestymistapojen positiivinen vaikutus puheentunnistuksen tarkkuuteen selvitetään kokeilla, joissa tehtävänä on laajan sanaston jatkuvan puheen tunnistus.

Piirteiden muokkauksen lisäksi väitöskirjassa tutkitaan tapoja saada ja hyödyntää tietoa siitä, miten luotettavia tuloksena saadut piirteet ovat. Vaikka mikään järjestelmä ei pysty täysin poistamaan kaikkia melun jälkiä puheen piirteistä, on usein mahdollista arvioida paikallisesti, kuinka tarkkoja käsitellyt piirteet ovat. Tätä arviota voidaan hyödyntää puheentunnistuksen dekoodausvaiheessa vähentämällä epäluotettavien alueiden painoarvoa. Väitöskirjassa esitetään ja arvioidaan sellaisia piirteiden luotettavuuden heuristisia mittareita, jotka ovat yhteensopivia puuttuvaan tietoon sekä ei-negatiivisen matriisihajotelman käyttöön perustuvien menetelmien kanssa.

Avainsanat automaattinen puheentunnistus, melusietoinen puheentunnistus, ei-negatiivinen matriisihajotelma, havaintojen epävarmuus

ISBN (painettu) 978-952-60-6665-3 **ISBN (pdf)** 978-952-60-6666-0

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2016

Sivumäärä 214 **urn** http://urn.fi/URN:ISBN:978-952-60-6666-0

Preface

This thesis has its origins at the Department of Information and Computer Science of the Helsinki University of Technology. After a merger of the university into Aalto University, and a short organizational (and physical) leap sideways, it was finished at the Department of Signal Processing and Acoustics, part of the School of Electrical Engineering of Aalto University. Along the way, the work has received external funding from multiple sources. I am thankful for the support of the Helsinki Graduate School in Computer Science and Engineering (Hecse) and the Academy of Finland grants for the CSI-Speech project (135003), projects on noise robust automatic speech recognition (114369, 140969, 272710), the "Say it again, kid!" project (274075) and the Finnish Centre of Excellence in Computational Inference Research (COIN, 251170). The experimental work in this thesis was also supported by the computational resources of the Aalto Science-IT project.

I would like to thank my primary thesis instructor, Doc. Kalle Palomäki, for all the guidance during the research leading to this thesis, as well as valuable suggestions and comments given during the preparation of this manuscript. By leading the Noise Robust Speech Recognition team within the Speech Recognition research group, he has fostered fruitful collaboration, inspiring several of the publications included in this thesis. He has also been directly involved as a co-author in each of the included publications.

I am grateful for the support of my supervisor, Prof. Mikko Kurimo, who also acted as my instructor at the Department of Information and Computer Science. As the head of the Speech Recognition research group, he has been essential in building the framework of speech recognition research at Aalto University, within which this thesis was written. He has also been instrumental in arranging the funding for this work, allowing

me to focus on the research itself. I would also like to thank Prof. Erkki Oja for being my supervisor during the early stages of my doctoral studies.

This work would not exist in its current form without the support of several co-authors of the included publications. I am thankful to Prof. Paavo Alku, Dr. Jouni Pohjalainen and Mr. Carlo Magi (in memoriam, 1980-2008) for including me in the weighted linear prediction research project. I would also like to thank Dr. Jort Gemmeke and Prof. Tuomas Virtanen for the collaboration on topics involving non-negative matrix factorization, as well as Prof. Guy Brown and Dr. Amy Beeston for their contributions to missing data processing. For their formal contributions in the form of co-authorship, but also for informal discussions on topics occasionally ranging far outside the scope of this thesis, I would especially like to thank fellow residents of ~~cell~~ office room G437, Mr. Sami Keronen and Ms. Ulpu Remes, as well as other past and current members of the Speech Recognition research group.

I am grateful to the preliminary examiners of this thesis, Doc. Tomi Kinnunen and Dr. Emmanuel Vincent, for their time and effort to evaluate this work, as well as their suggestions on how to improve it.

Finally, I wish to thank all my friends and relatives for the company and support. I also appreciate the interest my colleagues at Google have shown towards this thesis (and particularly its completion date). Most of all, I thank my wife Kati for invaluable opinions, support and motivation, and, especially during the final stages of finishing this thesis, for enduring many a nice weekend spent inside listening to the sound of typing. Without her, the following date would likely be much, much farther in the future.

London, December 13, 2015,

Heikki Kallasjoki

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Heikki Kallasjoki, Kalle J. Palomäki, Carlo Magi, Paavo Alku and Mikko Kurimo. Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. In *Proceedings of the 13th International Conference on Speech and Computer (SPECOM 2009)*, pages 221–225, St. Petersburg, Russia, June 2009.

II Jouni Pohjalainen, Heikki Kallasjoki, Paavo Alku, Kalle J. Palomäki and Mikko Kurimo. Weighted linear prediction for speech analysis in noisy conditions. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, pages 1315–1318, Brighton, UK, September 2009.

III Heikki Kallasjoki, Ulpu Remes, Jort F. Gemmeke, Tuomas Virtanen and Kalle Palomäki. Uncertainty measures for improving exemplar-based source separation. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pages 469–472, Florence, Italy, August 2011.

IV Heikki Kallasjoki, Sami Keronen, Guy J. Brown, Jort F. Gemmeke, Ulpu Remes and Kalle J. Palomäki. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, pages 58–63, Florence, Italy, September 2011.

V Sami Keronen, Heikki Kallasjoki, Ulpu Remes, Guy J. Brown, Jort F. Gemmeke, Kalle J. Palomäki. Mask estimation and imputation

methods for missing data speech recognition in a multisource reverberant environment. *Computer Speech and Language*, volume 27, issue 3, pages 798–819, May 2013.

VI Heikki Kallasjoki, Jort F. Gemmeke and Kalle J. Palomäki. Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 22, issue 2, pages 368–380, February 2014.

VII Heikki Kallasjoki, Jort F. Gemmeke, Kalle J. Palomäki, Amy V. Beeston, Guy J. Brown. Recognition of reverberant speech by missing data imputation and NMF feature enhancement. In *Proceedings of the REVERB Workshop*, Florence, Italy, May 2014.

VIII Kalle J. Palomäki, Heikki Kallasjoki. Reverberation robust speech recognition by matching distributions of spectrally and temporally decorrelated features. In *Proceedings of the REVERB Workshop*, Florence, Italy, May 2014.

IX Sami Keronen, Heikki Kallasjoki, Kalle J. Palomäki, Guy J. Brown, Jort F. Gemmeke. Feature enhancement of reverberant speech by distribution matching and non-negative matrix factorization. *EURASIP Journal on Advances in Signal Processing*, volume 2015, article 76, August 2015.

Author's Contribution

Publication I: "Noise robust LVCSR feature extraction based on stabilized weighted linear prediction"

The article evaluates the use of a spectral envelope estimation method based on stabilized weighted linear prediction for speech feature extraction, in the context of a large vocabulary continuous speech recognition task using recordings from real noisy environments. The author extended the feature extraction module of a speech recognition system to support the method, performed the experiments, and was the main writer of the paper.

Publication II: "Weighted linear prediction for speech analysis in noisy conditions"

The article investigates the spectrum analysis performance of stabilized and unstabilized weighted linear prediction, including large vocabulary noisy speech recognition experiments to demonstrate the increased robustness of the proposed methods. The author was responsible for performing the speech recognition experiments and wrote the corresponding portion of the paper.

Publication III: "Uncertainty measures for improving exemplar-based source separation"

The article proposes heuristic estimators of observation uncertainty to improve the performance of a noisy speech feature enhancement method based on sparse source separation. The author designed and implemented

the heuristic metrics, performed the speech recognition experiments, and was the main writer of the paper.

Publication IV: "Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments"

The article describes a submission to the 2011 CHiME challenge for noise robust speech recognition. The proposed system combines two existing algorithms, a binaural missing data mask estimation method and a missing data imputation scheme based on sparse non-negative matrix factorization, extended by heuristic observation uncertainty estimates. The author adapted the existing systems for the CHiME challenge framework, performed the experiments, and was the main writer of the paper.

Publication V: "Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment"

The article extends the work in Publication IV by proposing an improved mask estimation method based on a Gaussian mixture model classifier using a diverse set of acoustic features. The sparse imputation is also compared against a cluster-based imputation scheme. The author contributed to conducting the experiments involving the reference systems and wrote the descriptions of some of the acoustic features and the sparse imputation method.

Publication VI: "Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition"

The article presents an in-depth study of the system proposed in Publication III, augmented by new uncertainty heuristics and a channel normalization step to improve the match between signal and the speech and noise dictionaries. The recognition performance is also evaluated using real noisy speech and compared against the SPLICE feature enhancement method. The author was responsible for defining and implementing the uncertainty estimates, conducting the experiments, and was the main writer of the paper.

Publication VII: "Recognition of reverberant speech by missing data imputation and NMF feature enhancement"

The article describes a submission to the 2014 REVERB challenge for speech dereverberation. Two systems are proposed: missing data imputation with various reverberation-specific mask estimation methods, and a novel extension of the non-negative matrix factorization source separation that incorporates the estimation of a convolutional filter as a part of the matrix factorization. The author designed and implemented the extended source separation method, implemented the missing data mask estimation methods, conducted the experiments, and was the main writer of the paper.

Publication VIII: "Reverberation robust speech recognition by matching distributions of spectrally and temporally decorrelated features"

The article proposes a dereverberation method submitted to the 2014 REVERB challenge. The method is based on the histogram equalization principle, applied to a feature representation that includes long time context. The author contributed to the running of the experiments and wrote parts of the description of the experimental system.

Publication IX: "Feature enhancement of reverberant speech by distribution matching and non-negative matrix factorization"

The article conducts an extensive evaluation of the non-negative matrix factorization dereverberation method proposed in Publication VII with several advanced back-end speech recognition systems. In addition, the distribution matching dereverberation described in Publication VIII is evaluated as a method of providing an initial estimate. The author assisted in conducting the experiments and wrote the description of the non-negative matrix dereverberation algorithm.

List of Abbreviations

ASR	automatic speech recognition
BCMI	bounded conditional mean imputation
CI	cluster-based imputation
CMLLR	constrained maximum likelihood linear regression
CMS	cepstral mean subtraction
DFT	discrete Fourier transformation
DNN	deep neural network
EM	expectation maximization
FFT	fast Fourier transformation
FIR	finite impulse response
GMM	Gaussian mixture model
GRBM	Gaussian-Bernoulli restricted Boltzmann machine
HMM	hidden Markov model
ILD	interaural level difference
ITD	interaural time difference
LER	letter error rate
LP	linear prediction
LVCSR	large vocabulary continuous speech recognition
MAP	maximum *a posteriori*
MFCC	mel-frequency cepstral coefficient
ML	maximum likelihood
MLLR	maximum likelihood linear regression
MLLT	maximum likelihood linear transformation
MLP	multilayer perceptron
MVDR	minimum variance distortionless response
NAT	noise adaptive training
NMF	non-negative matrix factorization
PCA	principal component analysis

PLP	perceptual linear prediction
PMC	parallel model combination
RIR	room impulse response
SI	sparse imputation
SNR	signal-to-noise ratio
STE	short-time energy
STFT	short-time Fourier transformation
SVM	support vector machine
SWLP	stabilized weighted linear prediction
VTS	vector Taylor series
WER	word error rate
WFST	weighted finite state transducer
WLP	weighted linear prediction

List of Symbols

a	a scalar
$\boldsymbol{a}$	a vector
$\boldsymbol{A}$	a matrix
$\odot$	elementwise multiplication
$(\cdot)^{\mathsf{T}}$	transpose
$(\cdot)^{\mathsf{H}}$	conjugate transpose
$\|\boldsymbol{a}\|_1$	L^1-norm of vector $\boldsymbol{a}$
$[\boldsymbol{A}]_{r,c}$	row r, column c of matrix $\boldsymbol{A}$
$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	probability density of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ at $\boldsymbol{x}$

a_k	the k'th linear prediction coefficient
$\boldsymbol{a}$	linear prediction coefficient vector
$\boldsymbol{A}$	activation matrix
c	frequency band index
C	number of frequency bands
$e(n)$	linear prediction residual
E	linear prediction error
K	number of dictionary atoms
l	mixture model component index
L	number of mixture model components
M	weight function window width
n	sample index for a time domain signal
p	model order for linear predictive models
q	acoustic model state
$\boldsymbol{R}$	autocorrelation matrix (Chapter 3)
$\boldsymbol{R}$	filter matrix (Chapter 6)
$\boldsymbol{S}$	dictionary matrix

t	time frame index
T	number of time frames in a window
w_n	time-domain weight function
$\boldsymbol{x}_t$	clean speech features
$\hat{\boldsymbol{x}}_t$	point estimate of the clean speech features
$y(n)$	time-domain signal (Chapter 3)
$\boldsymbol{y}_n$	time-domain signal vector (Chapter 3)
$\boldsymbol{y}_t$	noisy observation features (Chapter 4, 5, 6)
$\boldsymbol{Y}$	observation matrix
$\boldsymbol{\Gamma}$	Gaussian mixture model parameters
$\boldsymbol{\theta}$	feature extraction parameters
λ	sparsity coefficient
$\boldsymbol{\lambda}$	sparsity coefficient vector
$\boldsymbol{\mu}$	mean vector
$\boldsymbol{\sigma}$	oracle variance in the acoustic model domain
$\boldsymbol{\Sigma}$	covariance matrix

1. Introduction

1.1 Speech Recognition in Challenging Conditions

Speech, for most of us, is a natural, effortless way of communicating with one another. Given its convenience over alternatives such as typing, it is no surprise that a robust method for a computer to extract the meaning of a spoken phrase is a powerful tool, invaluable for many applications involving human-computer interaction. Recent years, after decades of both sustained research effort and exponential increases in available computing power, have seen a major rise in the widespread adoption of speech recognition technology in everyday life.

While advances in the technology of computing have led to significantly improved speech recognition systems, they have also brought on additional challenges. The trend towards increasingly ubiquitous and mobile computing devices increases the demands for a speech recognition system to work under noisy conditions, such as on the street or in a busy restaurant. The spread of computers from the desk to the living room calls for new kinds of applications, where it is not feasible to require the speaker to be close to the microphone. As a result, speech recognition must now cope with sources of noise equally loud as the person speaking, as well as the complexity of sounds reflecting from the walls and other surfaces in the environment.

For humans, the above scenarios pose little difficulty. We are able to adapt to the environment and focus on the speech of a single speaker without any conscious effort. Conventional computer systems, based on a predetermined model of what speech sounds like, are less lucky, and often perform very poorly even under only moderately challenging conditions.

Even though our tools for modeling the patterns of sound that speech

consists of have grown more sophisticated, up to a degree they still share an unavoidable assumption: the training data used to build the system defines what the system can understand. To collect a set of recordings representing all the variation there is in speech produced by the people who speak a particular language is already a hard task. To do the same for every possible environment that speech may occur in is yet harder.

Many solutions have been proposed to address these problems. In this work, the main focus is on *front-end* methods, where the goal is to process the signals used as input by the speech recognition system in such a way that they are affected as little as possible by any distortion caused by noise and reflected sound. While the resulting processed signals are never completely free of environmental effects, the methods combine well with complementary approaches of dealing with distorted speech, and help to enable new applications of speech recognition.

1.2 Background of the Thesis

The research leading to this thesis took place in 2009–2014, within the speech recognition research group initially part of the Department of Information and Computer Science at the Helsinki University of Technology. During that time, the Helsinki University of Technology became part of Aalto University (in 2010) and the research group moved to the Department of Acoustics and Signal Processing (in 2013). The Department of Information and Computer Science, and its organizational predecessors, has a long history of research in the field of automatic speech recognition and machine learning, leading back to Prof. Teuvo Kohonen's pioneering 1970s work on pattern recognition using artificial neural networks.

Much of the work of the speech recognition research group has focused on the recognition of large vocabulary continuous speech in Finnish (Hirsimäki et al., 2006, 2009; Siivola et al., 2007), and on the development of the *AaltoASR* automatic speech recognition system (Hirsimäki and Kurimo, 2004; Pylkkönen, 2005), described in more detail in Section 2.5 of this thesis. Recent topics of interest include discriminative training of hidden Markov models (Pylkkönen and Kurimo, 2012), morphological analysis (Grönroos et al., 2014) and adaptation (Mansikkaniemi and Kurimo, 2015) for language modeling, information retrieval (Turunen and Kurimo, 2011), and recognition of noisy and reverberant speech, the topic of this thesis. In addition, the related field of parametric speech synthesis based

on hidden Markov models has received attention (Karhila et al., 2014).

1.3 Scope of the Thesis

This thesis studies the task of automatic speech recognition in the presence of distortions caused by environmental noise and reverberation. In particular, while a general introduction of the topic is included in this overview, the majority of its contents relate to *front-end* methods, where the goal is to minimize the effect of distortion on the stream of features observed by the rest of the system.

The publications and research included in this thesis focus on four main topics. Methods for estimating the spectral envelope based on the concept of weighted linear prediction are investigated for use in extracting robust features for noisy speech. In the field of feature enhancement for speech corrupted by additive noise, the primary concepts covered are the *missing data* framework, and the use of the non-negative matrix factorization model for source separation. As no feature enhancement method results in an exact restoration of the original uncorrupted speech, this thesis investigates how the use of information about the uncertainty of the enhanced features can benefit speech recognition. Finally, the extension of the presented systems to the special case of the convolutional distortion caused by reverberation is covered.

1.4 Contributions of the Thesis

The main research contributions of the publications included in this thesis are:

- Experimental evaluation of the use of (stabilized) weighted linear prediction for the feature extraction of large vocabulary continuous noisy speech (Publication I, Publication II).
- A study of mask estimation and feature imputation methods for missing data feature enhancement, evaluated in the context of the 2011 CHiME challenge (Publication IV, Publication V).
- A comprehensive evaluation of the applicability of source separation based on non-negative matrix factorization to the feature enhancement of noisy speech (Publication III, Publication VI).

- Heuristic uncertainty estimators for observation uncertainty processing to enhance the sparse imputation (Publication IV) and sparse source separation (Publication III, Publication VI) feature enhancement methods.
- A novel extension to the non-negative matrix factorization source separation model to address convolutional distortion present in reverberant speech, evaluated in conjunction with missing data dereverberation methods for the 2014 REVERB challenge (Publication VII, Publication IX).
- A speech dereverberation method based on the histogram normalization principle, extended to include long time context information, also participating in the 2014 REVERB challenge (Publication VIII, Publication IX).

While full details of the above contributions can be found in the original publications, included as a part of this thesis, an overview of the methods and key experiment results are presented in the following chapters.

1.5 Contents of the Thesis

Chapter 2 presents an overview of the speech recognition problem, the building blocks of a conventional speech recognition system, and the classes of methods that have been proposed to address issues caused by noise and reverberation. The evaluation methods and data sets used in the experiments of the publications of this thesis are also described at the end of the chapter.

Chapter 3 focuses on the topic of extracting features that are more robust in difficult environmental conditions. Specifically, methods based on estimating the spectral envelope are presented. Experimental evaluation is performed on systems using a weighted linear prediction algorithm for speech feature extraction.

Chapter 4 is dedicated to the investigation of feature enhancement in the important scenario of speech corrupted by additive background noise. Two separate classes of algorithms are covered in detail: missing data imputation and the use of non-negative matrix factorization for single-channel source separation. In addition, the combination of the two in the form of sparse imputation is also examined. Finally, results for speech recognition experiments involving both kinds of systems are presented.

As no feature enhancement method is perfect, Chapter 5 looks into the use of signals representing the uncertainty of feature enhancement to aid in the recognition process. Heuristic uncertainty estimators are proposed for both the missing data and source separation algorithms presented in Chapter 4, and their impact on speech recognition performance is evaluated.

Chapter 6 focuses on the particular type of distortion present in speech recorded far away from the speaker in a reverberant environment. Topics covered include the use of missing data methods for reverberant speech, a proposed novel extension of the non-negative matrix factorization source separation model that incorporates a convolutional filter, and a method of histogram normalization that includes long time context information to reduce the effects of reverberation. Evaluation results from the REVERB challenge are presented for all three systems.

Final conclusions related to the contributions of this thesis are drawn in Chapter 7.

2. Robust Automatic Speech Recognition

2.1 The Speech Recognition Problem

The problem of building systems that understand human speech has received considerable research attention since the 1950s. As progress has been made over the years, the category of feasible tasks has expanded, from recognizing isolated digits or words selected from a small set, through connected digit sequences and command languages with restricted grammars, all the way to large (or unlimited) vocabulary continuous speech recognition (LVCSR) tasks such as unrestricted dictation (Reddy, 1976; Young, 1996). While previous problems have moved from the realm of research systems to having commercially viable solutions, achieving performance close to matching human listeners, more research has been focused on tasks where automated systems still fall significantly short of human performance levels. Prominent among these are recognizing speech distorted by the environment and informal, conversational speech.

The main focus of this thesis is on the problem of producing a text transcription from a recording of continuous, unlimited vocabulary speech corrupted by either additive noise or reverberation produced by reflected sounds in the recording environment.

2.2 Overview of an ASR System

A typical speech recognition system shares several properties with other pattern recognition tasks. The input of the system takes the form of a time sequence of *features* extracted from data recorded by sensors. Statistical models of the acoustic patterns of the target language are used to classify the features. In addition, an informative prior over the recognition output,

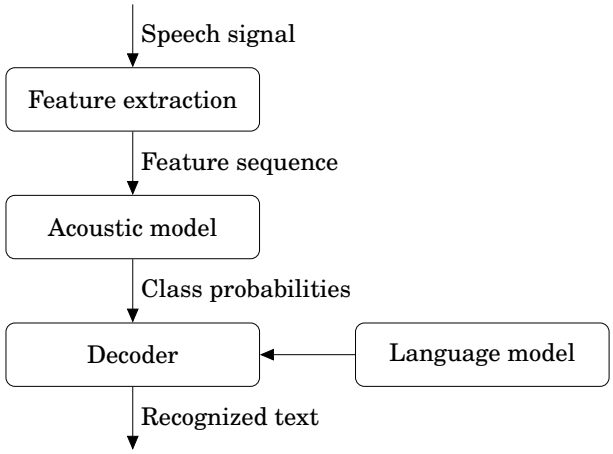


Figure 2.1. Overall structure of a conventional automatic speech recognition system.

based on higher-level models of the target language, is usually required in order to obtain acceptable performance, especially for more challenging inputs such as large vocabulary continuous speech.

Figure 2.1 presents a conceptual summary of the major components of a conventional LVCSR system, and the information flows connecting them. In practice, actual implementations often contain more complicated interactions between the components than indicated in the figure, due to, e.g., optimization reasons. The individual components are described in detail in the following sections.

While the long-sustained exponential growth of processing power has made it possible to use ever larger models and more advanced methods, the general framework depicted in Figure 2.1 has been flexible enough to accommodate the changes. Even the current state of the art systems, for the most part, follow the same outline, although recent trends toward the use of deep neural networks have blurred the line between feature extraction and acoustic modeling.

2.2.1 Feature Extraction

Although research still continues on finding feature representations more suitable for speech recognition purposes, some general principles have become well-established (Picone, 1993). The digitized waveform is typically split into partially overlapping frames approximately 10–25 ms in duration. Based on the physical limits of the human speech production

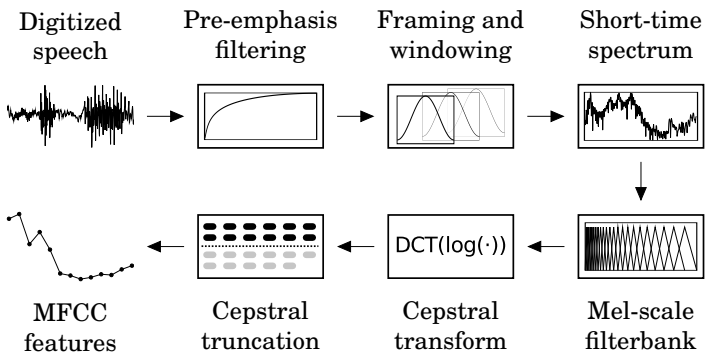


Figure 2.2. Feature extraction process of conventional MFCC features.

mechanisms, the frequency content within a single frame is assumed to be sufficiently stationary. Each individual frame is then characterized by an estimate of its spectrum.

Mel-frequency cepstral coefficients (MFCCs) have long been a popular choice as a feature representation in ASR systems (Davis and Mermelstein, 1980). The process for extracting conventional MFCC features is illustrated in Figure 2.2. The high-pass pre-emphasis filtering step compensates for the spectral tilt of the speech spectrum, while the mel-scale filterbank approximates the varying frequency resolution of the human hearing system. Finally, truncating the cepstral coefficients effectively isolates the spectral envelope information, which is crucial in discerning the formants, from the spectral fine structure, which is predominantly speaker-dependent.

Motivated by studies on human speech perception, the static MFCC coefficients are commonly extended by first- and second-order differences (*delta* and *delta-delta* features) to capture information about the dynamics of the signal (Furui, 1986). Other typical post-processing steps include feature normalization schemes, such as cepstral mean subtraction (CMS), designed to minimize the effect the recording channel has on the features (Atal, 1974).

The recent trend towards the increasing use of deep neural networks (DNNs) in conjunction with ever larger data sets in speech recognition systems has also changed the way feature extraction is viewed. In particular, successful methods have been proposed that start with a relatively low-level feature representation, such as the linear or the mel-scale spectrum, and consider the feature extraction stage only a single part of the DNN (Seide et al., 2011; Sainath et al., 2014). This approach has the ben-

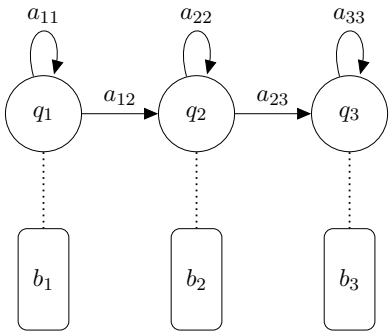


Figure 2.3. A state transition diagram for a three-state left-to-right HMM. Non-zero transition probabilities a_{ij} from state s_i to s_j are indicated by solid edges. Each state q_i is further associated with a conditional emission distribution $b_i(\boldsymbol{x}) = p(\boldsymbol{x} \mid q_i)$ over the observation vector $\boldsymbol{x}$.

efit of being able to optimize the feature representation itself, as a part of the acoustic model training. However, the training data needs to be sufficiently diverse to avoid generalization problems for unseen data (Yu et al., 2013).

2.2.2 Acoustic Model

The role of the acoustic model is to provide a statistical model for sequences of extracted feature vectors, to enable their classification into classes corresponding to meaningful units of speech, such as phonemes, words or sentences. While in theory the individual feature vectors can be classified in isolation, for accurate speech recognition, the temporal dependencies inherent to the speech signal need to be accounted for in the acoustic modeling. The hidden Markov model (HMM) framework is the predominant tool for this purpose.

A HMM consists of a first-order discrete Markov chain, characterized by its *transition probabilities* between the states, combined with a set of conditional *emission distributions* that give the conditional probability of an observed feature vector when the current state of the model is known. The number of parameters, and consequently the expressiveness of the model, can be further reduced by disallowing certain transitions, i.e., forcing their probability to zero. A common choice in speech recognition tasks is the *left-to-right HMM* illustrated in Figure 2.3.

Given a sequence of T HMM states q_1^T, the model defines a distribution over a corresponding sequence of observation vectors $\boldsymbol{x}_1^T$. The components of this distribution are the transition probabilities a_{qr} between states q

and r, and the emission distributions $b_q(\boldsymbol{x}) = p(\boldsymbol{x} \mid q)$. Mathematically, the distribution has the form

$$p(\boldsymbol{x}_1^T \mid q_1^T) = \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \prod_{t=1}^{T} b_{q_t}(\boldsymbol{x}_t). \tag{2.1}$$

Several efficient algorithms have been devised to solve problems related to the HMM framework. The acoustic model training problem — finding the transition probabilities a_{qr} and emission distributions b_q that maximize the likelihood $p(\boldsymbol{x} \mid q)$ of a corpus of labeled training data — is addressed by the *Baum-Welch (forward-backward) algorithm* (Baum et al., 1970), which belongs to the family of iterative expectation-maximization (EM) algorithms (Bilmes, 1998). Given a trained acoustic model and an unknown utterance $\boldsymbol{x}_1^T$, the *Viterbi algorithm* (Viterbi, 1967) can be used to solve the decoding problem

$$\arg\max_{q_1^T} \; p(\boldsymbol{x}_1^T \mid q_1^T), \tag{2.2}$$

to find the most probable sequence of HMM states corresponding to the observation.

If the vocabulary to recognize is limited, individual HMM models can be constructed for each word, with the number of states used per word based on, e.g., the complexity of the word or the number of available training samples (Rabiner, 1989). This approach is infeasible for an unlimited-vocabulary task, for which a typical solution is to instead model units such as phonemes, and construct word models by concatenating the models of their constituent phonemes. To account for context-sensitive phonetic variation, separate models can be used depending on the surrounding phonemes in the word. The de-facto standard approach in large vocabulary continuous speech recognition problems is the three-state left-to-right triphone HMM (Schwartz et al., 1984), which uses one phoneme of both preceding and succeeding context, and models the structure within individual phonemes with three HMM states. In order to avoid data sparsity issues for triphones that occur rarely in the training set, some model parameters are normally shared on the triphone model (Lee, 1990) or HMM state (Young et al., 1994) level.

The form of the emission distributions must be specified to fully define the HMM model. Popular choices in early work, primarily for reasons of computational efficiency, were multivariate Gaussian distributions, and the nonparametric discrete distribution (Schwartz et al., 1984). However,

the acoustic features are not necessarily well modeled by a single Gaussian distribution.

A mixture of Gaussian distributions is capable of approximating a large class of probability density functions (Sorenson and Alspach, 1971). As the estimation of parameters of Gaussian mixture model (GMM) emission distributions in an HMM can be performed efficiently (Juang et al., 1986), the combined GMM-HMM model has enjoyed widespread popularity in speech recognition systems. State-of-the-art systems extend the basic GMM-HMM model with a number of further refinements, such as feature space projections, structured covariance matrices and discriminative training criteria (Gales and Young, 2008).

A popular recent alternative for the GMM-HMM model is the context-dependent deep neural network hidden Markov model (CD-DNN-HMM) hybrid architecture (Dahl et al., 2012). The key concept is to train the deep neural network to predict the posterior distribution $p(q_t \mid \boldsymbol{x}_t)$ of the context-dependent HMM states q_t under observation $\boldsymbol{x}_t$, which can then be used to derive the observation probabilities $p(\boldsymbol{x}_t \mid q_t)$ used by the HMM. In order to cope with large numbers of hidden layers, the individual layers are commonly "pre-trained" as a stack of separate feature detectors before applying conventional backpropagation training (Hinton et al., 2006). The combination of rectified linear units (ReLU) as the network nonlinearity and the "dropout" regularization scheme is also a widely used training method for deep neural networks in speech recognition (Dahl et al., 2013).

2.2.3 Language Model

The acoustic model is not always sufficient to distinguish between similar candidate words, especially when the vocabulary size is large. In addition, in continuous speech the acoustic cues delimiting the word boundaries can be faint or nonexistent, leading to a further increase in the number of potential matches. While the acoustic models already encode some information about the target language, this information is inherently limited to a short time frame. For accurate speech recognition, separate models that make use of longer-term dependencies in the target word sequence are needed.

For tasks which are limited to a fixed grammar, knowledge of allowed word sequences can be readily incorporated in the HMM structure. However, much of the research in speech recognition has focused on tasks

which involve the full complexity of natural language, such as dictation. In this field, *N-gram models* are the predominant framework.

The underlying simplifying assumption in an N-gram model is that the probability of each word w_i in a k-word sequence $\boldsymbol{w}_1^k$ depends only on the $N-1$ preceding words: (Bahl et al., 1983)

$$p(\boldsymbol{w}_1^k) = \prod_{i=1}^{k} p(w_i \mid \boldsymbol{w}_1^{i-1}) \approx \prod_{i=1}^{k} p(w_i \mid \boldsymbol{w}_{i-N+1}^{i-1}). \tag{2.3}$$

Increasing the model order N makes the model more detailed, but also requires more memory to store the model. In addition, the problem of data sparsity is exacerbated: as the number of potential N-grams grows exponentially, it becomes increasingly likely that some valid N-grams occur only very rarely or never in the training data, leading to unreliable frequency statistics. However, higher-order models can be required for high accuracy, especially if sub-word units such as statistical morphemes are used (Hirsimäki et al., 2006, 2009).

To address the data sparsity problem, several *smoothing* methods that adjust the estimated probabilities have been proposed (Chen and Goodman, 1999), with Kneser–Ney smoothing (Kneser and Ney, 1995) being the conventional choice. For more compact models, irrelevant N-grams may be pruned from the model, or a variable-order model grown by incrementally adding N-grams (Siivola et al., 2007).

2.2.4 Decoder

The decoding problem can be stated as

$$\hat{W} = \arg\max_{W} p(\boldsymbol{x}_1^T \mid W)\, p(W), \tag{2.4}$$

where $W = w_1^k$ is an arbitrary sequence of k words, and $\boldsymbol{x}_1^T$ is the sequence of observed feature vectors. The language model captures the linguistic knowledge to determine $p(W)$, while the acoustic model defines the term $p(\boldsymbol{x}_1^T \mid W)$. For a conventional HMM acoustic model using context-dependent phoneme units, each word w_i can be expanded to a set of possible sequences of HMM states corresponding to that word, allowing Equation (2.4) to be equivalently written as (Aubert, 2002)

$$\hat{W} = \arg\max_{W} \left\{ \left(\sum_{q_1^T} p(\boldsymbol{x}_1^T, q_1^T \mid W) \right) p(W) \right\}, \tag{2.5}$$

where q_1^T is any sequence of HMM states that can represent the word sequence W.

Direct evaluation of Equation (2.5) is not generally computationally tractable, due to the exponential number of possible word and HMM state sequences. Practical systems typically simplify the problem to

$$\hat{W} = \arg\max_{W} \left\{ \left(\max_{q_1^T} p(\boldsymbol{x}_1^T, q_1^T \mid W) \right) (p(W))^{\alpha} \right\}, \tag{2.6}$$

which takes into account only the most probable state sequence, and adds a heuristic scale factor α to the language model probability term. In this case, the task can be formulated as a search problem for the most probable path through a network constructed based on the acoustic and language models.

For simple tasks with a limited vocabulary and language model, it is viable to statically expand the entire network into what is essentially a single HMM, and the standard Viterbi algorithm (Viterbi, 1967) is applicable. However, more heuristic search strategies need to be adopted for large vocabulary continuous speech recognition, as the potential search space grows too large for simple static expansion. Asynchronous, stack-based decoders, such as the one proposed by Jelinek (1969), were typical in early approaches, constrained strongly by limited computational resources. The prevailing methods in later work have been time-synchronous beam search algorithms, with several design features to reduce the computational burden, such as organizing the lexicon as a (phonetic) prefix tree (Ney et al., 1992). Young et al. (1989) provides a unifying conceptual view of such algorithms as variations of propagating tokens in the search network. An overview and a taxonomy of decoding techniques are given by Aubert (2002).

More recently, static search network expansion under the weighted finite state transducer (WFST) framework (Pereira et al., 1994) has emerged as a prominent method in state-of-the-art systems. The framework provides a common representation for the acoustic and language models. It is sufficiently flexible to accommodate both HMMs and N-gram grammars, including characteristics like context-dependency and pronunciation variants, and provides efficient operations for combining the models (Mohri et al., 2002; Hori and Nakamura, 2013).

2.3 Speech in Noisy Environments

Noisy and unpredictable environments pose difficult problems when speech recognition technology is brought from the laboratory to real world appli-

cations. The observed sequence of features no longer corresponds directly to the utterance of the speaker we wish to recognize. Instead, they are distorted by the transmission path of the signal from the speaker to the microphone and, depending on the application, may be mixed together with noise sources such as traffic noise, music or overlapping speech. While human listeners are relatively insensitive to such effects, the recognition performance of many automatic systems degrades dramatically even under mild or moderate levels of distortion and noise (Lippmann, 1997).

A large amount of research has sought to address the problem, and the proposed solutions involve refinements to all components of the speech recognition system. A summary of the major directions is given here, while the latter chapters elaborate on those lines of research on which this thesis is focused.

2.3.1 Multi-condition Training

The primary goal in recognizing distorted speech is to reduce the mismatch between the observations and the statistical models learned from the training data. A natural approach to achieve this is to include both clean and distorted speech in the training set, in a process called *multi-condition training*. If it is possible to construct a training set matching the test condition, significant improvements can be achieved, though usually at the cost of some degradation of performance when recognizing clean speech.

2.3.2 Robust Feature Extraction

When collecting a suitable matching training set is infeasible due to, e.g., the unpredictability of the test environment, alternative methods of compensating for the distortion need to be found. Considering, at first, only the initial feature extraction step of an ASR system, it is possible to derive feature representations that are less affected by the changes caused by typical environmental noises than the standard MFCC features.

Methods in this category include those that replace the discrete Fourier transformation (DFT) spectrum in conventional MFCC feature extraction with a more robust spectral envelope estimate, based on techniques such as temporally weighted linear prediction (Ma et al., 1993; Magi et al., 2009; Pohjalainen et al., 2010) or the minimum variance distortionless response (MVDR) spectrum (Murthi and Rao, 2000; Dharanipragada et al.,

2007; Yapanel and Hansen, 2008). Publication I and Publication II investigate systems in this family, and the topic is covered in more detail in Chapter 3.

Robust feature representations based on filtering the samples of the individual frequency channels of short-time spectral analysis have also been proposed. In RASTA-PLP processing (Hermansky and Morgan, 1994), the channels of a perceptually motivated power spectral representation are filtered using a band-pass filter to suppress slowly varying components, which significantly reduces the effect of mostly stationary background noise. The low-pass components of the band-pass filters additionally suppress high-speed variation between successive frames caused by spectral analysis artefacts. The modulation spectrogram framework proposed by Kingsbury et al. (1998) further generalizes the RASTA-PLP algorithm.

2.3.3 Feature Enhancement

While more robust feature extraction can yield recognition performance improvements under noisy conditions, *feature enhancement* systems that attempt to compensate for distortion by adaptively transforming the features have more flexibility. In many cases such methods can still be implemented strictly in the front-end, requiring no changes to the acoustic modeling or later stages of the speech recognition system.

The *missing data* model is a general framework inspired by studies of the human auditory system (Cooke et al., 2001). The key concept is to consider only those spectro-temporal regions that are dominated by the uncorrupted target speaker to be valid observations, and treat the rest of the data as missing. Depending on the implementation, this may involve changing the acoustic modeling back-end of a speech recognition system to cope with missing values (*marginalization*), or alternatively direct reconstruction of the missing data based on the surrounding context (*imputation*). In the latter case, the method falls under the feature enhancement paradigm. Publication IV and Publication V apply missing data imputation methods for noise robust recognition, and the topic is elaborated on in Chapter 4.

Non-negative matrix factorization (NMF) algorithms (Lee and Seung, 2001) are general tools that have found applications in diverse fields of machine learning, including text mining, spectral data analysis (Berry et al., 2007) and image processing (Hoyer, 2004). In the field of speech processing, NMF has been as a tool for feature enhancement by source

separation (Gemmeke et al., 2011), but also for direct acoustic modeling within the wider context of *exemplar-based* speech processing (Sainath et al., 2012). The use of NMF algorithms for noise robust speech processing is also covered in Chapter 4.

In real-world applications, no feature enhancement scheme is capable of perfectly reconstructing the true clean speech features. Furthermore, the error between the enhanced and clean features generally varies across time and the frequency channels. For many systems, it is possible to quantify the level of this residual uncertainty about the correct result after feature enhancement. Tools such as the observation uncertainty framework (Deng et al., 2005) and uncertainty decoding (Droppo et al., 2002; Liao and Gales, 2005) can use this information to enable the ASR system to give a larger weight to the signal regions considered more reliable. Publication III and Publication VI propose and analyze a method for applying the observation uncertainty concepts in the context of an NMF-based feature enhancement system, while Publication IV considers the context of missing data feature enhancement. This topic is discussed in more depth in Chapter 5.

2.3.4 Model-based Approaches

Substituting or complementing feature enhancement, it is also possible to reduce the mismatch between the observation and the models by adapting the model parameters to the prevailing conditions. While some adaptation methods, such as vocal tract length normalization (e.g., Lee and Rose, 1998), are fundamentally limited to speaker adaptation, more generic transformations can account for the environment too.

In maximum likelihood linear regression (MLLR) adaptation, linear transformations of either only the mean (Leggetter and Woodland, 1995) or both the mean and the variance parameters (Gales and Woodland, 1996) of the Gaussian distributions of a GMM acoustic model are estimated to maximize the likelihood of the adaptation data. As an important special case, in constrained MLLR (CMLLR) the mean and variance transformations of one distribution are required to correspond. If a single transformation is further used for all distributions in the model, the adaptation can be seen as a feature enhancement method, as it can be implemented by applying a complementary transformation to the feature vectors. The maximum *a posteriori* (MAP) framework provides an alternative approach to model adaptation (Gauvain and Lee, 1994).

Prior knowledge and assumptions about the environmental noise can also be incorporated in the back-end in a more explicit form. In parallel model combination (PMC), an HMM matching the noisy observation is constructed by combining the existing clean speech HMM with a (typically single-state) HMM estimated from non-speech noise samples (Gales and Young, 1996). Alternatively, the noise statistics can be derived from the clean speech model and a noisy observation. Due to the nonlinearity of the problem, analytical approaches must employ approximations such as a truncated vector Taylor series (VTS) expansion. VTS-based noise robustness methods have been derived both for model-based adaptation (Acero et al., 2000) as well as feature compensation (Moreno et al., 1996).

2.3.5 Training Robust Acoustic Models

No compensation method perfectly removes the mismatch between models trained only on clean speech and a noisy observation, and many methods introduce new artefacts to the processed signal. For front-end methods, a straightforward way to account for these artefacts and mismatch during acoustic model training is to use a multi-condition training set processed with the feature enhancement method in question. In the case of back-end methods, noise adaptive training (NAT) approaches modify the training scheme itself to be aware of the noise model, as proposed by Kalinli et al. (2010) in the NAT algorithm for VTS-based model adaptation.

The use of information about the uncertainty of feature enhancement processing is not limited to the decoding stage, as discussed in Section 2.3.3. The acoustic model training process can also be adapted to incorporate uncertainty information from both front-end (Ozerov et al., 2013) and back-end (Liao and Gales, 2007) methods.

2.4 Recognizing Reverberant Speech

Many real-world applications would benefit from accurate *far-field* speech recognition, where the path between the speaker and the recording device is comparatively long. However, in addition to inherently lower signal-to-noise ratios, increased effects of room reverberation are a major problem in this scenario.

Reverberation has inherent characteristics that can be exploited by ex-

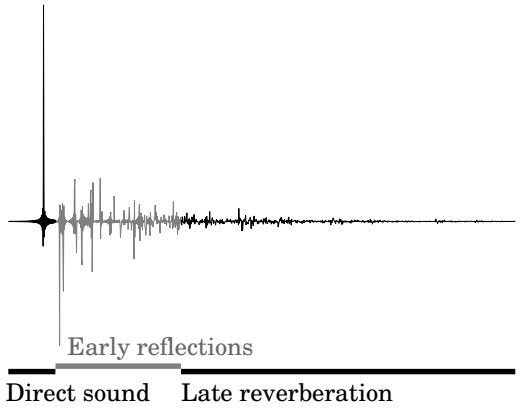


Figure 2.4. Components of a typical room impulse response simulated using an image-source model (Lehmann and Johansson, 2008).

plicitly designed solutions (Yoshioka et al., 2012). The effect of reverberation on the signal is captured by the *room impulse response* (RIR), which is often conceptually divided into three distinct components: the *direct sound* component, the *early reflections*, and the *late reverberation*. These components are illustrated on a simulated room impulse response in Figure 2.4.

The delays and amplitudes of the early reflections, up to approximately 50 ms after the direct sound, depend strongly on the positions of sound sources and microphones. By contrast, late reverberation is relatively insensitive to positioning, and its magnitude can be modeled as an exponential decay specific to the environment, often characterized by T_{60}, the time it takes for late reverberation to diminish to -60 dB compared to the direct sound. Accordingly, the reverberant signal can be approximately modeled as the convolution of the clean speech signal and a short impulse response corresponding to the early reflections, corrupted by an additive noise signal formed by the late reverberation.

If an impulse response is short compared to the feature extraction frame length, convolution of the time domain signal corresponds to multiplication in the frequency domain, and further to addition of a constant offset after the logarithmic compression common in MFCC-like feature representations. Standard feature normalization and model adaptation techniques can therefore suppress some of the effect of the early reflections.

In contrast to typical simple noise sources such as traffic noise, when considered as additive noise, the late reverberation is a highly nonstationary signal, rendering many conventional noise robustness approaches less

effective. Unlike other difficult noise sources such as interfering speakers, however, the late reverberation can be predicted from the past history of the reverberant signal. Accounting for this long-term temporal dependency is crucial to the performance of reverberation-specific solutions.

Similarly to noise robustness methods, approaches for reverberation-robust speech recognition can be classified to two main groups, depending on whether they are primarily based on modifying the front-end or the back-end (Yoshioka et al., 2012). In the former category, there is a further subdivision into three broad classes: phase-sensitive methods that operate directly on time-domain signals or (complex) short-time Fourier (STFT) spectra; methods that work on magnitude-based spectrograms, often in the mel scale; and methods that enhance features in or near the acoustic model feature domain.

In the time domain, dereverberation is a special case of the general *blind deconvolution* signal processing problem: the decomposition of an observed signal into an unknown input signal and the impulse response of an unknown convolutional filter. To avoid an ill-posed problem, some prior assumptions of either the signal, the filter, or both must be made. In the case of speech, these may include, e.g., the nonstationarity of the speech signal and the stationarity of the reverberant distortion (Hopgood and Rayner, 2003), or a linear predictive generative model for the speech production process (Nakatani et al., 2010; Kinoshita et al., 2009). Alternatively, multi-microphone methods such as LIMABEAM (Seltzer et al., 2004b) can be used.

While useful information is potentially lost as the phase information is discarded, the short-time magnitude spectra of reverberant speech are also largely invariant to speaker and microphone positioning, and therefore inherently robust to speaker motion. The effects of reverberation on magnitude spectra can be modeled by methods such as non-negative matrix factor deconvolution (NMFD) (Kameoka et al., 2009; Kumar et al., 2011), or simpler statistical models of late reverberation (Lebart et al., 2001). With suitable mask estimation, reverberation can also be compensated for by missing data methods (Palomäki et al., 2004, 2006).

Despite the inherent nonlinearity, it is also possible to model reverberation in the log-spectral domain common to speech recognition features such as MFCCs (Wölfel, 2009). In principle, convolutional distortions are approximately represented by a constant additive term in the short-term log-spectral domain. In the case of reverberation, however, typical analy-

sis window lengths used for speech feature extraction are far too short for this property to hold. Gelbart and Morgan (2002) propose a dereverberation method based on log-spectral subtraction using a long, one second window for the discrete Fourier transformation. With the exception of the long analysis window, the proposed method is similar to cepstral mean subtraction: the average log-magnitude spectrum across 45 consecutive frames, centered around each frame, is subtracted from the log-magnitude spectrum of that frame. The dereverberation is performed as a separate preprocessing step, followed by a resynthesis of the time-domain signal for use in a conventional ASR system.

The systems proposed by Publication VII and Publication VIII perform dereverberation of speech in the spectral and log-spectral domains. Both publications were submitted to the 2014 REVERB challenge, a framework for evaluating dereverberation methods (Kinoshita et al., 2013). They are presented in more detail in Chapter 6. The combination of the two systems is investigated in Publication IX.

Considering the back-end of an ASR system, while general HMM adaptation methods can reduce errors induced by reverberation, they do not explicitly make use of the long-term temporal dependencies. In order to model those dependencies, the acoustic model likelihood computation can be modified to depend on the preceding context instead of just the current time frame (Takiguchi et al., 2006).

Analogously to the case of additive noise, uncertainty information from dereverberation processing can be used by the ASR system to improve recognition (Delcroix et al., 2009; Krueger and Haeb-Umbach, 2010).

2.5 Experimental Evaluation

The common overall goal of all the publications included in this thesis is to improve the performance of ASR systems in challenging conditions. Accordingly, error rates of ASR systems have been used as a major performance metric in the empirical validation of the methods presented in the publications.

The most common measure for ASR performance is the word error rate (WER), defined as

$$\mathrm{WER} = \frac{W_\mathrm{I} + W_\mathrm{S} + W_\mathrm{D}}{W_\mathrm{R}} \cdot 100\%, \tag{2.7}$$

where W_I, W_S and W_D are the number of required insertions, substitu-

tions and deletions to convert the hypothesis to the correct result, while W_{R} is the total number of words in the reference transcript. The word error rate does not, however, make a difference between a confusion of a single phoneme in a long word, and the substitution of an entirely incorrect word. The letter error rate (LER) has therefore been considered a more accurate metric for agglutinative languages such as Finnish, where a single complex word with several suffixes often represents multiple English words[1]. Accordingly, LER is used as the primary performance metric for all ASR experiments involving Finnish in the publications of this thesis. Depending on the application, specific metrics such as the keyword accuracy used by the CHiME challenge corpus (Christensen et al., 2010) can also be more relevant in terms of final system performance.

In a majority of the publications, Publication I through Publication VI, the experiments have been performed using the Aalto University *AaltoASR* speech recognition system. While the system for the most part applies standard ASR methods, some of its design choices have been motivated by certain properties of the Finnish language. In particular, the language modeling is performed on sub-word units called *statistical morphemes*, discovered from large text corpora with an unsupervised method (Siivola et al., 2003; Hirsimäki et al., 2006). As a consequence, the system uses high-order n-gram models (Hirsimäki et al., 2009) trained with the *VariKN* toolkit (Siivola et al., 2007).

The acoustic model of the system is a state-tied GMM-HMM model over cross-word triphones. As the Finnish language commonly differentiates word meaning based on phoneme duration, the HMMs are extended with explicit gamma distribution state duration models. The decoder is based on a one-pass time-synchronous Viterbi beam search algorithm (Pylkkönen, 2005).

Publications VII and VIII were submitted to the REVERB challenge (Kinoshita et al., 2013), and make use of the baseline recognizer built on the HTK hidden Markov model toolkit (Young, 1993). The further experiments carried out in Publication IX are based on the Kaldi toolkit (Povey et al., 2011).

The following speech corpora were used in the experimental evaluations.

[1]For example, the Finnish word ‘kahvi|n|juoja|lle|kin’, where the | symbols represent morpheme boundaries, corresponds to the English fragment “also for a drinker of coffee”.

Finnish SPEECON

The Finnish SPEECON corpus is part of the SPEECON project (Iskra et al., 2002) designed to aid in building speech interfaces for consumer devices. The corpus contains 550 adult speakers, each uttering various types of both prompted (read) and free speech. The corpus has been locally divided into training, development and evaluation subsets. While the acoustic model training set makes use of most utterance types, all test sets are based on phonetically rich sentences collected from the Internet. Each speaker read a sample of 30 sentences out of the total set of 4487, and no sentence was repeated more than 5 times.

The SPEECON corpus is particularly well suited to the evaluation of noise robust methods, as it contains large parts recorded in realistic noisy environments. The four defined recording environments are a quiet office, a living room, a public place and a car. With the exception of the office environment, all contain varying levels of background noise. Furthermore, each utterance has been simultaneously recorded at four different distances: using a headset microphone (channel 1), using a lapel microphone (channel 2), and using two far-field microphones approximately $0.5 - 1$ m and $1 - 2$ m away from the speaker (channels 3 and 4, respectively).

CHiME Challenge Corpora

The Computational Hearing in Multisource Environments (CHiME) challenges, organized in 2011 (Christensen et al., 2010) and 2013 (Vincent et al., 2013), provide a standard benchmark for noise robust speech processing. The data set consists of binaural recordings of speech, recorded with a dummy head, artificially mixed with highly nonstationary noise sources recorded in a real living room.

REVERB Challenge Corpora

The REverberant Voice Enhancement and Recognition Benchmark (REVERB) challenge, organized in 2014 (Kinoshita et al., 2013), focuses on the specific problem of reverberant speech. The data consists of 8-channel microphone array recordings, with selected channels designated for the evaluation of systems based on 1- and 2-channel input. The corpus includes separate data sets of artificially distorted speech based on measured room impulse responses and speech recorded in a real reverberant room. Both sets include a low level of mostly stationary background noise. The WSJCAM0 Wall Street Journal British English corpus (Robinson et al., 1995) is used as the source of clean speech material.

3. Noise Robust Features

The dividing line between a robust feature extraction scheme and a feature enhancement method for noisy speech is not firmly fixed, as both classes can be seen as an example of each other. For the purposes of this chapter, the discussion will be limited to the class of systems that Publication I and Publication II belong to, namely those that replace the Fourier magnitude spectrum with a more robust spectral envelope estimate in conventional speech recognition feature extraction. Prominent tools for this use are different variants of linear predictive (LP) modeling, and the minimum variance distortionless response (MVDR) spectrum.

3.1 Linear Predictive Models

The key concept in LP modeling is the assumption that the process generating the values of the output time domain signal $y(n)$ can be modeled well as a linear combination of the p preceding values of $y(n)$ and an input signal $x(n)$ (Makhoul, 1975):

$$y(n) = -\sum_{k=1}^{p} a_k y(n-k) + Gx(n), \tag{3.1}$$

where p is the *model order*, a_k are the *predictor coefficients*, and G is the *gain* of the input signal. This is an *all-pole model*, as the corresponding frequency domain transfer function $H(z)$ has the form

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \tag{3.2}$$

and therefore the transfer function has only poles and no zeros.

In a typical application, the input signal $x(n)$ is unknown, and the assumption is that the observed signal $y(n)$ can be predicted as

$$y(n) = -\sum_{k=1}^{p} a_k y(n-k) + e(n), \tag{3.3}$$

where the *residual* $e(n)$ is the error between the actual and predicted values of $y(n)$. The coefficients a_k are then commonly obtained by minimizing the total squared error E, where

$$E = \sum_n e(n)^2, \tag{3.4}$$

$$e(n) = y(n) + \sum_{k=1}^{p} a_k y(n-k). \tag{3.5}$$

The two main variants of LP analysis, the *autocorrelation method* and the *covariance method*, are distinguished by their particular choices for the summation index n.

In the autocorrelation method, the summation is conceptually over the infinite interval, $-\infty < n < \infty$. As generally only a finite portion of the signal is of interest, the signal is typically multiplied by a window function such that it is zero outside the interval $1 \leq n \leq N$. By denoting $\boldsymbol{a} = \begin{bmatrix} 1 & a_1 & \cdots & a_p \end{bmatrix}^\mathsf{T}$ and $\boldsymbol{y}_n = \begin{bmatrix} y(n) & \cdots & y(n-p) \end{bmatrix}^\mathsf{T}$, Equation (3.5) and Equation (3.4) for the autocorrelation method can then be concisely written as

$$e(n) = \boldsymbol{a}^\mathsf{T} \boldsymbol{y}_n, \tag{3.6}$$

$$E = \boldsymbol{a}^\mathsf{T} \left(\sum_{n=1}^{N+p} \boldsymbol{y}_n \boldsymbol{y}_n^\mathsf{T} \right) \boldsymbol{a} = \boldsymbol{a}^\mathsf{T} \boldsymbol{R} \boldsymbol{a}, \tag{3.7}$$

where $\boldsymbol{R} = \sum_{n=1}^{N+p} \boldsymbol{y}_n \boldsymbol{y}_n^\mathsf{T}$ is the *autocorrelation matrix* of signal $y(n)$. The values of $\boldsymbol{a}$ that minimize E are given by the system of equations

$$\boldsymbol{R}\boldsymbol{a} = \begin{bmatrix} E & 0 & \cdots & 0 \end{bmatrix}^\mathsf{T}. \tag{3.8}$$

When the LP coefficients $\boldsymbol{a}$ are computed with the autocorrelation method, the corresponding *synthesis filter* of Equation (3.2) is guaranteed to be stable (Makhoul, 1975). This property of LP analysis is critical in some applications, such as speech coding, where the synthesis filter is directly used to generate time-domain signals. Other applications that do not use the synthesis filter in this manner, such as spectral envelope estimation for ASR feature extraction, may tolerate potentially unstable filters, as discussed in Publication II.

The *source-filter model* of speech production (Fant, 1960) provides a physical justification for the use of linear prediction in speech signal analysis. In the model, human speech production apparatus is approximated as the combination of an excitation signal generated at the vocal chords and a linear filter implemented by the vocal tract, with resonant peaks

at the formant frequencies. The LP model is less general, as it restricts the filter to have no zeros, but still compatible with this view. The input signal $x(n)$ corresponds to the excitation, and the predictor coefficients a_k encode the vocal tract filter properties.

As the all-pole vocal tract filter assumption is relatively accurate for nonnasal voiced speech, and multiple poles can also approximate the antiresonances of unvoiced and nasal sounds, LP has been widely used for speech processing tasks such as formant and pitch analysis (Atal and Hanauer, 1971). Pitch analysis of voiced speech can be performed by inspecting the error residual $e(n)$ of Equation (3.3). Under the assumption that the vocal tract filter is adequately captured by the prediction coefficients, the residual corresponds to the excitation signal. For voiced sounds, the excitation has the form of a pulse train and the fundamental frequency can be estimated by locating the peaks of the residual.

While LP analysis is a versatile tool, some of its properties are less desirable in a speech processing context. The spectral approximation of the model is equally accurate at all frequencies. However, the frequency resolution and amplitude sensitivity depend strongly on the frequency. As a consequence, the LP coefficients may model perceptually irrelevant information at the cost of discarding more pertinent features of the source spectrum (Hermansky, 1990).

LP modeling is also not robust towards some types of input speech signals. Additive noise can strongly distort the modeled spectrum (Sambur and Jayant, 1976). In addition, the model is liable to follow the harmonic structure of the excitation signal instead of extracting the true spectral envelope corresponding to the effects of the vocal tract. This bias is especially noticeable for high-pitch speech, where the harmonic peaks are spaced more sparsely, or when the model order is increased (Makhoul, 1975).

Figure 3.1 compares the LP spectra of both low-pitch and high-pitch voiced speech to a spectral envelope estimate extracted using the MVDR method with the same, relatively high model order of 30. It can be noted that, especially for high-pitch speech, the MVDR model is less prone to modeling the harmonic structure, and yields a more reliable estimate of the spectral envelope.

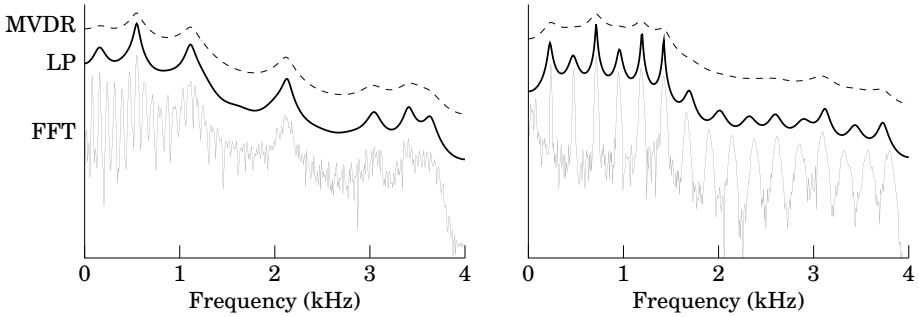


Figure 3.1. Fourier spectra and corresponding LP and MVDR spectral envelope estimates for both low-pitch (left) and high-pitch (right) realizations of the Finnish vowel /a/. The spectra are shown in logarithmic magnitude domain, with arbitrary normalization applied to isolate them. The same model order of 30 was used for both LP and MVDR models.

3.1.1 Perceptual Linear Prediction

Studies of human speech perception are an important source of inspiration for speech processing approaches. In the context of LP, the perceptual linear predictive (PLP) analysis method incorporates three major concepts: frequency-dependent spectral resolution and sensitivity, and the nonlinear relationship between sound intensity and perceived loudness (Hermansky, 1990).

Due to the physics of human hearing (Fletcher, 1940), when considered in the linear frequency scale, the spectral resolution of hearing is lower for high-frequency sounds, as the bandwidth of the masking effect caused by a pure tone at a particular frequency increases. The sound pressure level required to reach particular perceived loudness also varies as a function of frequency. Further, the relationship between the sound pressure level and perceived loudness at a given frequency is also not linear, and is typically modeled as a power law (Stevens, 1957).

PLP analysis begins with the short-time power spectrum, obtained by applying the discrete Fourier transformation (DFT) to a windowed speech segment. Conceptually, the spectral resolution of human hearing is accounted for by warping the frequency axis of the DFT power spectrum into the Bark frequency scale and convolving with a curve approximating the critical-band masking effect. The resulting spectrum is then sampled at approximately 1 Bark intervals, and each sample is pre-emphasized with an approximation of the equal-loudness curve to mimic the sensitivity of human hearing at different frequencies. Finally, the sample amplitudes

are compressed by taking the cube root, approximating the nonlinearity of perceived loudness. In practice, the effects of the frequency warping, convolution, pre-emphasis and compression are incorporated in a precomputed filterbank, similar to the mel-scale filterbank in MFCC feature extraction.

The final step in PLP analysis is closely analogous to the autocorrelation method of LP modeling. However, a modified autocorrelation matrix $\boldsymbol{R}$ is used in Equation (3.7), obtained as the inverse DFT of the PLP filterbank outputs.

The choice of the LP model order strongly affects the behavior of PLP analysis (Hermansky, 1990). In phoneme and isolated word recognition experiments, with a speaker mismatch between the training and test data, the optimal model order for PLP is very low: a 5th or 6th order PLP model outperforms conventional LP analysis, while increasing the PLP model order actually degrades the performance. When matched speaker data is used, accuracy increases with model order for both PLP and conventional LP. However, PLP reaches close to optimal accuracy already at the 6th order, after which further increase in the model order yields only minor improvement in accuracy. This suggests that the overall shape of the spectrum captured by a low-order PLP model successfully captures most of the speaker-independent information about the linguistic content of the signal.

3.1.2 Weighted Linear Prediction

The key insight of weighted linear prediction (WLP) is to apply a temporal weight function to the prediction error residual. As proposed by Ma et al. (1993), the motivation for the weighting is to improve the modeling of clean speech signals by focusing on a subset of the samples based on a model of speech production. In the context of noisy signals, the weights make it possible to suppress the effect of noise on the final predictor coefficients, and focus on the less noisy regions of the signal (Magi et al., 2009).

To incorporate the weight function into LP analysis, the cost function E of Equation (3.4) is modified to include a temporal weight term w_n,

$$E_w = \sum_n w_n e(n)^2. \tag{3.9}$$

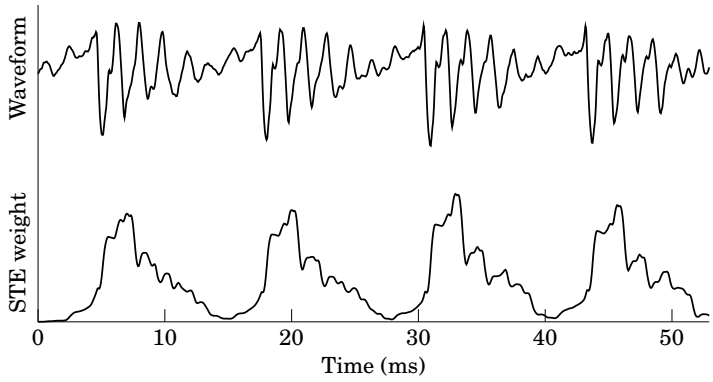


Figure 3.2. Four pitch periods of clean speech waveform of the Finnish vowel /a/ (top), and the corresponding STE weight function with a window size of 2.5 ms (bottom).

The matrix form corresponding to Equation (3.7) is then

$$E_w = \boldsymbol{a}^\mathsf{T} \left(\sum_{n=1}^{N+p} w_n \boldsymbol{y}_n \boldsymbol{y}_n^\mathsf{T} \right) \boldsymbol{a} = \boldsymbol{a}^\mathsf{T} \boldsymbol{R}_w \boldsymbol{a}, \tag{3.10}$$

where $\boldsymbol{R}_w = \sum_{n=1}^{N+p} w_n \boldsymbol{y}_n \boldsymbol{y}_n^\mathsf{T}$ is a weighted autocorrelation matrix. Notably, for a constant weight function $w_n = w$, WLP analysis reduces to conventional LP.

Ma et al. (1993) propose a weight function based on the *short-time energy* (STE):

$$w_n = \sum_{i=1}^{M} y(n-i)^2, \tag{3.11}$$

where M is a window size parameter controlling the sharpness of the STE function. Considering a general noisy signal $y(n)$, the use of STE weights can be justified by noting that the weight function emphasizes those regions of the signal where the samples are high in amplitude, and therefore less affected by additive noise. Additionally, in the particular case of voiced speech, the formant resonances induced by the shape of the vocal tract are most prominently expressed during the closed stage of the glottal cycle (Wong et al., 1979), which the STE weight function approximately coincides with. The STE weight function for a clean voiced speech segment, with a window size of $M = 40$ samples (2.5 ms), is illustrated in Figure 3.2.

Unlike autocorrelation LP analysis, the synthesis filter corresponding to the WLP solution is not guaranteed to be stable. If a stable filter is required for the application of interest, the stabilized WLP (SWLP) algorithm proposed by Magi et al. (2009) can be used. The SWLP stabiliza-

tion process modifies the weighted autocorrelation matrix $\boldsymbol{R}_w$ of Equation (3.10) to enforce the filter stability, while having only a moderate effect on the resulting spectral model.

Experimental Results and Discussion

Publication I of this thesis investigates the applicability of SWLP analysis with the STE weight function for speech recognition feature extraction. Experiments were performed on the car and public place data sets of the Finnish SPEECON corpus. With careful tuning of the STE window width parameter M, improved letter error rates were obtained for noisy speech when using acoustic models trained on clean speech data. Considerable gains were also noted in an oracle experiment where the M value was selected individually for each utterance, based on the reference transcription, to minimize the letter error rate.

As the stability of the filter corresponding to the LP coefficients is not of primary importance for the ASR application, Publication II extends the work by including also MFCC features based on unstabilized WLP. For this application, WLP was found both to perform better and to be less sensitive to the window width parameter M.

The overall relative letter error rate improvements of the linear predictive systems proposed in both publications are summarized in Table 3.1. The numbers were computed over a joint evaluation data set of both car and public place noise from the SPEECON corpus. Details of the speech recognition system and the corpus, including the definitions of the recording channels, are given in Section 2.5. The WLP-MFCC system is seen to outperform the other systems for the noisy data of channels 2 and 3, recorded with a lapel and a far-field microphone, respectively. However, the increase in performance is obtained at the cost of more severe degradation of results for the close talk signal of channel 1, which corresponds in practice to clean speech.

In the oracle experiment of per-utterance optimization of the M parameter presented in Publication I, compared with a conventional MFCC baseline system, relative improvements of 17.3%, 22.2% and 29.6% in the letter error rate were achieved by the LP-MFCC system, the SWLP-MFCC system with the best fixed M value, and the SWLP-MFCC system with the best per-utterance M setting, respectively. This experiment was performed on channel 3 of the SPEECON development set data, which has similar properties but is not directly comparable to channel 3 of the eval-

	Recording channel		
System	1	2	3
LP-MFCC	−1.7	9.8	17.3
WLP-MFCC	−32.0	18.1	23.7
SWLP-MFCC	−0.7	11.1	20.6

Table 3.1. Relative improvement (in percents) over a baseline MFCC system in the measured letter error rate for each of the systems proposed in Publication I and Publication II for the SPEECON corpus (Section 2.5). Recording channels 1, 2 and 3 correspond to a close-talk, lapel and far-field microphone, respectively.

uation data set used for Table 3.1. The automatic adaptation of the M parameter based on the features of the input signal and feedback from the speech recognition system remains a potential topic for future work. Instead of being limited to a single M value per utterance, such methods could update M more frequently within the utterance, possibly outperforming even the oracle experiment.

Further Extensions

The temporal weighting of the LP error residual can be further generalized by allowing either a unique weight $z_{n,k}$ for each sample position n and lag k (Pohjalainen et al., 2010), or even a weight $q_{n,i,j}$ for each individual element i, j in each of the terms $\boldsymbol{y}_n \boldsymbol{y}_n^{\mathsf{T}}$ making up the autocorrelation matrix (Pohjalainen and Alku, 2013). In the first case, the extended weighted linear prediction (XLP) cost function then has the form

$$E_{\mathrm{XLP}} = \sum_n \left(z_{n,0} y(n) - \sum_{k=1}^{p} z_{n,k} a_k y(n-k) \right). \tag{3.12}$$

The SWLP filter stabilization method is equally applicable to this form of XLP, yielding a stabilized XLP (SXLP) variant.

Conventional WLP and LP are obtained as special cases of XLP when the weight is constant for each sample ($z_{n,k} = w_n$) or overall ($z_{n,k} = w$), respectively. The later *autocorrelation snapshot* formulation of Pohjalainen and Alku (2013) is a further generalization, containing regular XLP as a special case when the weight function $q_{n,i,j}$ factorizes as $q_{n,i,j} = z_{n,i} z_{n,j}$.

In order to take advantage of the added flexibility, Pohjalainen et al. (2010) propose a new weight function, *absolute value sum* (AVS):

$$z_{n,k} = \frac{M-1}{M} z_{n-1,k} + \frac{1}{M} \left(|y(n)| + |y(n-k)| \right), \tag{3.13}$$

$$z_{0,k} = 0. \tag{3.14}$$

The AVS weight function is analogous to the STE weights in that it emphasizes signal samples with high amplitudes, but additionally, within each prediction, it further emphasizes those delays where the corresponding sample had a large amplitude. The M parameter controls the length of the moving average memory. It has a similar role as the M parameter of the STE weight function.

Alternative weighting schemes for both variants of XLP have been proposed by Pohjalainen and Alku (2012, 2013). The extended linear prediction framework has shown improved performance in noisy conditions for both speaker verification (Pohjalainen et al., 2010) and speech recognition (Keronen et al., 2011) tasks.

3.2 Minimum Variance Distortionless Response

The minimum variance distortionless response (MVDR) spectrum, also known as the *Capon spectrum*, was originally proposed by Capon (1969) for use in the signal analysis of sensor arrays. It has later been widely applied for all-pole modeling of speech spectra (Murthi and Rao, 2000).

Conceptually, the value of a Mth order MVDR spectrum at frequency ω_k is given by the output power of a Mth order FIR filter $h_k(n)$. The filter is designed to minimize the output power (hence, *minimum variance*), given the observed input signal, under the constraint that it has unity gain at frequency ω_k (the *distortionless response* criterion). Denoting $\boldsymbol{\omega}_k = \begin{bmatrix} 1 & e^{j\omega_k} & e^{j2\omega_k} & \cdots & e^{jM\omega_k} \end{bmatrix}^{\mathsf{T}}$ and $\boldsymbol{h}_k = \begin{bmatrix} h_k(0) & h_k(1) & \cdots & h_k(M) \end{bmatrix}^{\mathsf{T}}$, $\boldsymbol{h}_k$ is therefore given by the solution of the optimization problem

$$\min_{\boldsymbol{h}_k} \boldsymbol{h}_k^{\mathsf{H}} \boldsymbol{R} \boldsymbol{h}_k \quad \text{subject to } H_k(e^{j\omega_k}) = \boldsymbol{\omega}_k^{\mathsf{H}} \boldsymbol{h}_k = 1, \tag{3.15}$$

where $\boldsymbol{R}$ is the $(M+1) \times (M+1)$ autocorrelation matrix.

The formulation of the MVDR spectrum as the output of a filterbank of distortionless filters is conceptually alike to periodogram-based spectrum estimation methods such as FFT. The major advantage of MVDR is that the individual filters are not fixed, but allowed to depend both on the input data as well as the target frequency of the filter. As with linear prediction, increasing the model order M increases the modeling power, but the MVDR method produces a more robust spectral envelope estimate that is less prone to modeling the harmonic structure of speech.

Notably, however, there is no need to explicitly design the filters $h_k(n)$, as the MVDR spectrum at any frequency ω can be computed as (Murthi

and Rao, 2000)

$$P_{\mathrm{MVDR}}(\omega) = \frac{1}{\boldsymbol{\omega}^{\mathsf{H}} \boldsymbol{R}^{-1} \boldsymbol{\omega}}. \tag{3.16}$$

An alternative parametrization can also be used (Musicus, 1985), where

$$P_{\mathrm{MVDR}}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu_k e^{-jk\omega}}, \tag{3.17}$$

and the coefficients μ_k can be obtained directly from the LP coefficients a_k and prediction error variance E of Equation (3.4) and Equation (3.5) as

$$\mu_k = \begin{cases} \frac{1}{E} \sum_{l=0}^{M-k} (M+1-k-2l) a_l a_{l+k}^*, & \text{for } k = 0, \ldots, M, \\ \mu_{-k}^*, & \text{for } k = -M, \ldots, -1. \end{cases} \tag{3.18}$$

Properties of the human speech perception, such as those used by PLP modeling and MFCC feature extraction, can be also incorporated directly in MVDR-based spectral envelope estimation methods. Perceptual MVDR-based cepstral coefficients (PMCCs), similarly to PLP, are based on performing the all-pole modeling using a modified autocorrelation function, derived from the outputs of a perceptually motivated mel-scale filterbank (Dharanipragada et al., 2007). Perceptual-MVDR (PMVDR) accomplishes the same task by direct warping of the Fourier spectrum with a first-order all-pass filter (Yapanel and Hansen, 2008). The warping parameter α of PMVDR can be used to approximate both the mel and Bark frequency scales.

4. Feature Enhancement of Noisy Speech

Methods of reducing the effect of noise on speech features can be used to enhance speech both for human and machine listeners. This chapter considers the case of feature enhancement used in the front-end of an ASR system. The concepts covered are those employed by the publications in this thesis: missing data approaches (Publication IV, Publication V) and non-negative matrix factorization (Publication III, Publication VI). *Sparse imputation*, which applies the tools of non-negative matrix factorization in a missing data framework, is also discussed.

4.1 Missing Data Methods

The missing data framework (Cooke et al., 1994, 2001) is inspired by the observation that human speech recognition is remarkably robust against missing information: speech remains intelligible to humans even if large parts of the signal are deliberately removed, as long as the removal is either spectrally or temporally localized. While the common scenario of additive environmental noise does not in itself involve missing data, studies on human hearing show that noise sources can entirely *mask* out the spectro-temporal components they dominate, as far as further auditory processing is concerned (Moore, 2012).

Several variations on the idea of modeling these effects in a speech processing system have been proposed (Raj and Stern, 2005). A common approach is to estimate a spectro-temporal mask, denoting which components of the input signal contain reliable information originating from the sound source of interest. Depending on how it will be used, the estimated mask can be a *binary mask* which labels certain components as entirely missing, or a *soft mask* which is able to represent the notion of partially obstructed components. In both cases, the method of estimating the mask

is generally critical to the performance of the resulting system, and can be based on application-specific assumptions of the nature of the speech and noise sources present in the signal.

After identifying the unreliable regions, the speech recognition system must be modified to disregard them from the input. The two most common approaches for this are *marginalization* and *imputation* (Raj and Stern, 2005). In marginalization, the back-end decoder is modified to ignore the elements. In imputation, by contrast, the missing elements are replaced with a best estimate of their underlying values, based on the surrounding context of reliable data, and passed to a conventional speech recognition back-end. In both cases, under the additive noise assumption, the actual observed value is often used as supplementary information about the upper bound of the true value. Continuing the theme of front-end based approaches, the publications related to missing data processing in this thesis focus on imputation-based methods exclusively.

Publication IV presents a missing data system using binary masks and sparse imputation. The work compares masks based on binaural cues derived from stereo data (Harding et al., 2006) with masks generated by a support vector machine (SVM) classifier based on a number of single-frame features (Gemmeke et al., 2009). Observation uncertainty estimates, discussed in detail in Chapter 5, are used to improve the speech recognition accuracy further. Publication V extends the classifier-based mask construction by considering a wide set of both monaural and binaural acoustic features as inputs to a Gaussian mixture model (GMM) based classifier. An alternative missing data imputation method using GMM clustering (Raj et al., 2004) is also evaluated.

The following sections describe these concepts in more detail.

4.1.1 Classifier-based Missing Data Mask Generation

The task of estimating a binary mask for missing data processing can be quite naturally formulated as a two-class classification problem: given information about the speech signal, a particular spectro-temporal component can be classified as being dominated by the target speech signal or noise. Of the many binary classification algorithms available, the publications in this thesis have applied the GMM and SVM classifiers.

An L-component GMM is defined as a weighted mixture of L multivari-

ate Gaussian distributions, with the probability density function

$$p(\boldsymbol{x} \mid \boldsymbol{\Gamma}) = \sum_{l=1}^{L} w_l \, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{4.1}$$

where $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of the multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The GMM parameter set $\boldsymbol{\Gamma} = \{w_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l \mid l = 1, \ldots, L\}$ consists of the scalar *mixture weights* w_l and the mean vectors $\boldsymbol{\mu}_l$ and covariance matrices $\boldsymbol{\Sigma}_l$ for each of the component distributions. The weights are constrained as $\sum_{l=1}^{L} w_l = 1$, and the number of parameters in the model can be reduced by restricting the covariance matrices $\boldsymbol{\Sigma}_l$ to be diagonal or even spherical, or tied between the mixture components. The expectation-maximization (EM) algorithm can be used to efficiently learn the maximum likelihood (ML) estimate of the parameters $\boldsymbol{\Gamma}$ for a given data set.

A GMM-based classifier for missing data mask generation can be constructed as follows (Seltzer et al., 2004a; Publication V). A feature vector $\boldsymbol{x}(t, c)$ is associated with each spectrogram element, for the frequency band c of time frame t. For a paired training data set where both clean and noisy signals are known, *oracle masks* are generated, based on the SNR within each element, with the elements having sufficiently high SNR labeled as reliable. For each frequency band c, two GMM parameter sets $\boldsymbol{\Gamma}_{c,\mathrm{r}}$ and $\boldsymbol{\Gamma}_{c,\mathrm{u}}$ are computed as the ML estimates, given, respectively, the reliable and unreliable samples in the training data as classified by the oracle masks. Given a feature vector $\boldsymbol{x}(t, c)$ of an element in an observed spectrogram, the element is then classified as reliable if

$$K \, p(\boldsymbol{x}(t, c) \mid \boldsymbol{\Gamma}_{c,\mathrm{r}}) > p(\boldsymbol{x}(t, c) \mid \boldsymbol{\Gamma}_{c,\mathrm{u}}), \tag{4.2}$$

where K is a scaling factor optimized using a development data set. The purpose of the scaling factor is to account for the unequal relative costs of false negatives and positives in the mask estimation, which depend on the particulars of the missing feature method in use. The special case of $K = 1$ corresponds to conventional maximum-likelihood classification with equal prior probabilities.

Alternative binary classification methods, such as SVM classifiers, can be used for missing data mask generation in a manner analogous to the GMM-based classifier described above (Gemmeke et al., 2009; Keronen, Cho, Raiko, Ilin and Palomäki, 2013; Publication IV).

For accurate classification, the output classes must be well separable in the space defined by the input features of the classifier. While a larger

set of input features generally contains more information about the signal, adding irrelevant data can increase the computational cost of classification, and, depending on the classifier, also reduce the classification accuracy. The selected feature set should therefore be concise but informative with respect to whether the corresponding element is dominated by speech or noise.

Publication V defines and analyzes a set of 14 acoustic features, both monaural and binaural, suitable for missing data mask estimation. Alternatively, neural network models such as GRBM can be used to perform unsupervised learning of efficient feature representations from existing data. This approach is used for mask estimation and described in more detail by Keronen, Remes, Kallasjoki and Palomäki (2013); Keronen, Cho, Raiko, Ilin and Palomäki (2013).

4.1.2 Gaussian Mixture Models for Missing Data Imputation

In order to take advantage of context provided by temporally adjacent spectrogram frames in missing data imputation, the input data is processed in *windows* of T contiguous spectrogram frames. The noisy observation window $\boldsymbol{y}$ is represented as a TC-dimensional vector, where C is the number of frequency bands in the (typically mel-scale) spectrogram representation. The corresponding, unknown clean speech window is denoted by $\boldsymbol{s}$. Given an estimated binary mask over the window, the elements of both vectors can be divided into two disjoint subsets of reliable and unreliable components, denoted by $\boldsymbol{y}_\mathrm{r}$, $\boldsymbol{y}_\mathrm{u}$, $\boldsymbol{s}_\mathrm{r}$ and $\boldsymbol{s}_\mathrm{u}$, respectively. For the reliable components, assuming that the mask is accurate, the observation itself can be used as the reconstruction by setting $\boldsymbol{s}_\mathrm{r} = \boldsymbol{y}_\mathrm{r}$. The unreliable components $\boldsymbol{s}_\mathrm{u}$ must be estimated by imputation. If the processing windows overlap (i.e., a window step less than T is used), the final imputation results for the utterance can be obtained by averaging over all overlapping estimates.

Clustering of clean speech data can be used as a basis for the imputation step (Raj et al., 2004; Raj and Stern, 2005; Publication V). To apply the cluster-based imputation, the clean speech $\boldsymbol{s}$ is assumed to be sampled from an L-component full covariance GMM. The parameter set $\boldsymbol{\Gamma}$ of the GMM is estimated from clean speech training data. The resulting mixture model can be seen as a clustering of the data into L Gaussian clusters. Given the known value of the reliable components and the upper bound provided by the observation, a bounded maximum *a posteriori*

(MAP) estimate $\hat{s}_l$ of the clean speech can be computed for each cluster l. This estimate is defined as

$$\hat{\boldsymbol{s}}_l = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s} \mid \boldsymbol{s}_\mathrm{r} = \boldsymbol{y}_\mathrm{r}, \boldsymbol{s}_\mathrm{u} \leq \boldsymbol{y}_\mathrm{u}, \boldsymbol{\Gamma}_l), \tag{4.3}$$

where $\boldsymbol{\Gamma}_l$ denotes the mean and covariance parameters of cluster l. A combined estimate can be obtained as a linear combination of the cluster-specific estimates,

$$\hat{\boldsymbol{s}} = \sum_{l=1}^{L} p(l \mid \boldsymbol{s}_\mathrm{r} = \boldsymbol{y}_\mathrm{r}, \boldsymbol{s}_\mathrm{u} \leq \boldsymbol{y}_\mathrm{u}, \boldsymbol{\Gamma})\, \hat{\boldsymbol{s}}_l, \tag{4.4}$$

where $p(l \mid \boldsymbol{s}_\mathrm{r} = \boldsymbol{y}_\mathrm{r}, \boldsymbol{s}_\mathrm{u} \leq \boldsymbol{y}_\mathrm{u}, \boldsymbol{\Gamma})$, the posterior probability of cluster l, is computed by an iterative process, as described in Raj and Stern (2005).

As an alternative to the cluster-based imputation method described above, Remes et al. (2015) define a bounded conditional mean imputation (BCMI) scheme, where the posterior distribution of clean speech, $p(\boldsymbol{s} \mid \boldsymbol{y}_\mathrm{r})$, is approximated by a single Gaussian distribution, which is then truncated based on the upper bound $\boldsymbol{s}_\mathrm{u} \leq \boldsymbol{y}_\mathrm{u}$. Point estimates of the mean of the posterior distribution can be used for conventional missing data reconstruction, but the full posterior is also available for computing observation uncertainty information. This method is used for missing data imputation by Keronen, Remes, Kallasjoki and Palomäki (2013).

4.2 Non-negative Matrix Factorization

In non-negative matrix factorization (NMF), given a $D \times N$ matrix $\boldsymbol{Y}$ with non-negative elements, we are interested in finding suitable $D \times K$ matrix $\boldsymbol{S}$ and $K \times N$ matrix $\boldsymbol{A}$, also non-negative, such that the product

$$\boldsymbol{S}\boldsymbol{A} \approx \boldsymbol{Y}. \tag{4.5}$$

The details depend on the application in question. For example, K may be constrained to be small compared to the size of $\boldsymbol{Y}$ in order to find a less complex representation of $\boldsymbol{Y}$. Alternatively, K may be large but $\boldsymbol{A}$ constrained to be sparse.

For the publications in this thesis, $\boldsymbol{S}$ is a *dictionary*, or *basis* matrix of K *atoms*, or *exemplars*: fixed-length samples of the kinds of sounds modeled by the dictionary. Each atom is a $C \times T$ mel-scale magnitude spectrogram, having C frequency bands and T time frames. Individual spectrograms are stacked to column vectors of length $D = CT$ to form the

matrix $\boldsymbol{S}$. Similarly, the $\boldsymbol{Y}$ matrix represents an observed sound signal. To model an observation of arbitrary length using atoms of fixed size, a sliding window technique is used. Groups of T consecutive spectrogram frames, in overlapping windows, are stacked to form the columns of the $D \times N$ matrix $\boldsymbol{Y}$. Finally, each column of the *activation* matrix $\boldsymbol{A}$ contains the weights needed to approximate the corresponding column of the observation $\boldsymbol{Y}$ with a weighted sum of the spectrogram atoms in the dictionary $\boldsymbol{S}$.

Given a fixed observation $\boldsymbol{Y}$, efficient iterative algorithms exist for performing the factorization (Lee and Seung, 2001). The $\boldsymbol{S}$ and $\boldsymbol{A}$ matrices can be jointly estimated to minimize $d(\boldsymbol{SA}, \boldsymbol{Y})$, where d is, e.g., the Euclidian distance, or the Kullback-Leibler divergence. The dictionary matrix $\boldsymbol{S}$ can also be either fully or partially fixed, as a way to incorporate prior knowledge about the expected sound sources in the system. The representation of the observation in the activation matrix $\boldsymbol{A}$ can be made sparse by using a cost function with a penalty term for non-zero activation weights.

The stacked spectrogram vectors permit the use of atoms with temporal structure with the standard NMF algorithms. Alternatively, the NMF optimization algorithms can be extended to allow the direct use of atoms with temporal context, as is done in non-negative matrix factor deconvolution (Smaragdis, 2000) or non-negative tensor factorization (Fitzgerald et al., 2005; Barker and Virtanen, 2013). For some applications, the special case of purely spectral atoms, where $T = 1$, may also suffice.

This section discusses the application of NMF to source separation, as used in Publication III and Publication VI. Section 4.3 considers the same model in the context of missing data imputation, while Section 6.2 extends the model to account for the effects of reverberation.

4.2.1 Source Separation

If the sound sources are assumed to mix approximately additively, a straightforward way to apply NMF methods to source separation is to associate each dictionary atom with a particular sound source. For the case of a noisy speech signal considered in this work, combining a clean speech dictionary $\boldsymbol{S}_\mathrm{s}$ of size K_s and a noise dictionary $\boldsymbol{S}_\mathrm{n}$ of size K_n yields the model

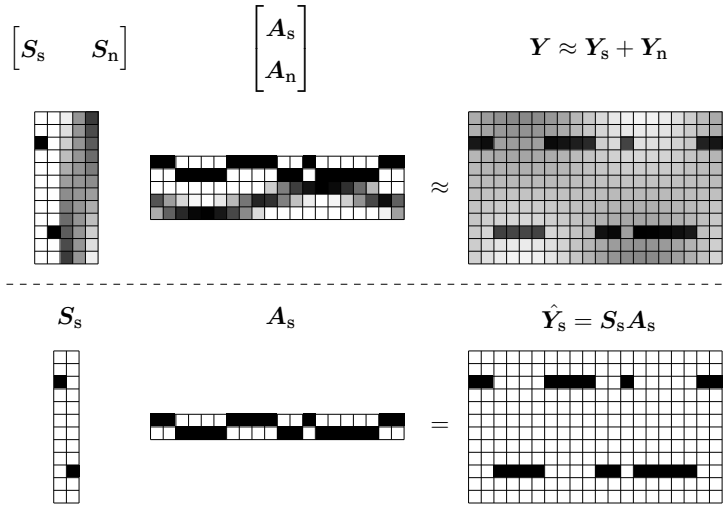


Figure 4.1. A toy example demonstrating the signal model for the NMF-based source separation. For simplicity of visualization, the special case of $T = 1$, when the dictionary atoms are simple spectra, is used. Large values in the matrices are indicated by a dark shade. In this example, a signal formed of two pure tones, represented by the two columns of the matrix $\boldsymbol{S}_\mathrm{s}$, is observed over a background of additive noise with a varying spectral tilt. The noise dictionary $\boldsymbol{S}_\mathrm{n}$ contains three atoms, and the smoothly changing background in the observation $\boldsymbol{Y}$ can be represented with their weighted combination. The bottom matrices illustrate the clean signal reconstruction in Equation (4.11).

(Gemmeke et al., 2011)

$$\boldsymbol{Y} \approx \boldsymbol{Y}_\mathrm{s} + \boldsymbol{Y}_\mathrm{n} \tag{4.6}$$

$$\approx \boldsymbol{S}_\mathrm{s}\boldsymbol{A}_\mathrm{s} + \boldsymbol{S}_\mathrm{n}\boldsymbol{A}_\mathrm{n} \tag{4.7}$$

$$= \begin{bmatrix} \boldsymbol{S}_\mathrm{s} & \boldsymbol{S}_\mathrm{n} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_\mathrm{s} \\ \boldsymbol{A}_\mathrm{n} \end{bmatrix} = \boldsymbol{S}\boldsymbol{A}, \tag{4.8}$$

where $\boldsymbol{Y}$ is the stacked mel-spectrogram representation of the noisy observation, and $\boldsymbol{Y}_\mathrm{s}$ and $\boldsymbol{Y}_\mathrm{n}$ are the unknown clean speech and noise components, respectively. Figure 4.1 illustrates this model.

The combined activation matrix $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_\mathrm{s} & \boldsymbol{A}_\mathrm{n} \end{bmatrix}^\mathsf{T}$ in Equation (4.8) is obtained by solving, for each column $\boldsymbol{a}$ of $\boldsymbol{A}$ and the corresponding column $\boldsymbol{y}$ of $\boldsymbol{Y}$, the constrained optimization problem

$$\min_{\boldsymbol{a}} d(\boldsymbol{y}, \boldsymbol{S}\boldsymbol{a}) + \|\boldsymbol{\lambda} \odot \boldsymbol{a}\|_1 \quad \text{s.t. the elements of } \boldsymbol{a} \text{ are non-negative}, \tag{4.9}$$

where d is the Kullback-Leibler divergence, and the $\odot$ operator denotes elementwise multiplication of the activations with a *sparsity coefficient* vector $\boldsymbol{\lambda}$.

The sparsity coefficients can be used to adjust both the overall level of sparsity of the activations and the relative costs of using the speech and

noise dictionaries. Accordingly, in the publications of this thesis, $\boldsymbol{\lambda}$ has the form

$$\boldsymbol{\lambda} = \left[\underbrace{\lambda \ \cdots \ \lambda}_{K_\mathrm{s}} \ \ \underbrace{\lambda_r\lambda \ \cdots \ \lambda_r\lambda}_{K_\mathrm{n}} \right], \tag{4.10}$$

where the λ parameter controls the overall sparsity, and λ_r is the relative cost for using the noise dictionary atoms, compared to the speech dictionary.

Reconstruction of the speech and noise components from the factorization provides simple estimates for the clean speech and noise components,

$$\hat{\boldsymbol{Y}}_\mathrm{s} = \boldsymbol{S}_\mathrm{s}\boldsymbol{A}_\mathrm{s}, \tag{4.11}$$

$$\hat{\boldsymbol{Y}}_\mathrm{n} = \boldsymbol{S}_\mathrm{n}\boldsymbol{A}_\mathrm{n}. \tag{4.12}$$

Averaging across the overlapping time frames of the stacked representation can be used to produce corresponding single spectrograms $\bar{\boldsymbol{Y}}_\mathrm{s}$ and $\bar{\boldsymbol{Y}}_\mathrm{n}$.

However, if the observed signal does not have a good match with the specified dictionaries, the accuracy of the direct reconstruction $\bar{\boldsymbol{Y}}_\mathrm{s}$ may be low. In addition, fine details in the noisy spectrogram are not fully modeled by the sparse representation in matrix $\boldsymbol{A}$, and are further smoothed by the averaging over time in the reconstructions $\bar{\boldsymbol{Y}}_\mathrm{s}$ and $\bar{\boldsymbol{Y}}_\mathrm{n}$. In practice, better results are obtained by using the filtered estimate $\tilde{\boldsymbol{Y}}_\mathrm{s}$, defined as

$$\left[\tilde{\boldsymbol{Y}}_\mathrm{s}\right]_{c,t} = \frac{\left[\bar{\boldsymbol{Y}}_\mathrm{s}\right]_{c,t}}{\left[\bar{\boldsymbol{Y}}_\mathrm{s}\right]_{c,t} + \left[\bar{\boldsymbol{Y}}_\mathrm{n}\right]_{c,t}} \left[\bar{\boldsymbol{Y}}\right]_{c,t}, \tag{4.13}$$

where $\bar{\boldsymbol{Y}}$ is the original noisy mel-spectrogram. This amounts to mel-spectral domain Wiener-like filtering with a filter based on the reconstructed speech and noise components $\bar{\boldsymbol{Y}}_\mathrm{s}$ and $\bar{\boldsymbol{Y}}_\mathrm{n}$.

4.2.2 Dictionary Construction

When some prior information about all the components of the noisy mixture is available, a fixed dictionary can be used. The publications of this thesis are based on using *overcomplete* dictionaries, constructed by selecting, at random, a sufficiently large number of atoms — in the order of ten thousand — from sample recordings to cover all the patterns that are present in the data. Despite the number of atoms, the resulting signal representations in terms of activations remain sparse, due to the activation cost term in Equation (4.9). While the large size of the dictionary does raise the computational cost of standard NMF factorization algo-

rithms, specific methods exist for efficiently utilizing large dictionaries (Gemmeke and Van hamme, 2011; Virtanen et al., 2014).

In many real applications, it is infeasible to collect a representative sample containing all the types of noise that may be encountered. Several approaches have been proposed to account for noise types not present in the training data (Gemmeke, Remes and Palomäki, 2010; Hurmalainen et al., 2013). The noise dictionary, or a subset of its atoms, can be jointly estimated with the activation matrix during the NMF factorization. Alternatively, a synthetic noise dictionary consisting of atoms with a single frequency component can be used. Both of these methods require the match between the observed speech and the clean speech dictionary to be relatively good, as they essentially attribute all information in the signal not represented by the speech atoms to the noise component. Finally, the noise dictionary can be augmented online by including samples of observed non-speech segments as atoms, which is effective at capturing unseen but stationary noise sources. Publication VI takes advantage of this *noise sniffing* approach.

4.3 Sparse Imputation

The sparse dictionary-based representations produced by NMF processing can also be applied within the missing data framework. The general concept of *sparse imputation* (Gemmeke, Van hamme, Cranen and Boves, 2010) is to obtain the activation coefficients based on the portion of the observation that is considered reliable. The reconstruction as a linear combination of the selected dictionary atoms can then be used to provide estimated values for the unreliable elements of the observation.

Essentially, the model for the clean speech is identical to that of Section 4.2.1, with two exceptions: the dictionary contains only clean speech atoms, and some of the elements of the observation matrix $\boldsymbol{Y}$ are considered missing or unreliable. As a consequence, when solving the optimization problem of Equation (4.9), the unreliable elements are ignored. In practice, this can be done by removing the unreliable elements from the observation column vector $\boldsymbol{y}$, and the corresponding rows from the dictionary matrix $\boldsymbol{S}$.

Solving the modified Equation (4.9) yields a complete activation matrix $\boldsymbol{A}$. A reconstructed spectrogram $\bar{\boldsymbol{Y}}$ can then be defined by averaging the product $\boldsymbol{SA}$ across the overlapping windows. Finally, the unreliable el-

	SNR (dB)					
System	9	6	3	0	−3	−6
Baseline	**83.3**	**77.9**	**67.3**	52.9	42.2	36.8
ILD/ITD	75.8	71.9	64.8	**55.1**	**49.5**	**43.1**
SVM	76.5	69.4	54.6	44.4	37.6	34.6
Oracle	92.7	93.1	90.3	90.6	89.3	88.4

Table 4.1. Keyword accuracy rates (%) obtained using sparse missing data imputation for the CHiME 2011 development set in Publication IV. The baseline system performs no feature enhancement processing. The evaluated systems use the Harding et al. (2006) ILD/ITD-based mask ("ILD/ITD"), a mask generated by an SVM classifier based on 7 monaural acoustic features ("SVM"), or the oracle SNR mask constructed using knowledge of the true underlying clean speech ("Oracle").

ements of the original observation can be filled in from $\bar{Y}$ whenever the value is smaller than the observation. This elementwise upper bound reflects the additive noise assumption, under which the true clean speech elements can never be greater than those of the observation.

4.4 Experimental Results and Discussion

Missing Data Methods

Publications IV and V of this thesis apply missing data methods for improving speech recognition performance. In Publication IV, two mask estimation methods are compared: the Harding et al. (2006) binary mask estimation based on interaural level and arrival time differences (ILD/ITD), and the classifier-based mask estimation described in Section 4.1.1 using an SVM classifier and a set of 8 monaural acoustic features. Feature reconstruction is performed using the sparse imputation scheme outlined in Section 4.3 (Gemmeke, Van hamme, Cranen and Boves, 2010). All evaluations are done within the context of the first (2011) CHiME challenge corpus (Christensen et al., 2010).

Table 4.1 lists the CHiME keyword accuracy for the different mask estimation methods. In these experiments, the ILD/ITD-based binaural mask clearly outperforms the mask based on the SVM classifier using monaural features only, at all tested SNR levels. Even with the better mask, however, missing data imputation is beneficial only for the very noisy scenarios having an SNR of 0 dB or below. While oracle masks are unsuitable

	SNR (dB)						
System	9	6	3	0	−3	−6	Avg.
Baseline	86.3	78.3	68.5	53.9	44.3	41.9	62.2
ILD/ITD	88.6	79.8	70.8	58.9	47.4	46.3	65.3
ILD/ITD GMM	88.5	83.2	73.5	63.6	54.9	48.6	68.7
Full GMM	**90.3**	84.3	76.9	68.2	58.2	56.3	72.3
MC+A	89.6	**86.7**	**83.2**	**75.4**	**65.7**	**62.3**	**77.1**

Table 4.2. Keyword accuracy rates (%) obtained using GMM cluster-based missing data imputation on the CHiME 2011 evaluation set in Publication V. The baseline system performs no feature enhancement processing. The evaluated systems use the Harding et al. (2006) ILD/ITD-based mask ("ILD/ITD"), a mask generated by a GMM classifier with ILD/ITD pairs as the input feature ("ILD/ITD GMM"), or a GMM classifier using a 14-feature set of both binaural and monaural acoustic features ("Full GMM"). Further improvement over the last system is obtained by using multi-condition training on the imputed features and speaker adaptation ("MC+A").

for real applications, as they are based on knowledge of the underlying clean speech signal, the significantly higher performance obtained by using them shows that missing data imputation can be highly efficient when accurate masks can be estimated.

Publication IV also investigates the use of uncertainty estimates for the reconstructed features to further improve the recognition performance. These methods and results are discussed in Chapter 5.

In Publication V, the classifier-based mask construction approach is investigated in more depth. A set of 14 monaural and binaural acoustic features are used to generate missing data masks with a GMM-based binary classifier, and the behavior and classification importance of the defined features are analyzed. Both cluster-based (Raj et al., 2004) and sparse imputation methods using the generated masks are evaluated on the CHiME 2011 data set.

Table 4.2 compares the Harding et al. (2006) ILD/ITD masks also used in Publication IV with masks estimated with a binary GMM-based classifier and evaluated on the CHiME 2011 evaluation data set. Classifier-based masks are generated both with the ILD/ITD pairs as the only feature, as well as using the full 14-feature set. The classifier-based mask construction outperforms the simple histogram-based mask estimation even when limited to the same input features. Including the rest of the defined features results in even higher recognition accuracy. With the more sophisticated masks, missing data imputation improves performance over

the baseline at all tested SNR levels. In noisy scenarios (SNR of 6 dB or lower), performing model training on an imputed multi-condition training set and applying unsupervised speaker adaptation further improves the accuracy.

There are multiple explanations for the improved performance of the masks generated with the full feature set. One of the new features measures the interaural coherence, which allows for the classifier to de-emphasize the potentially misleading ILD and ITD features when the coherence is low. The ILD and ITD features are also easily degraded by reverberation, unlike, e.g., the modulation-filtered spectrogram feature included in the full 14-feature set. Finally, the ILD and ITD features cannot differentiate between the speaker and the noise if their positions overlap, unlike the monaural cues based on the spectral content of the signal.

Publication V also compares two missing data imputation methods: the NMF-based sparse imputation (SI), and the GMM cluster-based imputation (CI). While SI outperforms CI when oracle masks are used, for the estimated masks using the full feature set and the GMM-based classifier, the situation is reversed. Average keyword accuracy rates across all tested SNR levels for SI and CI are, respectively, 91.5% and 90.3% for oracle masks, and 63.6% and 72.7% for estimated masks. The CI and SI methods have been compared also by Remes et al. (2011), who found CI to outperform SI for impulsive noise, but vice versa for babble noise, in experiments using masks derived from noise estimates. SI was found to be more sensitive to features that are falsely labeled as reliable, as they easily lead to the selection of incorrect clean speech dictionary atoms for the sparse representation.

For the second (2013) CHiME challenge (Vincent et al., 2013), Keronen, Remes, Kallasjoki and Palomäki (2013) propose a missing data mask estimation scheme using binary classification with input features generated by Gaussian-Bernoulli restricted Boltzmann machines (GRBM). The proposed system is directly compared against the 14-feature classifier-based masks of Publication V by Keronen, Cho, Raiko, Ilin and Palomäki (2013).

In earlier studies (Gemmeke et al., 2011), the SVM-based masks used in Publication IV were found to perform well on the SPEECON corpus. The contrast to the results in this work may be explained by the effect of the noise type on the features used by the classifier. In the CHiME corpus, the noise is often the voice of an interfering talker. The classifier features

	Car			Public		
System	1	2	3	1	2	3
Baseline	**4.3**	29.5	69.0	**3.4**	23.5	39.5
SPLICE	**4.3**	**12.0**	36.0	**3.4**	14.1	23.7
NMF SS	5.7	13.3	**32.5**	5.1	**10.1**	**15.3**
Baseline, MC	4.2	6.7	18.9	**3.6**	6.5	12.1
NMF SS, MC	**4.1**	**5.7**	**12.8**	3.7	**5.4**	**8.4**

Table 4.3. Letter error rates (%) for the NMF-based source separation feature enhancement system on real noisy speech recorded in a moving car ("Car") or at a public place ("Public"). For acoustic models trained on clean speech, the evaluated system ("NMF SS") is compared against baseline methods of no feature processing ("Baseline") as well as the SPLICE feature enhancement ("SPLICE"). Performance is also evaluated for acoustic models trained on a multi-condition training set, with no processing ("Baseline, MC") or with both the training and test sets enhanced with the NMF-based source separation ("NMF SS, MC"). The best system is indicated individually for both sets of results.

used in these experiments are primarily based on harmonic decomposition and are not specifically designed for the case of interfering speech. In the experiments of Publication V, summarized in Table 4.2, classifier-based masks using the binaural ILD and ITD features outperformed the histogram-based ILD/ITD masks used in Publication IV.

Source Separation by Non-negative Matrix Factorization

The performance of the NMF-based source separation model for additive noise, described in Section 4.2.1, is evaluated for a realistic large vocabulary noisy speech recognition task in Publication III and VI. Table 4.3 summarizes the main results of Publication VI pertaining to the source separation. When clean speech acoustic models are used, the method generally outperforms a strong baseline system using the SPLICE (Deng et al., 2001) feature enhancement method, especially for the public place noise type. Best overall results are obtained using a combination of multi-condition training and NMF-based feature enhancement.

Source separation feature enhancement has a detrimental effect for the practically clean close-talk recordings of channel 1, when acoustic models trained on clean speech only are used. This is due to the match between the fixed clean speech dictionary and the observed signal, which is never fully perfect. As a result, small artefacts are generated by the feature enhancement. The difference is no longer significant when the enhanced multi-condition training set is used, as the acoustic model training in that

case includes also the artefacts introduced by the enhancement.

Publication VI extends the sparse source separation by incorporating a channel normalization scheme specific to noisy speech, originally proposed by Palomäki et al. (2004). The goal of the channel normalization is to reduce the mismatch between the dictionary atoms and the observed signal caused by different recording conditions. Applying the normalization has a clear beneficial effect on the speech recognition performance in noisy conditions, especially for the noisy car environment. The car recordings have a strong low-frequency noise component, and the microphones used for the different recording channels have very different frequency responses at the corresponding frequencies.

A key contribution and focus area of Publication III and VI is the use of heuristic uncertainty estimates in conjunction with NMF-based feature enhancement. Along with the corresponding work related to benefiting from uncertainty estimates in the missing data imputation context, these methods and results are discussed in Chapter 5. Publication VII extends the NMF-based model to handle speech distorted by reverberation. This use case is covered in Chapter 6.

5. Use of Uncertainty Information

Given real noisy speech signals, no existing feature enhancement method is capable of recovering the true clean speech perfectly under all conditions. In many cases, however, the magnitude of the error made by the feature enhancement process varies across time or spectral bands, and can be estimated. In addition, many feature enhancement methods are based on probabilistic models, and naturally yield a full posterior distribution as a result, the mean of which is then typically used as the enhanced features for an ASR system. Both ad hoc error estimates as well as clean speech estimates in the form of a posterior distribution contain valuable information about the *uncertainty* of a feature enhancement process, and this information can be used during the decoding process to focus more on the subset of input features likely to be correct, and to give a lower weight to less certain values. Such uncertainty information has been successfully used to augment various feature enhancement methods, including the SPLICE feature transformation (Droppo et al., 2002), independent component analysis (Kolossa et al., 2010), computational auditory scene analysis (Shao et al., 2010) and multichannel signal beamforming (Astudillo et al., 2013).

Conventionally, a front-end feature enhancement method produces, for each time frame t, a point estimate $\hat{\boldsymbol{x}}_t$ of the clean speech features:

$$\mathrm{FE}(\boldsymbol{y}_t, \boldsymbol{\theta}) = \hat{\boldsymbol{x}}_t, \tag{5.1}$$

where $\mathrm{FE}(\boldsymbol{y}_t, \boldsymbol{\theta})$ denotes the result of the feature enhancement methods on noisy observation $\boldsymbol{y}_t$, and $\boldsymbol{\theta}$ denotes the parameters for the feature enhancement. These estimates are then used in place of the original features by the rest of the ASR system. For an HMM-based ASR system, the decoding process builds on state likelihood values $L(q) = p(\hat{\boldsymbol{x}}_t \mid \mathcal{M}, q)$ defined by a state q of the acoustic model $\mathcal{M}$.

In the *observation uncertainty* framework, the underlying assumption

is that the result of the feature enhancement is a posterior distribution representing an estimate of the clean speech:

$$\mathrm{FE}(\boldsymbol{y}_t, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \boldsymbol{y}_t, \boldsymbol{\theta}). \tag{5.2}$$

In order to use the posterior distribution in the decoding stage, the state likelihood values $L(q)$ are calculated by integrating over all possible values of $\boldsymbol{x}$, yielding

$$L(q) = \int p(\boldsymbol{x} \mid \boldsymbol{y}_t, \boldsymbol{\theta}) p(\boldsymbol{x} \mid \mathcal{M}, q) \, \mathrm{d}\boldsymbol{x}. \tag{5.3}$$

An important simplification occurs when the acoustic model likelihood $p(\boldsymbol{x} \mid \mathcal{M}, q)$ is defined by a Gaussian mixture model, and the posterior distribution is restricted to be (or approximated as) a Gaussian distribution with an estimated mean $\hat{\boldsymbol{x}}_t$ and covariance $\boldsymbol{\Sigma}_t$, derived from the observation $\boldsymbol{y}_t$:

$$p(\boldsymbol{x} \mid \boldsymbol{y}_t, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x} \mid \hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_t). \tag{5.4}$$

In this case, the state likelihood $L(q)$ is a weighted sum of likelihoods $L(l)$ for each Gaussian component l, and the per-Gaussian likelihood computations reduce to

$$L(l) = \int \mathcal{N}(\boldsymbol{x} \mid \hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_t) \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}) \, \mathrm{d}\boldsymbol{x} \tag{5.5}$$

$$= \mathcal{N}(\hat{\boldsymbol{x}}_t \mid \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)} + \boldsymbol{\Sigma}_t), \tag{5.6}$$

where $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$ are the mean and covariance of mixture component l in the acoustic model. Decoding with observation uncertainties can therefore be done similarly to conventional feature enhancement, based on the mean $\hat{\boldsymbol{x}}_t$ of the posterior distribution, with the exception of an additional time-varying covariance compensation term $\boldsymbol{\Sigma}_t$ that is added to every Gaussian distribution of the acoustic model.

An alternative way of using uncertainty information in an ASR system is provided by the *uncertainty decoding* model (Droppo et al., 2002; Liao and Gales, 2005, 2008). The main assumption for uncertainty decoding is the existence of a probabilistic model for the formation of the corrupted noisy observation $\boldsymbol{y}_t$. This model defines the conditional distribution $p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta})$ in terms of the unseen clean speech $\boldsymbol{x}_t$ and the feature enhancement parameters $\boldsymbol{\theta}$. In decoding, the acoustic likelihood of an observed frame is obtained similarly to Equation (5.3), by marginalization over the clean speech $\boldsymbol{y}_t$, as

$$L(q) = \int p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}) p(\boldsymbol{x}_t \mid \mathcal{M}, q) \, \mathrm{d}\boldsymbol{x}_t. \tag{5.7}$$

In pure *front-end uncertainty decoding*, the distribution $p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta})$ is given a form where the included noise model is independent of the acoustic model state. Typical assumptions, made for the computational efficiency of the implementation, include representing the distribution with a Gaussian mixture model, and further selecting only the most likely component of the mixture model for each input frame, based on the noisy observation $\boldsymbol{y}_t$. In this case, the acoustic likelihood computation will have an exactly identical form to Equation (5.6), consisting of a single global variance offset per frame. The derivation of the parameters, however, has a fundamentally different theoretical framework.

5.1 Heuristic Uncertainty Estimates

The NMF-based source separation and missing data imputation systems discussed in, respectively, Section 4.2 and Section 4.3, do not inherently produce a probabilistic model of the variance of the estimated clean speech features. While generic methods such as the Wiener filter posterior distribution proposed by Astudillo (2010) may be applicable, it is also possible to define quantities specific to the NMF representation that are likely to correlate with the correctness of the resulting estimates. In general, such heuristic metrics will not directly measure the variance of the enhanced features, and therefore will need some form of transformation in order to be usable within the observation uncertainty framework.

In this thesis, the use of heuristic uncertainty metrics for observation uncertainty processing is investigated by Publication IV in the context of the NMF-based sparse imputation presented in Section 4.3. Correspondingly, Publication III and Publication VI apply the same principle to the NMF-based sparse source separation described in Section 4.2.

Some heuristic estimates are applicable to any front-end feature enhancement method. In particular, the difference between the noisy observation and the enhanced features can be used, under the assumption that in spectro-temporal regions where the feature enhancement processing has made large modifications to the input signal, the local SNR level is low, and the enhanced features are also more likely to differ from the true clean signal (Arrowood, 2003). This estimate is denoted as H1 in Publication III, and, with slight variations in the definition, as H1 through H3 in Publication VI.

With any feature enhancement method based on missing data process-

ing, the missing data mask can be used as a source of information. For methods based on a binary missing data mask, such as the imputation approaches evaluated in this work, a straightforward metric is the count of reliable components near the imputed area, as more reliable components are likely to improve the performance of the imputation. The M5 metric in Publication IV is based on this principle.

Correspondingly, for the NMF-based algorithms presented in this work, the sparse representation of the input observation encoded by the activation matrix is available. The degree of match between the input and the predefined dictionary matrix can be quantified by the number of dictionary atoms with significantly non-zero activation weights, since a representation based on multiple dictionary atoms is indicative of a poor match. As the quality of the reconstruction depends on the match between the observation and the dictionary, the count of active dictionary atoms in a frame can be assumed to negatively correlate with the quality of enhanced features for that frame. This heuristic is denoted by M4 in Publication IV, and by H4 in both Publication III and Publication VI.

Finally, in the case of the NMF-based source separation, the division of the dictionary matrix into distinct clean speech and noise atoms can be used for further heuristic uncertainty estimates. Instead of using the activation count for the entire dictionary, as above, the activation count heuristic can be restricted to the clean speech atoms only, which may be appropriate if some degree of mismatch between the noise dictionary and the observed noise is expected. This approach is denoted by H3 in Publication III and by H5 in Publication VI.

Alternatively, rather than counting activations, the ratio of the sums of the activation vector components corresponding to clean speech and noise atoms can be seen as a kind of a local SNR estimate for a single frame. For speech-dominated frames, some clean speech atoms will have large activation coefficients, while the noise atom activations will be close to zero, leading to a high ratio; and vice versa for noise-dominated frames. The H2 metric of Publication III and the corresponding H6 metric of Publication VI are based on this ratio.

5.2 Uncertainty Propagation

Feature enhancement methods commonly operate in the spectral domain, and therefore produce information about the variance of the enhanced

spectral feature components. On the other hand, spectral features are not optimal for the acoustic modeling used by conventional HMM-GMM-based ASR systems. For the ASR decoding in the observation uncertainty model, variance estimates in the feature domain used by the acoustic models are needed. In order to enable the use of features such as MFCCs in conjunction with uncertainty processing, the spectral variance information must be propagated through the feature transformation. In addition, the heuristic uncertainty metrics do not directly provide any variance estimates.

Proposed methods for uncertainty propagation can be split into two major classes: *data-driven* and *model-driven* (Kolossa et al., 2010). In the former category, a parallel data set containing computed variance estimates (or heuristic metrics) along with corresponding oracle variance information in the acoustic model domain is used by a machine learning system to learn the mapping between the domains. By contrast, the methods in the latter category are based directly on the mathematical model of the feature transformation when applied to a random variable. While this has the advantage of requiring no special training, it is only applicable when the source of information is already in the form of a variance estimate.

5.2.1 Learning Mappings from Data

In order to learn a mapping from an uncertainty estimate to a variance estimate in the feature domain of the acoustic model of an ASR system, a training data set with known feature variances is needed. Given a parallel data set containing corresponding clean and noisy speech signals, *oracle variances* can be computed for enhanced features as the squared difference between the true and estimated clean speech features, and used as training data for the mapping. Due to the need for known clean speech, the training data set is generally based on artificially corrupted noisy speech. In practice, the mapping has been observed to be relatively robust to the choice of training data (Publication VI). Many classes of transformations have been proposed, including simple linear mappings (Gemmeke, Remes and Palomäki, 2010), multilayer perceptron neural networks (Srinivasan and Wang, 2006) and regression trees (Srinivasan and Wang, 2007).

In the publications of this thesis, a GMM-based mapping is used to transform the heuristic metrics to variance estimates. The mapping is similar in principle to the cluster-based imputation algorithm described

in Section 4.1.2, although the lack of the upper bound constraints simplifies the implementation. Formally, the mapping is constructed as follows.

Denoting a heuristic uncertainty estimate by $\boldsymbol{h}$, and the corresponding oracle variances in the acoustic model domain by $\boldsymbol{\sigma}$, a GMM is trained to model the concatenated vectors $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{h} & \log \boldsymbol{\sigma} \end{bmatrix}^{\mathsf{T}}$. The logarithmic compression of the oracle variances $\boldsymbol{\sigma}$ is intended to make the distribution easier to model by Gaussian components. This transformation was motivated by observing histograms of computed oracle variance values in small-scale experiments for Publication III, which introduced the use of a Gaussian mixture model for mapping uncertainty values to the acoustic model domain. The distribution of $\boldsymbol{z}$ given by the GMM has the form

$$p(\boldsymbol{z}) = \sum_l p(l) \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}), \tag{5.8}$$

where l is a mixture component index, $p(l)$ the mixture component weight, and $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$ are the per-component mean and covariance, respectively.

For unseen data, given the heuristic estimate $\boldsymbol{h}$ and the GMM parameters $\boldsymbol{\Lambda}$, the minimum mean square error (MMSE) estimate for the corresponding (log-compressed) acoustic model domain variance $E\{\log \boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{\Lambda}\}$ has the solution

$$E\{\log \boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{\Lambda}\} = \sum_l p(l|\boldsymbol{h}, \boldsymbol{\Lambda}) E\{\log \boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{\Lambda}, l\}, \tag{5.9}$$

$$E\{\log \boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{\Lambda}, l\} = \boldsymbol{\mu}_\sigma^{(l)} + \boldsymbol{\Sigma}_{h\sigma}^{(l)} \left(\boldsymbol{\Sigma}_{hh}^{(l)}\right)^{-1} \left(\boldsymbol{h} - \boldsymbol{\mu}_h^{(l)}\right). \tag{5.10}$$

The cluster membership term $p(l|\boldsymbol{h}, \boldsymbol{\Lambda})$ in Equation (5.9) can be computed from the component weight $p(k)$ and the likelihood $p(\boldsymbol{h} \mid l, \boldsymbol{\Lambda})$. In Equation (5.10), the terms $\boldsymbol{\mu}_h^{(l)}$ and $\boldsymbol{\mu}_\sigma^{(l)}$ denote the components of the mean vector $\boldsymbol{\mu}^{(l)}$ corresponding to the heuristic estimate and the oracle variances, respectively, while $\boldsymbol{\Sigma}_{hh}^{(l)}$ is the covariance of the heuristic estimates, and $\boldsymbol{\Sigma}_{h\sigma}^{(l)}$ the cross-covariance.

5.2.2 Propagation Through Feature Transformations

When variance information is available for, e.g., spectral features, the propagation of this information through many feature transformations can be handled mathematically, by treating the original features as random variables drawn from particular distributions. Due to the nonlinear transformations and operations on multiple feature components involved in extracting features such as MFCCs, finding a closed form solution for the distribution of the transformed features may be infeasible.

While propagating only the first and second moments of the distribution is sufficient for taking advantage of uncertainty information in an ASR system, various approximations for the specific feature transformation steps may still be required (Astudillo, 2010; Astudillo et al., 2010). Uncertainty propagation has been successfully applied to several well-known speech recognition features, such as MFCCs and RASTA-PLP features (Kolossa et al., 2010), as well as the ETSI advanced front-end feature extraction (Astudillo et al., 2010).

This model-driven approach to uncertainty propagation has not been used for the heuristic uncertainty estimates proposed by the publications of this thesis, since a data-driven transformation step would still be required to convert the defined heuristics into variance estimates in any domain. However, early unpublished small-scale experiments have yielded promising results for a limited form of uncertainty propagation in conjunction with the system described in Publication IV. In these experiments, the GMM-based mapping was still used to generate variance estimates for the static MFCC coefficients, which were then propagated through the simple operations of generating the first and second-order differential ("delta" and "delta-delta") MFCC terms, as well as the linear MLLT transformation used by the ASR system.

5.3 Experimental Results and Discussion

This section contains key experimental results of the three publications in this thesis involving feature enhancement extended by uncertainty processing (Publication III, IV, VI). For legibility, Publication IV, concerning a missing data imputation system, is denoted here by IMP, while publications III and VI, based on a sparse source separation feature enhancement method, are denoted by SEP1 and SEP2, respectively.

Common to all three publications, the change in speech recognition accuracy caused by the inclusion of uncertainties is small compared to the effect of the feature enhancement. The effect, however, is in general uniformly positive in all environments and at all noise levels, with the exception of multi-condition training on enhanced features. The results of using heuristic uncertainty metrics also fall far short of the upper bound established by experiments involving oracle variances. The oracle results, if achievable, would represent a significant performance improvement over the use of feature enhancement without uncertainty processing.

	IMP	SEP1	SEP2			
			Artificial		Real	
			Car	Pub.	Car	Pub.
Baseline	44.0	45.9	62.2	41.2	34.3	22.1
FE only	49.2	18.3	26.8	19.3	17.2	10.2
Diff	—	**16.2**	26.7	19.0	15.3	**9.2**
#Act	—	16.7	25.7	18.8	**13.9**	**9.2**
#CleanAct	**53.4**	16.5	**25.4**	**18.7**	**13.9**	**9.2**
ActRatio	—	16.7	25.7	19.2	15.0	9.7
Mask	52.8	—	—	—	—	—
Oracle	74.2	13.1	18.0	14.6	—	—
Oracle, spec.	—	15.2	—	—	—	—

Table 5.1. Recognition performance achieved by the different uncertainty heuristics. Results for using oracle variances are also included when available, i.e., for experiments involving known clean speech artificially corrupted by noise. Note that the different publications use entirely different feature enhancement systems, data sets and evaluation metric, so the results are not directly comparable. For all publications, the results in this table refer to acoustic models trained on clean speech and tested on noisy speech. Tested systems include the baseline system ("Baseline"); noisy speech feature enhancement (IMP: NMF-based sparse imputation; SEP1, SEP2: NMF-based source separation) with no uncertainty information ("FE only"); the same feature enhancement with various heuristic uncertainty estimates; and finally with oracle variances derived from known clean speech, either in the acoustic model feature domain ("Oracle") or in the mel-spectral domain ("Oracle, spec."). The uncertainty heuristics investigated here are based on the difference between noisy and enhanced features ("Diff"), the number of clean and noisy dictionary atom activations ("#Act"), the number of clean dictionary atom activations only ("#CleanAct"), the ratio between clean and noisy atom activations ("ActRatio") or the number of reliable components in the missing data mask ("Mask"). For IMP, the presented values are the keyword accuracy (%) averages of the CHiME 1 development set for the noisier conditions (SNRs of 0 dB, -3 dB, -6 dB) where the missing data imputation was beneficial. For SEP1, shown is the letter error rate (%) averaged over all tested data sets. For SEP2, shown are again letter error rate (%) averages over all SNR levels or recording channels, for both the artificial and real noisy data sets in the car ("Car") and public place ("Pub.") environments.

A summary of results for different heuristics, across all publications, is provided in Table 5.1. While the differences between individual heuristics are small, the choice of optimal heuristic appears to depend, not only on the feature enhancement system, but on the test data. The heuristic based on the difference of observed and enhanced features ("Diff") performs best in SEP1, but the overall best heuristic for the experiments reported in SEP2 is based on the clean speech dictionary atom activation counts. Although the feature enhancement and ASR systems in SEP1 and SEP2 are highly alike, the noisy speech in SEP1 is clean speech artificially corrupted with different types of factory noise, while in SEP2 both artificial and real noisy data from public place and car environments are used. Notably, too, the highest performance increase achieved by the use of heuristic uncertainties in SEP2 is seen for the real noisy recordings in the car environment; a result that is not observed in the corresponding artificial car noise data set.

Both SEP1 and SEP2 include experiments that combine multiple heuristic estimates. In SEP1, recognition performance using the concatenation of all tested uncertainty heuristics is found to be no better than the best-performing heuristic alone. This result is confirmed also in SEP2. In addition, a similar experiment using a feedforward MLP neural network to map the concatenated heuristic directly to an acoustic model domain variance estimate achieves performance closely matching the combination of the single best heuristic and the GMM-based mapping. SEP2 includes an analysis of the correlation between errors in the final feature variance estimates, suggesting that all the proposed heuristics make very similar errors.

Oracle experiments both with and without the GMM-based mapping between domains were compared in SEP1. Oracle variances estimated directly in the acoustic model feature domain were used for the case with no mapping ("Oracle"). Results with the GMM-based mapping were obtained by computing the variance estimates in the mel-spectral domain used for the feature enhancement processing and transforming them similarly to the heuristic estimates ("Oracle, spec."). The considerably lower performance in the latter experiment suggests that imperfections in the data-driven uncertainty propagation are one of the causes why the gains provided by the heuristic uncertainty estimates remain much lower than the acoustic model domain oracle variances.

The feature enhancement system submitted to the first CHiME chal-

	SNR (dB)					
System	9	6	3	0	−3	−6
Baseline	**85.6**	**77.9**	66.3	51.2	40.0	38.7
SI only	74.3	70.1	64.3	54.7	45.3	42.8
SI and uct.	77.3	73.5	**67.3**	**58.5**	**47.8**	**46.8**

Table 5.2. CHiME challenge evaluation set results for the missing data method proposed in IMP. The shown values indicate the CHiME keyword accuracy (%) at all tested SNR levels. The baseline system with no feature enhancement ("Baseline") is compared to the system performing only NMF-based sparse missing data imputation ("SI only") and the system applying sparse imputation and heuristic uncertainty information ("SI and uct.").

lenge in IMP, using NMF-based sparse imputation with binaural masks, is beneficial only in the noisy evaluation sets, with an SNR level of 0 dB or lower, as indicated by the final CHiME evaluation results shown in Table 5.2. However, the addition of observation uncertainty processing using a heuristic estimate improves results at all SNR levels, and the improved system is sufficient to outperform the baseline also at the 3 dB SNR level.

The major results of SEP2, on real noisy speech recordings, are represented in Table 5.3. As established in Section 4.4, applying the source separation with acoustic models trained on unprocessed data degrades the performance for the close-talk speech (practically clean condition) of channel 1. However, the use of the uncertainty information is able to partially compensate for the degradation.

For multi-condition training with feature-enhanced data, the system using uncertainty information ("FE-UC") performs more poorly than the feature enhancement alone ("FE-SS"). This result is expected, as the use case is contrary to the observation uncertainty framework. The feature variance term $\boldsymbol{\Sigma}_t$ in Equation (5.6) is added to the model variances under the assumption that the acoustic models reflect the underlying clean speech. With multi-condition training, the effect of the feature enhancement on noisy data is already accounted for by the acoustic model parameters, leading to an overestimation of the variance. A more principled *uncertainty training* method that accounts for feature enhancement uncertainty during the acoustic model training is presented in Ozerov et al. (2013).

The performance of the sparse source separation on clean speech acoustic models, denoted by the abbreviations of the form "CL-x" in Table 5.3, is also compared against the SPLICE feature enhancement method in

	Car			Public			
System	1	2	3	1	2	3	Avg.
CL-NO	4.3	29.5	69.0	**3.4**	23.5	39.5	28.2
CL-SS	5.7	13.3	32.5	5.1	10.1	15.3	13.7
CL-UC	4.8	**10.4**	**26.5**	4.1	**9.4**	**14.0**	**11.5**
CL-SP	4.3	12.0	36.0	**3.4**	14.1	23.7	15.6
CL-SPU	**4.0**	11.2	35.5	**3.4**	13.1	21.3	14.8
MC-NO	**4.2**	**6.7**	18.9	**3.6**	**6.5**	12.1	**8.7**
MC-SS	7.0	8.8	16.4	7.3	10.0	11.7	10.2
MC-UC	5.9	7.6	**12.9**	5.9	9.1	**10.5**	**8.7**
FE-SS	**4.1**	**5.7**	12.8	**3.7**	**5.4**	**8.4**	**6.7**
FE-UC	4.2	5.9	**12.5**	**3.7**	6.0	8.8	6.9

Table 5.3. Letter error rate (%) for the systems evaluated in SEP2. This table shows the results for the real noisy recordings in the car ("Car") and public place ("Public") environments, with the numbers (1–3) denoting the recording channel. The first component of the system name denotes the data set used for acoustic model training: clean speech ("CL"), multi-condition clean and noisy speech ("MC"), or the multi-condition data processed by the sparse source separation feature enhancement method ("FE"). The second component denotes the feature enhancement processing performed during evaluation: no processing ("NO"), source separation ("SS"), source separation with heuristic uncertainty information ("UC"), SPLICE feature enhancement ("SP") or SPLICE with uncertainty ("SPU").

SEP2. As the SPLICE method has a natural extension for estimating the variance of the enhanced features (Droppo et al., 2002), it is evaluated both as-is ("CL-SP") and by including the feature variance estimates as observation uncertainties, mapped with the same GMM-based system ("CL-SPU"). Due to the use of an environment selection scheme adapted from that proposed by Droppo et al. (2001), the SPLICE method does no processing for the almost clean recording channel 1, and therefore outperforms the NMF-based feature enhancement in that case. For the noisy recording channels 2 and 3, however, source separation with observation uncertainty processing performs better than the SPLICE baseline both in the car and public place environments.

6. Handling Reverberant Speech

In this thesis, the case of reverberant speech is considered in Publication VII and Publication VIII, both of which were submitted to the REVERB challenge (Kinoshita et al., 2013), and in Publication IX, which describes the combination of the above methods.

Missing data and non-negative matrix factorization methods, introduced in the context of additive noise in Section 4.1 and Section 4.2, respectively, are compared in Publication VII. Publication VIII proposes a novel *distribution matching* method, based on the concept of histogram equalization, incorporating a long time context via the use of stacked input vectors consisting of windows of several frames. In Publication IX, a thorough evaluation of the NMF-based dereverberation algorithm is carried out. In addition, the distribution matching method is investigated as an alternative way to produce an initial estimate for the NMF approach.

6.1 Missing Data Dereverberation

Both the imputation and marginalization approaches to missing data speech recognition are suitable for handling input corrupted by reverberation, if masks that label each time-frequency component as dominated by direct or indirect sound can be generated. However, accuracy of the input masks is a critical factor for the performance of the resulting system. Mask estimation is therefore the major challenge in designing a missing data system to handle reverberant speech.

While missing data processing has been extensively investigated in the context of additive noise, the problem of estimating masks suitable for convolutional distortion has received relatively little attention. Palomäki et al. (2004, 2006) propose using reverberation-specific missing data masks based on *modulation filtering*, in which time domain filters are applied to

the samples of a single frequency band in a spectrogram representation. The proposed method uses a combination of a low-pass filter, to detect modulation relevant to speech, and a differentiator, to emphasize the fast onsets that are likely associated with the direct sound and early reflections and to de-emphasize the slowly changing areas dominated by late reverberation. A binary missing data mask can be produced by thresholding the output of the modulation filter.

The data-driven approach of producing missing data masks by using a classifier trained on oracle data can be applied in any context. However, the optimal features for differentiating between direct sound and late reverberation are likely to differ from the ones that have been found the most useful for differentiating between speech and non-speech noise sources.

6.1.1 Mask Estimation for Reverberant Speech

Publication VII investigates the performance of four missing data mask estimation methods in the REVERB challenge reverberant speech recognition task (Kinoshita et al., 2013). The m_{R} mask is derived from the modulation filter mask proposed in Palomäki et al. (2004, 2006). The m_{LP} mask is adapted from a method for estimating the overall level of reverberation, originally presented by Beeston and Brown (2013), and is also based on the modulation filtering concept. Finally, the m_{GMM} and m_{SVM} masks use the classifier-based mask estimation process, applying GMM and SVM classifiers, respectively.

In its original formulation, the Palomäki et al. (2004, 2006) mask was based on a gammatone filterbank as the source for the modulation-filtered signals. In its place, Publication VII uses the mel-scale filterbank from the feature extraction process of the REVERB challenge baseline speech recognition system, followed by a compression step of raising the output to the power 0.3. The resulting values for time frame t and frequency band b are denoted as $y(t, b)$.

The modulation filter for the m_{R} mask is a band-pass filter with 3 dB cutoff frequencies of 1.5 Hz and 8.2 Hz. Its output is processed by an automatic gain control block, and normalized by subtracting a per-utterance, per-frequency-band value selected so that the minimum value of each frequency band over an utterance is zero. Denoting this processed signal as

$y_{\text{bp}}^{\text{agc}}(t,b)$, the final mask is defined as

$$m_{\text{R}}(t,b) = \begin{cases} 1 & \text{if } y_{\text{bp}}^{\text{agc}}(t,b) > \theta(b), \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Following Palomäki et al. (2004), the frequency band threshold $\theta(b)$ is based on the 'blurredness' metric B, given by

$$B = \sum_b \frac{\frac{1}{T}\sum_t y(t,b)}{\max_t y(t,b)}, \tag{6.2}$$

where T is the total number of time frames. The threshold value for each channel b is set to a fraction of the channel's average energy in the modulation-filtered spectrogram, given by a sigmoid function

$$\theta(b) = \gamma \frac{\frac{1}{T}\sum_t y_{\text{bp}}^{\text{agc}}(t,b)}{1 + \exp\left(-\alpha\left(B - \beta\right)\right)} \tag{6.3}$$

where α, β and γ are tunable hyperparameters that can be set, e.g., by small-scale experiments.

In Beeston and Brown (2013), reverberation tails are located by using a 10 Hz low-pass modulation filter to estimate the smoothed temporal envelope of each frequency band. The regions where y'_{lp}, the difference between two consecutive filter outputs in time, is negative are identified as potentially reverberant. The level of reverberation in the input can then be estimated based on the average spectral energy in these regions. The principle behind this estimate is that, as reverberation increases, the reverberant tails following spectral peaks decay more slowly, causing a corresponding increase in the average energy.

For mask estimation, a reverberation estimate for each time-frequency component can be calculated as

$$L(t,b) = \begin{cases} \frac{1}{|n(t,b)|}\sum_{k \in n(t,b)} y(t,b) & \text{if } y'_{\text{lp}}(t - \tau_d, b) < 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6.4}$$

where τ_d corrects for the filter delay, and $n(t,b)$ is the set of contiguous time indices of frequency band b around t where y'_{lp} remains negative. The final mask is then obtained by thresholding with a fixed value θ_{LP}, as

$$m_{\text{R}}(t,b) = \begin{cases} 1 & \text{if } L(t,b) < \theta_{\text{LP}}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

For the classifier-based masks m_{GMM} and m_{SVM}, Publication VII uses a set of six features as the classifier input. The features include the compressed mel-spectral samples $y(t,b)$ and the GRAD feature, an estimate of

System	Simulated	Real
Baseline	51.81	89.04
Mask m_{R}	40.07	**67.88**
Mask m_{LP}	48.01	73.06
Mask m_{GMM}	**39.93**	70.87
Mask m_{SVM}	40.78	74.14

Table 6.1. Word error rates for the simulated and real data development sets of the REVERB challenge. Shown are the baseline recognizer results, using original data ("Baseline"), or with missing data imputation using one of the defined mask estimation methods.

the local slope of $y(t,b)$, used for noisy data imputation in Publication V. In addition, four features derived from the mask estimation methods m_{R} and m_{LP} are used: the modulation-filtered signal $y_{\mathrm{bp}}^{\mathrm{agc}}(t,b)$, the ratio $\frac{y_{\mathrm{bp}}^{\mathrm{agc}}(t,b)}{\theta(b)}$ and the 'blurredness' metric B defined for mask m_{R}, and the reverberant energy estimate $L(t,b)$ defined for mask m_{LP}.

6.1.2 Experimental Results and Discussion

For experimental evaluation, Publication VII uses the REVERB challenge baseline recognition system (Kinoshita et al., 2013). Missing data feature enhancement is performed by BCMI imputation, described in Section 4.1.2 in the context of additive noise.

A summary of the resulting letter error rates for the REVERB challenge development set using all four mask estimation methods are presented in Table 6.1. These results reflect performance with acoustic models trained on clean speech. Overall, the best performing mask estimation method is the m_{R} mask, achieving similar performance as the classifier-based m_{GMM} and m_{SVM} masks on simulated data, yet generalizing better to the real reverberant data set. A likely cause is the training of the mask classifiers on simulated data, as it requires the availability of oracle masks. The m_{LP} mask is efficient only for highly reverberant conditions, as can be seen in the more detailed results included in Table 1 of Publication VII.

Further results of missing data dereverberation using the m_{R} mask can be found in Table 6.2 in Section 6.2, where the missing data front-end is denoted by "MD". The dereverberation is seen to be beneficial for both simulated and real data sets, even in conjunction with multi-condition training and preprocessing 8-channel microphone array signals using delay-sum beamforming.

In a small-scale experiment on the simulated data set used for training the mask classifiers, applying missing data dereverberation with the oracle masks outperformed the generally better NMF-based dereverberation method described in Section 6.2. However, realistic mask estimation methods typically do not achieve results close to the upper bounds established by oracle experiments.

The classifier-based masks m_{GMM} and m_{SVM} presented in Publication VII rely on simple single-channel spectral features. As shown in Section 4.4 and Publication V, in the context of missing data feature enhancement for noisy speech, improved performance with classifier-based masks can be obtained with an extended set of monaural and binaural features. The use of multichannel acoustic features for generating missing data masks for reverberant speech is a potential topic for future research.

6.2 Extending NMF Feature Enhancement

The use of non-negative matrix factorization (NMF) algorithms in the context of speech degraded by additive noise sources is covered in Section 4.2. Following the terminology in Section 2.4, reverberation can be considered to consist of the direct sound, propagated directly from the sound source to the recording device, and the reflected components, consisting of the early reflections and late reverberation. While the reflected components of reverberant speech can be seen as additive to the direct sound, modeling them in the conventional NMF source separation model is impractical, as they can be spectrally similar to the dry speech and vary based on the environment and relative position of the speaker. In addition, treating the reflected components as a separate sound source would lose valuable information by ignoring the known dependency between the direct sound and its reflected counterparts.

Publication VII extends the NMF-based model of Section 4.2 to allow for the estimation of a separate per-utterance convolutional filter for each of the frequency bands in the mel-spectrogram domain. The proposed method takes advantage of the stacked vector representation of the fixed-length spectrograms, which permits the use of a matrix multiplication with a specifically designed matrix to perform the convolutive filtering of the spectrogram channels. The stacked vector representation also allows for a computationally efficient implementation of the non-negative matrix factor deconvolution (NMFD) model (Smaragdis, 2000), which is

important for avoiding reverberation-specific problems of the simple sliding window approach described earlier.

NMFD has been used as a blind deconvolution method to factorize reverberant speech into the convolution of a clean speech and room reverberation estimate, using the sparsity of speech spectra as a constraint to regularize the factorization (Kameoka et al., 2009; Kumar et al., 2011). In this model, the resulting factorization consists of two matrices, representing the filter and the clean speech signal. By contrast, the method proposed in Publication VII retains the model of factorizing speech into a sparse linear combination of predefined clean speech dictionary atoms, represented as a multiplication of a dictionary and an activation matrix. The convolutive filter is estimated in addition to this factorization.

6.2.1 NMF Model for Reverberant Speech

As the REVERB challenge corpus (Kinoshita et al., 2013) used in Publication VII contains very limited levels of additive noise, a simplified version of the compositional speech model of Section 4.2, with a dictionary matrix consisting of clean speech only, is used as the basis. In it, the observed stacked spectrogram $\boldsymbol{Y}$, a $TC \times N$ matrix of N overlapping windows of T-frame, C-band spectrograms, is approximated as

$$\boldsymbol{Y} \approx \boldsymbol{S}\boldsymbol{A}, \tag{6.6}$$

where $\boldsymbol{S}$ and $\boldsymbol{A}$ are again the $TC \times K$ dictionary and $K \times N$ activation matrices, respectively, with K individual dictionary atoms. The dictionary matrix $\boldsymbol{S}$ contains samples of clean, dry speech.

To account for the convolutional distortion caused by reverberation, the model of Equation (6.6) is modified to have the form

$$\boldsymbol{Y} \approx \boldsymbol{R}\boldsymbol{S}\boldsymbol{A}, \tag{6.7}$$

where the left multiplication by $\boldsymbol{R}$ performs the convolution of the channels of each stacked vector spectrogram with corresponding filters of length T_f. The result of the multiplication is the stacked representation of a spectrogram of length $T_r = T + T_f - 1$. Conceptually, the approximation $\boldsymbol{Y} \approx \boldsymbol{R}\boldsymbol{S}\boldsymbol{A}$ can be equivalently interpreted either as modeling the observation using an artificially reverberated dictionary, $(\boldsymbol{R}\boldsymbol{S})\,\boldsymbol{A}$, or by convolving the conventional NMF approximation, $\boldsymbol{R}\,(\boldsymbol{S}\boldsymbol{A})$.

Denoting the filter coefficients for mel band c as $\begin{bmatrix} r_{c,1} & r_{c,2} & \cdots & r_{c,T_f} \end{bmatrix}$, the convolution is performed by defining $\boldsymbol{R}$ as the $T_rC \times TC$ block-structured

matrix

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & 0 & \cdots & 0 \\ \boldsymbol{R}_2 & \boldsymbol{R}_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{R}_{T_f} & \cdots & \boldsymbol{R}_2 & \boldsymbol{R}_1 \\ 0 & \boldsymbol{R}_{T_f} & \cdots & \boldsymbol{R}_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{R}_{T_f} \end{bmatrix}, \tag{6.8}$$

where the $C \times C$ blocks $\boldsymbol{R}_t$ are diagonal matrices containing the t'th filter coefficient for each mel band,

$$\boldsymbol{R}_t = \operatorname{diag}\begin{bmatrix} r_{1,t} & r_{2,t} & \cdots & r_{C,t} \end{bmatrix}. \tag{6.9}$$

This construction is analogous to performing a time-domain convolution by multiplication with a Toeplitz matrix representation of the filter.

Similarly to the NMF-based source separation method for additive noise presented in Section 4.2, feature enhancement can be performed based on the dry and reverberant speech reconstructions

$$\hat{\boldsymbol{X}}_{\text{NMF}} = \boldsymbol{S}\boldsymbol{A}, \tag{6.10}$$

$$\hat{\boldsymbol{Y}}_{\text{NMF}} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{A}. \tag{6.11}$$

Denoting by $\bar{\boldsymbol{X}}_{\text{NMF}}$ and $\bar{\boldsymbol{Y}}_{\text{NMF}}$ the corresponding spectrograms after deconstructing the stacked vector representations, while the reconstruction $\bar{\boldsymbol{X}}_{\text{NMF}}$ could be directly used as a dry speech estimate, better results are again obtained by the filtered estimate $\tilde{\boldsymbol{X}}$, defined as

$$\left[\tilde{\boldsymbol{X}}\right]_{c,n} = \frac{\left[\bar{\boldsymbol{X}}_{\text{NMF}}\right]_{c,n}}{\left[\bar{\boldsymbol{Y}}_{\text{NMF}}\right]_{c,n}} \left[\bar{\boldsymbol{Y}}\right]_{c,n}, \tag{6.12}$$

where $\bar{\boldsymbol{Y}}$ is the original reverberant spectrogram. This Wiener-like filtering is directly analogous to that of Equations (4.11) to (4.13) in Section 4.2.1.

6.2.2 Optimizing the Factorization

As the dictionary matrix $\boldsymbol{S}$ is fixed, the multiplicative update rules of conventional iterative NMF algorithms can be adapted to optimize the $\boldsymbol{R}$ and $\boldsymbol{A}$ matrices. The corresponding cost function is

$$\sum_n \left(d([\boldsymbol{Y}]_n, [\boldsymbol{R}\boldsymbol{S}\boldsymbol{A}]_n) + \|\lambda [\boldsymbol{A}]_n\|_1 \right), \tag{6.13}$$

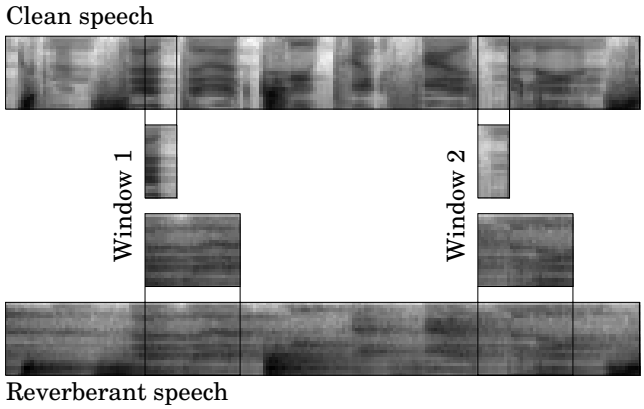


Figure 6.1. Logarithmic mel-spectrogram images of corresponding clean and artificially reverberated utterances, and two corresponding window pairs with lengths 10 and 30 for the clean and reverberated speech, respectively. Dark tones indicate regions with higher signal energy. Window 2 illustrates a segment that is problematical for the sliding window model: the true clean speech contains a period of silence, but in the reverberant observation, reverberation from preceding frames is present at the start of the window.

where d is again the Kullback-Leibler divergence, λ the sparsity cost coefficient, and $[\cdot]_n$ denotes the n'th column vector of a matrix. The resulting update rules, following the derivation in Gemmeke et al. (2011), are

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{(\boldsymbol{RS})^\mathsf{T} \frac{\boldsymbol{Y}}{\boldsymbol{RSA}}}{(\boldsymbol{RS})^\mathsf{T}\mathbf{1} + \lambda}, \tag{6.14}$$

$$\boldsymbol{R} \leftarrow \boldsymbol{R} \odot \frac{\frac{\boldsymbol{Y}}{\boldsymbol{RSA}}(\boldsymbol{SA})^\mathsf{T}}{\mathbf{1}(\boldsymbol{SA})^\mathsf{T}}. \tag{6.15}$$

Here $\odot$ denotes elementwise multiplication, division of matrices is likewise performed elementwise, and $\mathbf{1}$ is a $T_r C \times N$ all-one matrix.

These update rules implement the sliding window approach, where the activations for each window are optimized in isolation. For the presented model, this poses a problem under reverberant conditions, when the window is aligned so that it begins in a region that has a weak direct sound component but a strong reflected sound component in the reverberated spectrogram. As the clean speech dictionary atoms have been individually transformed with matrix $\boldsymbol{R}$, sound energy at the beginning of a window will be considered as direct sound, and will not be attenuated when the estimates for each overlapping window are averaged. This situation is illustrated in Figure 6.1.

The NMFD model is more robust against this issue, as a well-matching earlier activation can "explain away" the energy of the reverberant tail, as long as the reverberated dictionary atoms are sufficiently long. The

stacked vector representation of dictionary atoms makes it possible to implement the NMFD model by a slight adjustment of the multiplicative update rules. Denoting by o the operation of converting a stacked spectrogram into a regular spectrogram by adding together the overlapping regions of the windows and by s the operation of converting a spectrogram to the stacked form, the modified update rules are (Publication IX)

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{(\boldsymbol{RS})^{\mathsf{T}} s(\frac{\boldsymbol{Y}_0}{o(\boldsymbol{RSA})})}{(\boldsymbol{RS})^{\mathsf{T}}\boldsymbol{1} + \lambda}, \tag{6.16}$$

$$\boldsymbol{R} \leftarrow \boldsymbol{R} \odot \frac{s(\frac{\boldsymbol{Y}_0}{o(\boldsymbol{RSA})})(\boldsymbol{SA})^{\mathsf{T}}}{\boldsymbol{1}(\boldsymbol{SA})^{\mathsf{T}}}, \tag{6.17}$$

where $\boldsymbol{Y}_0$ is the original unstacked spectrogram form of $\boldsymbol{Y}$.

While the update rule of Equation (6.17) preserves some of the structure present in the $\boldsymbol{R}$ matrix in that only initially nonzero elements can ever be nonzero, it does not maintain the repetition of the filter coefficients in $C \times C$ blocks, nor does it necessarily result in a physically plausible filter. After every update, $\boldsymbol{R}$ is therefore reconstructed to have the form specified in Equation (6.8). The updated filter coefficients $r_{c,t}$ are obtained by averaging across their occurrences in the $\boldsymbol{R}$ matrix. Based on assumptions about the decay of the reverberation, they are further clamped to satisfy $r_{c,n+1} \leq r_{c,n}$ and normalized by uniform scaling to $\sum_{c,n} r_{c,n} = C$.

In conventional NMF optimization, the two matrices, here $\boldsymbol{A}$ and $\boldsymbol{R}$, are updated in alternating iterations. Based on early experiments, the conventional update scheme does not lead to acceptable results in this context: while it minimizes the cost function, the resulting factorization is not suitable for dereverberation. Publication VII defines three further refinements to the factorization process in order to mitigate this issue: initialization based on an alternative dereverberation method, a fixed iteration scheme, and filtering of the activation matrix. The resulting factorization algorithm is as follows:

1. Using an initial dry speech estimate $\hat{\boldsymbol{X}}_I$, the activation matrix $\boldsymbol{A}$ is updated for I_1 iterations to optimize $\hat{\boldsymbol{X}}_I \approx \boldsymbol{SA}$.

2. The activation sequences for each dictionary atom (rows of $\boldsymbol{A}$) are filtered with the filter H_A, with the output clamped to be non-negative.

3. The $\boldsymbol{R}$ matrix is initialized to contain the T_f-length filter $\frac{1}{T_f}\begin{bmatrix}1 & \cdots & 1\end{bmatrix}$ on each frequency band.

4. Keeping $\boldsymbol{A}$ fixed, $\boldsymbol{R}$ is updated for I_2 iterations to optimize $\boldsymbol{Y} \approx$

$\boldsymbol{RSA}$, performing the reconstruction of matrix $\boldsymbol{R}$ described above after each update.

5. Keeping $\boldsymbol{R}$ fixed, $\boldsymbol{A}$ is updated for I_3 iterations to optimize $\boldsymbol{Y} \approx \boldsymbol{RSA}$.

Based on numerous small-scale tests, Publication VII uses an activation matrix filter of $H_A(z) = 1 - 0.9z^{-1} - 0.8z^{-2} - 0.7z^{-3}$. The goal of the activation filtering in step 2 is to remove traces of reverberation present in the activation matrix $\boldsymbol{A}$, in order to bias the filter matrix $\boldsymbol{R}$ optimization (step 4) to account for the effects of reverberation using the convolutional filter. This is crucial for the efficiency of the feature enhancement, as any remaining effect of reverberation in matrix $\boldsymbol{A}$ will be present in both reconstructions $\boldsymbol{SA}$ and $\boldsymbol{RSA}$ used by Equation (6.12), and therefore will not be attenuated in the output. The chosen filter reduces the activations of a certain dictionary atom that occur immediately after an earlier activation of the same atom, as these are typical of reverberant speech. The filtering step is similar to dereverberation methods based on modulation filtering (Palomäki et al., 2004, 2006), with the exception of taking place in the domain of the sparse representation $\boldsymbol{A}$ instead of the conventional spectrogram domain.

The initial estimate $\hat{\boldsymbol{X}}_I$ in Publication VII is generated by the missing data system described in Section 6.1. In Publication IX, the missing data initialization is compared to using the distribution matching dereverberation method described in Section 6.3 to generate the initial estimate. In addition, the use of the method with no initial estimate, using the reverberant observation as the source of matrix $\hat{\boldsymbol{X}}_I$, is investigated. The selected iteration counts, based on small-scale experiments, are $I_1 = 50$, $I_2 = 50$ and $I_3 = 100$.

6.2.3 Experimental Results and Discussion

As the experiments are based on the same evaluation data, a combined summary of the key results of Publication VII and Publication IX for the REVERB challenge evaluation set is presented in Table 6.2.

Three systems described in Publication VII are shown: models based on both clean speech and multi-condition training sets, in addition to the overall best results obtained by including a delay-sum beamformer processing 8-channel microphone array signals, and a per-speaker CMLLR adaptation step. The NMF processing significantly improves the recog-

System	FE	Simulated	Real
VII, clean	None	51.82	89.04
	MD	39.14	71.67
	NMF-MD	**29.74**	**59.13**
VII, MC	None	29.60	56.58
	MD	27.25	51.31
	NMF-MD	**24.11**	**47.06**
VII, 8-ch.	None	19.76	40.21
	MD	19.40	38.28
	NMF-MD	**17.80**	**34.79**
IX, MC	None	11.69	28.99
	NMF	10.17	25.50
	NMF-MD	10.04	25.40
	NMF-DM	**9.94**	**25.29**
IX, 8-ch.	NMF-MD	7.66	**17.65**
	NMF-DM	**7.61**	18.03

Table 6.2. Word error rates (%) for the simulated and real data evaluation sets of the REVERB challenge. For Publication VII, shown are three separate systems: a baseline ASR system trained on clean speech acoustic models ("VII, clean"), a similar system trained on a multi-condition training set ("VII, MC") and a system trained on 8-channel microphone array signals preprocessed with a delay-sum beamformer and including CMLLR speaker adaptation ("VII, 8-ch."). Each system was tested with no feature enhancement ("None"), the missing data system described in Section 6.1 ("MD") and the NMF-based feature enhancement system initialized by the missing data estimate ("NMF-MD"). In the case of Publication IX, only multi-condition trained systems are included. Both the "IX, MC" and "IX, 8-ch." systems use the *LDA+MLLT+SAT+f-bMMI* back-end recognizer, described fully in the publication, with the latter system applying the delay-sum beamforming to the microphone array recordings before feature enhancement. Presented results are using no feature enhancement ("None"), or the NMF-based feature enhancement initialized with the reverberant observation ("NMF"), the missing data estimate ("NMF-MD"), or the distribution matching method ("NMF-DM").

nition performance over both the baseline system and the missing data dereverberation results that are used for the NMF initialization. While the highest relative reduction of the word error rate is obtained with models trained with clean speech, improvement is seen for all three systems.

A thorough evaluation using several different back-end ASR systems is carried out in Publication IX. Results from the *LDA+MLLT+SAT+f-bMMI* system are included in Table 6.2. This system uses a feature representation based on 13 mel-frequency cepstral coefficients of nine consecutive feature frames, reduced to a 40-dimensional feature vector with linear discriminant analysis (LDA), followed by a maximum likelihood linear transform (MLLT). Per-utterance speaker adaptive training (SAT), based on feature-space maximum-likelihood linear regression adaptation, is used to reduce the inter-speaker variability in the training set. The system is trained with discriminative training, using the feature-space boosted maximum mutual information (f-bMMI) criterion.

Both the missing data and distribution matching methods, described in Section 6.1 and Section 6.3, respectively, are investigated in Publication IX as possible sources for the initial estimate used by the NMF-based dereverberation algorithm. In the experiments, the distribution matching initialization generally outperforms the missing data approach by a small margin. Using the reverberant observation itself as the initial estimate also achieves relatively good performance.

Applying the NMF-based feature enhancement on dry speech with no reverberation degrades the speech recognition results, causing the system to perform worse than the baseline when dry speech is also used to train the acoustic models. This effect is not seen for the missing data dereverberation. With multi-condition training, both front-end methods outperform the baseline, as they decrease the mismatch between the training and test data sets. Detailed results on clean speech data are provided in Table 4 of Publication VII.

The activation matrix filtering step, which implicitly assumes the presence of some reverberation in the input observation, is a likely reason for the distortion of dry speech features, but also seems crucial for effective dereverberation. In small-scale experiments, omitting the filtering removes the observed distortion, but also causes the NMF-based method to perform no better than the missing data initialization alone. More flexible methods such as adjusting the strength of the activation matrix filter based on the input signal remain a possible topic for future work.

While the activation matrix filter does have built-in assumptions about temporal modulation patterns of the activation matrix typical of reverberation, it has not been necessary to adjust the filter for the different room conditions present in the REVERB challenge data sets. Similarly, the distribution matching method of Section 6.3, used for initialization, makes only weak assumptions about the effects of reverberation on the signal. As a result, the NMF-based system requires no supervised training data to cope with different environments, and compares favorably with respect to generalization against other proposed dereverberation methods in the experiments of Publication IX.

Some of the methods presented in Chapter 4 and Chapter 5 of this thesis, in the context of noisy speech, may be applicable as extensions to the NMF-based dereverberation. The sparse source separation feature enhancement described in Section 4.2 makes use of separate clean speech and noise dictionaries. A similar approach could be used to perform joint removal of noise and reverberation, although the shared reverberation filter matrix $\boldsymbol{R}$ may present difficulties. Similar heuristic uncertainty estimates as those presented in Chapter 5 could be applied also in the context of reverberant speech.

6.3 Distribution Matching

Publication VIII proposes a novel dereverberation method based on the principle of histogram equalization. This method is also used in Publication IX as an initial estimate for the NMF-based dereverberation system. In order to account for the long-term temporal effects of reverberation, the processing is done on stacked vectors of several contiguous temporal frames, as in Section 6.2. A decorrelating principal component analysis (PCA) transformation is also applied to the stacked vectors.

Dharanipragada and Padmanabhan (2000) and Saon et al. (2004) present similar approaches for speaker and environment adaptation by a nonlinear transformation of features. The transformations are specified to make the feature distribution match, respectively, that of the training data or a fixed Gaussian distribution. Both make the simplifying assumption that the dimensions of the features are statistically independent, in order to treat the histogram equalization as a one-dimensional problem. The primary difference of Publication VIII to these works is the long time context incorporated by the stacked vectors of multiple frames. As the consecu-

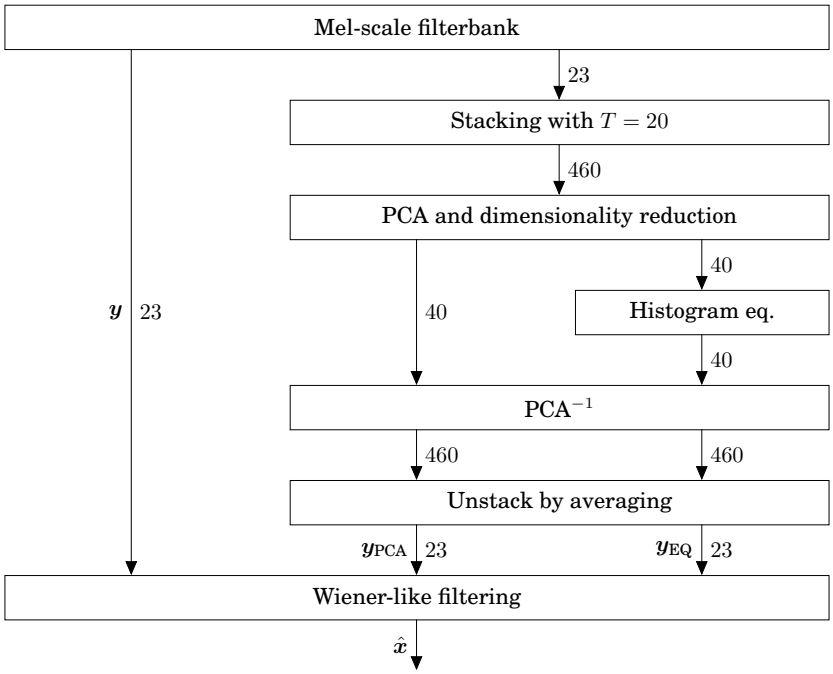


Figure 6.2. Data flow of the distribution matching dereverberation method. The edge labels indicate the dimensionality of data between each processing step. The symbols used in Equation (6.18) are also shown in the figure.

tive frames are highly correlated, the simplifying assumption of statistical independence is obviously violated for the stacked supervectors. Consequently, the histogram equalization is performed in a lower-dimensional decorrelated space.

Figure 6.2 outlines the processing steps of the distribution matching feature enhancement method. The input features are 23-dimensional mel-scale log-spectral vectors. The stacking operation is done in overlapping windows of $T = 20$ frames, with a frame step of 1, and the corresponding unstack operation averages over all windows containing a particular frame. The shared principal component subspace used by the "PCA" and the inverse "PCA^{-1}" blocks is obtained from clean, dry speech training data.

The histogram equalization step is performed by a set of separate look-up tables for each of the 40 PCA dimensions. The look-up tables map the input values so that the estimated cumulative distribution function (CDF) of the input data matches the estimate derived from the training data. Notably, as the process is based on overall distribution estimates only, no paired reverberant and dry speech data is required.

System	FE	Simulated	Real
Devel	None	51.81	88.51
	No filter	59.15	79.83
	$T = 1$	44.17	71.03
	Full PCA	38.92	67.75
	DM	**36.60**	**63.57**
Eval	None	51.82	89.04
	DM	**37.87**	**72.25**
Eval, CMLLR	None	39.40	81.68
	DM	**28.70**	**62.24**

Table 6.3. Word error rates (%) for the REVERB challenge data sets. The three systems all use the REVERB challenge baseline recognizer, trained and tested with either the development ("Devel") or the evaluation ("Eval") data set. In the latter case, results are shown both without and with per-utterance CMLLR adaptation ("Eval, CMLLR"). The tested feature enhancement methods are the baseline with no processing ("None") and the full distribution matching system ("DM"). For the development system, the contribution of the individual components of the system was investigated by leaving each component out: the Wiener-like filter ("No filter"), the long temporal context ("$T = 1$") or the dimensionality reduction by the PCA transformation ("Full PCA").

Finally, Wiener-like filtering is used to generate the enhanced features. While the histogram equalization step is efficient in reducing the effects of reverberation, the other processing steps, in particular the dimensionality reduction performed by the PCA, have the effect of smoothing out fine-grained temporal structure present in the original signal. The output features are therefore generated by filtering the original mel-spectral features with a per-frame filter based on the ratio between processed features with and without the histogram equalization. In the log-spectral domain, this operation has the form

$$\hat{\boldsymbol{x}} = \boldsymbol{y} + \left(\boldsymbol{y}_{\mathrm{EQ}} - \boldsymbol{y}_{\mathrm{PCA}}\right), \tag{6.18}$$

where, as indicated by Figure 6.2, $\boldsymbol{y}$ and $\hat{\boldsymbol{x}}$ are, respectively, the input and output features. The processed features with and without the histogram equalization are denoted as $\boldsymbol{y}_{\mathrm{EQ}}$ and $\boldsymbol{y}_{\mathrm{PCA}}$, respectively. This process is similar to the filtering performed for the NMF-based dereverberation system in Section 6.2, where a filter constructed from the NMF-based reconstructions of both the reverberant and dry speech components is used.

The key results of Publication VIII are summarized in Table 6.3. The development data set results highlight the importance of the individual components of the proposed system, as they show degraded speech recog-

nition performance when the multi-frame time context, dimensionality reduction in the PCA transformation, or the postprocessing filter is left out. The $T = 20$ frame temporal context also outperforms, e.g., the setting $T = 10$, demonstrating the utility of the long time context. The evaluation set results also show that the method combines well with CMLLR adaptation.

The implementation in Publication VIII uses *full-batch* processing, where the histogram of the reverberant test set is computed based on the entire test set. Based on small-scale development experiments, however, this is only marginally better than obtaining the histogram from a single utterance. The computational cost of the method could likely be reduced without significantly affecting the performance by using a smaller amount of adaptation data. After collecting a sufficient amount of data in the target environment, the method could also be applied for online processing, with a latency of only a single spectral frame.

An important advantage of the proposed method is that it makes very few assumptions about the nature and amount of reverberation present in the observed signal. In the dereverberation algorithm, the effect of reverberation is assumed to be convolutive, and the chosen time context T is expected to be long enough so that the reverberation can be described as a matrix multiplication on the stacked supervector. Finally, the PCA transformation learned from clean speech data is assumed to generalize so that the components of the transformed vectors of reverberant observations can be treated individually. Notably, no particular training is required for different environmental conditions, as the algorithm adapts based on the collected histogram data.

7. Conclusions

The existence of automatic speech recognition systems that achieve near human-level performance in favorable conditions, combined with the call for practical solutions for new real-world applications, has increased interest in research on speech recognition algorithms that are robust to noise and reverberation. Of the possible approaches, this work focuses on the methods of providing a speech recognition system with sequences of input features that change as little as possible in different environments, or, failing that, inform the system when the provided values are uncertain.

As they involve the basis of almost all speech processing algorithms, the spectral envelope estimation methods discussed in Chapter 3 incorporate robustness to distortion in a way that involves minimal changes to existing systems. The fact that the methods are generally entirely decoupled from the rest of the system and perform the same processing for each segment of the input signal imply certain limitations in how invariant to noise the resulting envelope estimate can be. However, the flexibility inherent in the selection of the weight function for the weighted linear prediction and its later extensions provides additional possibilities of adapting to the input signal. The behavior of the short-time energy and absolute value sum weight functions used with weighted and extended weighted linear prediction can be controlled by adjusting the window length parameter (Pohjalainen et al., 2010). In addition, Pohjalainen and Alku (2012, 2013) propose further weight functions that emphasize structures present in a single analysis frame.

By contrast, the feature enhancement paradigm is more general, allowing for arbitrary processing of the input features, while still being applicable to a wide class of back-end speech recognition systems. In this thesis, a major point of focus is on source separation by non-negative matrix fac-

torization. The use of the sparse representation of the input signal as a combination of selected sample spectrograms places the method in the class of *non-parametric*, or *exemplar-based* methods, in a marked departure from conventional tools used for modeling the acoustics of speech. While Chapter 4 and the publications of this work consider only the case of simple feature enhancement, the framework can also be used more directly to act as an acoustic model for a speech recognition system. In a *sparse classification* system, the state likelihoods used in the decoding stage are based on the activation vectors and state labels attached to the dictionary atoms (Gemmeke et al., 2011).

The ways to incorporate information about the uncertainty of the input, considered in Chapter 5, are a natural complement to a feature enhancement system, as any form of processing is bound to include some amount of residual error. Uncertainty processing also provides a way for speech recognition to focus on specific parts of the input signal, and therefore has close ties to the missing data framework. A major contribution of this thesis is the use of observation uncertainty methods for the source separation feature enhancement, based on heuristic uncertainty metrics derived from the sparse representation of speech. Due to the non-probabilistic nature of the feature enhancement system, the proposed heuristics are somewhat ad hoc. However, they can be combined with uncertainty propagation methods (Astudillo, 2010), in order to handle the feature transformations employed by the speech recognition system in a more principled way.

While reverberation is merely a special case of general distortion, it has an inherent structure that allows for more effective handling by specialized methods. It is also a case of particular importance for any application that requires accurate *far-field* speech recognition, where the microphone cannot be positioned close to the speaker. Prominent scenarios are speech recognition for meetings, or by an entertainment system in the living room. To cope with reverberation, this thesis proposes an extension of the non-negative matrix factorization model that incorporates a convolutional filter, described in Chapter 6. The filter is adapted to represent the effect of reverberation present in the input signal as part of the optimization of the factorization. No reverberant training data is needed by the method.

Due to the remarkable level of success achieved, the renaissance of neural networks observed in the field of speech recognition cannot be ig-

nored. While their rise can be partially attributed to increase in computing power, allowing for ever bigger models and larger data sets, the new advanced training and regularization methods are also a key factor. The results have raised the question of how useful built-in domain-specific knowledge about the properties of speech is, and to which degree such information can be learned simply by general neural network training procedures.

In theory, the front-end methods discussed in this work can be freely applied in conjunction with models of speech based on deep neural networks, simply by adjusting the input features. For example, enhancing speech using a system similar to the NMF source separation presented in Section 4.2 has been observed to improve the performance of such acoustic models (Baby et al., 2014). Uncertainty estimates can also be used to provide the network with information about the reliability of input (Abdelaziz et al., 2015; Huemmer et al., 2015; Tachioka and Watanabe, 2015). In practice, it remains an open question as to which methods will still prove beneficial, and which perform operations that can be automatically deduced from the available training data by the networks.

References

Abdelaziz, A. H., Watanabe, S., Hershey, J. R., Vincent, E. and Kolossa, D. (2015). Uncertainty propagation through deep neural networks, *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden, Germany.

Acero, A., Deng, L., Kristjansson, T. T. and Zhang, J. (2000). HMM adaptation using vector Taylor series for noisy speech recognition, *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000 / Interspeech 2000)*, Beijing, China, pp. 869–872.

Arrowood, J. A. (2003). *Using observation uncertainty for robust speech recognition*, PhD thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology.

Astudillo, R. F. (2010). *Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition*, PhD thesis, Technische Universität Berlin.

Astudillo, R. F., Kolossa, D., Abad, A., Zeiler, S., Saeidi, R., Mowlaee, P., da Silva Neto, J. P. and Martin, R. (2013). Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments, *Computer Speech & Language* **27**(3): 837–850. doi: 10.1016/j.csl.2012.07.009

Astudillo, R. F., Kolossa, D., Mandelartz, P. and Orglmeister, R. (2010). An uncertainty propagation approach to robust ASR using the ETSI advanced front-end, *IEEE Journal of Selected Topics in Signal Processing* **4**(5): 824–833. doi: 10.1109/JSTSP.2010.2057194

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *The Journal of the Acoustical Society of America* **55**(6): 1304–1312. doi: 10.1121/1.1914702

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave, *The Journal of the Acoustical Society of America* **50**(2B): 637–655. doi: 10.1121/1.1912679

Aubert, X. L. (2002). An overview of decoding techniques for large vocabulary continuous speech recognition, *Computer Speech & Language* **16**(1): 89–114. doi: 10.1006/csla.2001.0185

Baby, D., Gemmeke, J. F., Virtanen, T. and Van hamme, H. (2014). Exemplar-based speech enhancement for deep neural network based automatic speech

recognition, *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, pp. 4485–4489. doi: 10.1109/ICASSP.2015.7178819

Bahl, L. R., Jelinek, F. and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5**(2): 179–190. doi: 10.1109/TPAMI.1983.4767370

Barker, T. and Virtanen, T. (2013). Non-negative tensor factorization of modulation spectrograms for monaural sound source separation, *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics* **41**(1): 164–171. doi: 10.1214/aoms/1177697196

Beeston, A. V. and Brown, G. J. (2013). Modelling reverberation compensation effects in time-forward and time-reversed rooms, *UK Speech Conference*, Cambridge, UK.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization, *Computational Statistics & Data Analysis* **52**(1): 155–173. doi: 10.1016/j.csda.2006.11.006

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, *Technical Report TR-97-021*, International Computer Science Institute, Berkeley, CA, USA.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis, *Proceedings of the IEEE* **57**(8): 1408–1418. doi: 10.1109/PROC.1969.7278

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling, *Computer Speech & Language* **13**(4): 359–393. doi: 10.1006/csla.1999.0128

Christensen, H., Barker, J., Ma, N. and Green, P. (2010). The CHiME corpus: a resource and challenge for Computational Hearing in Multisource Environments, *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Chiba, Japan.

Cooke, M., Green, P. and Crawford, M. (1994). Handling missing data in speech recognition, *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, Yokohama, Japan.

Cooke, M., Green, P., Josifovski, L. and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Communication* **34**(3): 267–285. doi: 10.1016/S0167-6393(00)00034-0

Dahl, G. E., Sainath, T. N. and Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout, *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, pp. 8609–8613. doi: 10.1109/ICASSP.2013.6639346

Dahl, G. E., Yu, D., Deng, L. and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **20**(1): 30–42. doi: 10.1109/TASL.2011.2134090

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech and Signal Processing* **28**(4): 357–366. doi: 10.1109/TASSP.1980.1163420

Delcroix, M., Nakatani, T. and Watanabe, S. (2009). Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing, *IEEE Transactions on Audio, Speech, and Language Processing* **17**(2): 324–334. doi: 10.1109/TASL.2008.2010214

Deng, L., Acero, A., Jiang, L., Droppo, J. and Huang, X. (2001). High-performance robust speech recognition using stereo training data, *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, UT, USA, pp. 301–304. doi: 10.1109/ICASSP.2001.940827

Deng, L., Droppo, J. and Acero, A. (2005). Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion, *IEEE Transactions on Speech and Audio Processing* **13**(3): 412–421. doi: 10.1109/TSA.2005.845814

Dharanipragada, S. and Padmanabhan, M. (2000). A nonlinear unsupervised adaptation technique for speech recognition, *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000 / Interspeech 2000)*, Beijing, China, pp. 556–559.

Dharanipragada, S., Yapanel, U. H. and Rao, B. D. (2007). Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method, *IEEE Transactions on Audio, Speech, and Language Processing* **15**(1): 224–234. doi: 10.1109/TASL.2006.876776

Droppo, J., Acero, A. and Deng, L. (2001). Efficient online acoustic environment estimation for FCDCN in a continuous speech recognition system, *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Vol. I, Salt Lake City, UT, USA, pp. 209–212. doi: 10.1109/ICASSP.2001.940804

Droppo, J., Acero, A. and Deng, L. (2002). Uncertainty decoding with SPLICE for noise robust speech recognition, *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, Vol. I, Orlando, FL, USA, pp. 57–60. doi: 10.1109/ICASSP.2002.5743653

Fant, G. (1960). *Acoustic theory of speech production*, Mouton & Co. N.V., Publishers, The Hague.

Fitzgerald, D., Cranitch, M. and Coyle, E. (2005). Non-negative tensor factorisation for sound source separation, *Proceedings of the Irish Signals and Systems Conference*, Dublin, Ireland.

Fletcher, H. (1940). Auditory patterns, *Reviews of Modern Physics* **12**(1): 47–65. doi: 10.1103/RevModPhys.12.47

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum, *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**(1): 52–59. doi: 10.1109/TASSP.1986.1164788

Gales, M. J. F. and Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework, *Computer Speech & Language* **10**(4): 249–264. doi: 10.1006/csla.1996.0013

Gales, M. J. F. and Young, S. J. (1996). Robust continuous speech recognition using parallel model combination, *IEEE Transactions on Speech and Audio Processing* **4**(5): 352–359. doi: 10.1109/89.536929

Gales, M. and Young, S. (2008). The application of hidden Markov models in speech recognition, *Foundations and Trends in Signal Processing* **1**(3): 195–304. doi: 10.1561/2000000004

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains, *IEEE Transactions on Speech and Audio Processing* **2**(2): 291–298. doi: 10.1109/89.279278

Gelbart, D. and Morgan, N. (2002). Double the trouble: handling noise and reverberation in far-field automatic speech recognition, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002 / Interspeech 2002)*, Denver, CO, USA.

Gemmeke, J. F., Remes, U. and Palomäki, K. J. (2010). Observation uncertainty measures for sparse imputation, *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Chiba, Japan, pp. 2262–2265.

Gemmeke, J. F. and Van hamme, H. (2011). An hierarchical exemplar-based sparse model of speech, with an application to ASR, *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, USA, pp. 101–106. doi: 10.1109/ASRU.2011.6163913

Gemmeke, J. F., Van hamme, H., Cranen, B. and Boves, L. (2010). Compressive sensing for missing data imputation in noise robust speech recognition, *IEEE Journal of Selected Topics in Signal Processing* **4**: 272–287. doi: 10.1109/JSTSP.2009.2039171

Gemmeke, J. F., Virtanen, T. and Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(7): 2067–2080. doi: 10.1109/TASL.2011.2112350

Gemmeke, J. F., Wang, Y., Segbroeck, M. V., Cranen, B. and Van hamme, H. (2009). Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases, *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, pp. 1227–1230.

Grönroos, S.-A., Virpioja, S., Smit, P. and Kurimo, M. (2014). Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology, *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland.

Harding, S., Barker, J. and Brown, G. J. (2006). Mask estimation for missing data speech recognition based on statistics of binaural interaction, *IEEE Transactions on Audio, Speech, and Language Processing* **14**(1): 58–67. doi: 10.1109/TSA.2005.860354

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America* **87**(4): 1738–1752. doi: 10.1121/1.399423

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech, *IEEE Transactions on Speech and Audio Processing* **2**(4): 578–589. doi: 10.1109/89.326616

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets, *Neural Computation* **18**(7): 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S. and Pylkkönen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish, *Computer Speech & Language* **20**(4): 515–541. doi: 10.1016/j.csl.2005.07.002

Hirsimäki, T. and Kurimo, M. (2004). Decoder issues in unlimited Finnish speech recognition, *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG 2004)*, Espoo, Finland.

Hirsimäki, T., Pylkkönen, J. and Kurimo, M. (2009). Importance of high-order N-gram models in morph-based speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **17**(4): 724–732. doi: 10.1109/TASL.2008.2012323

Hopgood, J. R. and Rayner, P. J. W. (2003). Blind single channel deconvolution using nonstationary signal processing, *IEEE Transactions on Speech and Audio Processing* **11**(5): 476–488. doi: 10.1109/TSA.2003.815522

Hori, T. and Nakamura, A. (2013). *Speech Recognition Algorithms Using Weighted Finite-State Transducers*, Vol. 10 of *Synthesis Lectures on Speech and Audio Processing*, Morgan & Claypool, San Francisco, CA, USA. doi: 10.2200/S00462ED1V01Y201212SAP010

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints, *The Journal of Machine Learning Research* **5**: 1457–1469.

Huemmer, C., Maas, R., Schwarz, A., Astudillo, R. F. and Kellermann, W. (2015). Uncertainty decoding for DNN-HMM hybrid systems based on numerical sampling, *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden, Germany.

Hurmalainen, A., Gemmeke, J. F. and Virtanen, T. (2013). Modelling non-stationary noise with spectral factorisation in automatic speech recognition, *Computer Speech & Language* **27**(3): 763–779. doi: 10.1016/j.csl.2012.07.008

Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F. and Kiessling, A. (2002). SPEECON - speech databases for consumer devices: Database specification and validation, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, pp. 329–333.

Jelinek, F. (1969). Fast sequential decoding algorithm using a stack, *IBM Journal of Research and Development* **13**(6): 675–685. doi: 10.1147/rd.136.0675

Juang, B. H., Levinson, S. E. and Sondhi, M. M. (1986). Maximum likelihood estimation for multivariate mixture observations of Markov chains (corresp.), *IEEE Transactions on Information Theory* **32**(2): 307–309. doi: 10.1109/TIT.1986.1057145

Kalinli, O., Seltzer, M. L., Droppo, J. and Acero, A. (2010). Noise adaptive training for robust automatic speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **18**(8): 1889–1901. doi: 10.1109/TASL.2010.2040522

Kameoka, H., Nakatani, T. and Yoshioka, T. (2009). Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms, *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, pp. 45–48. doi: 10.1109/ICASSP.2009.4959516

Karhila, R., Remes, U. and Kurimo, M. (2014). Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments, *IEEE Journal of Selected Topics in Signal Processing* **8**(2): 285–295. doi: 10.1109/JSTSP.2013.2278492

Keronen, S., Cho, K., Raiko, T., Ilin, A. and Palomäki, K. (2013). Gaussian-Bernoulli restricted Boltzmann machines and automatic feature extraction for noise robust missing data mask estimation, *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, pp. 6729–6733. doi: 10.1109/ICASSP.2013.6638964

Keronen, S., Pohjalainen, J., Alku, P. and Kurimo, M. (2011). Noise robust feature extraction based on extended weighted linear prediction in LVCSR, *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, pp. 1265–1268.

Keronen, S., Remes, U., Kallasjoki, H. and Palomäki, K. (2013). Noise robust missing data mask estimation based on automatically learned features, *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME 2013)*, Vancouver, Canada.

Kingsbury, B. E., Morgan, N. and Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram, *Speech Communication* **25**(1-3): 117–132. doi: 10.1016/S0167-6393(98)00032-6

Kinoshita, K., Delcroix, M., Nakatani, T. and Miyoshi, M. (2009). Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction, *IEEE Transactions on Audio, Speech, and Language Processing* **17**(4): 534–545. doi: 10.1109/TASL.2008.2009015

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., Gannot, S. and Raj, B. (2013). The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech, *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2013)*, New Paltz, NY, USA, pp. 1–4. doi: 10.1109/WASPAA.2013.6701894

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling, *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, Detroit, MI, USA, pp. 181–184. doi: 10.1109/ICASSP.1995.479394

Kolossa, D., Astudillo, R. F., Hoffmann, E. and Orglmeister, R. (2010). Independent component analysis and time-frequency masking for speech recognition in multitalker conditions, *EURASIP Journal on Audio, Speech, and Music Processing* . Article ID 651420. doi: 10.1155/2010/651420

Krueger, A. and Haeb-Umbach, R. (2010). Model-based feature enhancement for reverberant speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **18**(7): 1692–1707. doi: 10.1109/TASL.2010.2049684

Kumar, K., Singh, R., Raj, B. and Stern, R. (2011). Gammatone subband magnitude-domain dereverberation for asr, *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011*, Prague, Czech Republic, pp. 4604–4607. doi: 10.1109/ICASSP.2011.5947380

Lebart, K., Boucher, J. M. and Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation, *Acta Acustica united with Acustica* **87**(3): 359–366.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, MIT Press, pp. 556–562.

Lee, K.-F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing* **38**(4): 599–609. doi: 10.1109/29.52701

Lee, L. and Rose, R. (1998). A frequency warping approach to speaker normalization, *IEEE Transactions on Speech and Audio Processing* **6**(1): 49–60. doi: 10.1109/89.650310

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech & Language* **9**(2): 171–185. doi: 10.1006/csla.1995.0010

Lehmann, E. A. and Johansson, A. M. (2008). Prediction of energy decay in room impulse responses simulated with an image-source model, *The Journal of the Acoustical Society of America* **124**(1): 269–277. doi: 10.1121/1.2936367

Liao, H. and Gales, M. (2007). Adaptive training with joint uncertainty decoding for robust recognition of noisy data, *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Vol. IV, Honolulu, HI, USA, pp. 389–392. doi: 10.1109/ICASSP.2007.366931

Liao, H. and Gales, M. (2008). Issues with uncertainty decoding for noise robust automatic speech recognition, *Speech Communication* **50**(4): 265–277. doi: 10.1016/j.specom.2007.10.004

Liao, H. and Gales, M. J. F. (2005). Joint uncertainty decoding for noise robust speech recognition, *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech 2005 / Interspeech 2005)*, Lisbon, Portugal, pp. 3129–3132.

Lippmann, R. P. (1997). Speech recognition by machines and humans, *Speech Communication* **22**(1): 1–15. doi: 10.1016/S0167-6393(97)00021-6

Ma, C., Kamp, Y. and Willems, L. F. (1993). Robust signal selection for linear prediction analysis of voiced speech, *Speech Communication* **12**(1): 69–81. doi: 10.1016/0167-6393(93)90019-H

Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P. (2009). Stabilised weighted linear prediction, *Speech Communication* **51**(5): 401–411. doi: 10.1016/j.specom.2008.12.005

Makhoul, J. (1975). Linear prediction: A tutorial review, *Proceedings of the IEEE* **63**(4): 561–580. doi: 10.1109/PROC.1975.9792

Mansikkaniemi, A. and Kurimo, M. (2015). Adaptation of morph-based speech recognition for foreign names and acronyms, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(5): 941–950. doi: 10.1109/TASLP.2015.2414818

Mohri, M., Pereira, F. and Riley, M. (2002). Weighted finite-state transducers in speech recognition, *Computer Speech & Language* **16**(1): 69–88. doi: 10.1006/csla.2001.0184

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing*, sixth edn, Brill.

Moreno, P. J., Raj, B. and Stern, R. M. (1996). A vector Taylor series approach for environment-independent speech recognition, *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, Vol. 2, Atlanta, GA, USA, pp. 733–736. doi: 10.1109/ICASSP.1996.543225

Murthi, M. N. and Rao, B. D. (2000). All-pole modeling of speech based on the minimum variance distortionless response spectrum, *IEEE Transactions on Speech and Audio Processing* **8**(3): 221–239. doi: 10.1109/89.841206

Musicus, B. R. (1985). Fast MLM power spectrum estimation from uniformly spaced correlations, *IEEE Transactions on Acoustics, Speech and Signal Processing* **33**(5): 1333–1335. doi: 10.1109/TASSP.1985.1164696

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M. and Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction, *IEEE Transactions on Audio, Speech, and Language Processing* **18**(7): 1717–1731. doi: 10.1109/TASL.2010.2052251

Ney, H., Haeb-Umbach, R., Tran, B.-H. and Oerder, M. (1992). Improvements in beam search for 10000-word continuous speech recognition, *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, San Francisco, CA, USA, pp. 9–12. doi: 10.1109/ICASSP.1992.225985

Ozerov, A., Lagrange, M. and Vincent, E. (2013). Uncertainty-based learning of acoustic models from noisy data, *Computer Speech & Language* **27**(3): 874–894. doi: 10.1016/j.csl.2012.07.002

Palomäki, K. J., Brown, G. J. and Barker, J. P. (2004). Techniques for handling convolutional distortion with 'missing data' automatic speech recognition, *Speech Communication* **43**(1-2): 123–142. doi: 10.1016/j.specom.2004.02.005

Palomäki, K. J., Brown, G. J. and Barker, J. P. (2006). Recognition of reverberant speech using full cepstral features and spectral missing data, *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, Vol. I, Toulouse, France, pp. 289–292. doi: 10.1109/ICASSP.2006.1660014

Pereira, F., Riley, M. and Sproat, R. (1994). Weighted rational transductions and their application to human language processing, *Proceedings of the Workshop on Human Language Technology (HLT '94)*, Plainsboro, NJ, USA, pp. 262–267. doi: 10.3115/1075812.1075870

Picone, J. W. (1993). Signal modeling techniques in speech recognition, *Proceedings of the IEEE* **81**(9): 1215–1247. doi: 10.1109/5.237532

Pohjalainen, J. and Alku, P. (2012). Robust speech analysis by lag-weighted linear prediction, *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, pp. 4453–4456. doi: 10.1109/ICASSP.2012.6288908

Pohjalainen, J. and Alku, P. (2013). Extended weighted linear prediction using the autocorrelation snapshot - a robust speech analysis method and its applications to recognition of vocal emotions, *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France.

Pohjalainen, J., Saeidi, R., Kinnunen, T. and Alku, P. (2010). Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions, *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Chiba, Japan, pp. 1477–1480.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwartz, P., Silovský, J., Stemmer, G. and Veselý, K. (2011). The Kaldi speech recognition toolkit, *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, USA.

Pylkkönen, J. (2005). An efficient one-pass decoder for finnish large vocabulary continuous speech recognition, *Proceedings of the 2nd Baltic Conference on Human Language Technologies (HLT 2005)*, Tallinn, Estonia, pp. 167–172.

Pylkkönen, J. and Kurimo, M. (2012). Analysis of extended Baum–Welch and constrained optimization for discriminative training of HMMs, *IEEE Transactions on Audio, Speech, and Language Processing* **20**(9): 2409–2419. doi: 10.1109/TASL.2012.2203805

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2): 257–286. doi: 10.1109/5.18626

Raj, B., Seltzer, M. L. and Stern, R. M. (2004). Reconstruction of missing features for robust speech recognition, *Speech Communication* **43**(4): 275–296. doi: 10.1016/j.specom.2004.03.007

Raj, B. and Stern, R. M. (2005). Missing-feature approaches in speech recognition, *IEEE Signal Processing Magazine* **22**(5): 101–116. doi: 10.1109/MSP.2005.1511828

Reddy, D. R. (1976). Speech recognition by machine: A review, *Proceedings of the IEEE* **64**(4): 501–531. doi: 10.1109/PROC.1976.10158

Remes, U., López, A. R., Palomäki, K. and Kurimo, M. (2015). Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(7): 1198–1208. doi: 10.1109/TASLP.2015.2424322

Remes, U., Nankaku, Y. and Tokuda, K. (2011). GMM-based missing-feature reconstruction on multi-frame windows, *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, pp. 1665–1668.

Robinson, T., Fransen, J., Pye, D., Foote, J. and Renals, S. (1995). WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition, *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, Detroit, MI, USA, pp. 81–84.

Sainath, T., Kingsbury, B., Mohamed, A.-R., Saon, G. and Ramabhadran, B. (2014). Improvements to filterbank and delta learning within a deep neural network framework, *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy.

Sainath, T. N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., Compernolle, D. V., Demunyck, K., Gemmeke, J. F., Bellegarda, J. R. and Sundaram, S. (2012). Exemplar-based processing for speech recognition: An overview, *IEEE Signal Processing Magazine* **29**(6): 98–113. doi: 10.1109/MSP.2012.2208663

Sambur, M. R. and Jayant, N. S. (1976). Lpc analysis/synthesis from speech inputs containing quantizing noise or additive white noise, *IEEE Transactions on Acoustics, Speech and Signal Processing* **24**(6): 488–494. doi: 10.1109/TASSP.1976.1162870

Saon, G., Dharanipragada, S. and Povey, D. (2004). Feature space Gaussianization, *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Canada, pp. 329–332. doi: 10.1109/ICASSP.2004.1325989

Schwartz, R., Chow, Y., Rouoos, S., Krasner, M. and Makhoul, J. (1984). Improved hidden Markov modeling of phonemes for continuous speech recognition, *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '84)*, Vol. 9, San Diego, CA, USA, pp. 21–24. doi: 10.1109/ICASSP.1984.1172751

Seide, F., Li, G., Chen, X. and Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription, *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, USA, pp. 24–29. doi: 10.1109/ASRU.2011.6163899

Seltzer, M. L., Raj, B. and Stern, R. M. (2004a). A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition, *Speech Communication* **43**(4): 379–393. doi: 10.1016/j.specom.2004.03.006

Seltzer, M. L., Raj, B. and Stern, R. M. (2004b). Likelihood-maximizing beamforming for robust hands-free speech recognition, *IEEE Transactions on Speech and Audio Processing* **12**(5): 489–498. doi: 10.1109/TSA.2004.832988

Shao, Y., Srinivasan, S., Jin, Z. and Wang, D. (2010). A computational auditory scene analysis system for speech segregation and robust speech recognition, *Computer Speech & Language* **24**(1): 77–93. doi: 10.1016/j.csl.2008.03.004

Siivola, V., Hirsimäki, T., Creutz, M. and Kurimo, M. (2003). Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner, *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003 / Interspeech 2003)*, Geneva, Switzerland, pp. 2293–2296.

Siivola, V., Hirsimäki, T. and Virpioja, S. (2007). On growing and pruning Kneser–Ney smoothed N-gram models, *IEEE Transactions on Audio, Speech, and Language Processing* **15**(5): 1617–1624. doi: 10.1109/TASL.2007.896666

Smaragdis, P. (2000). Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, *Independent Component Analysis and Blind Signal Separation*, Vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499. doi: 10.1007/978-3-540-30110-3_63

Sorenson, H. W. and Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums, *Automatica* **7**(4): 465–479. doi: 10.1016/0005-1098(71)90097-5

Srinivasan, S. and Wang, D. (2006). A supervised learning approach to uncertainty decoding for robust speech recognition, *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, Vol. I, Toulouse, France, pp. 297–300. doi: 10.1109/ICASSP.2006.1660016

Srinivasan, S. and Wang, D. (2007). Transforming binary uncertainties for robust speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* **15**(7): 2130–2140. doi: 10.1109/TASL.2007.901836

Stevens, S. S. (1957). On the psychophysical law, *The Psychological Review* **64**(3): 153–181. doi: 10.1037/h0046162

Tachioka, Y. and Watanabe, S. (2015). Uncertainty training and decoding methods of deep neural networks based on stochastic representation of enhanced features, *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden, Germany.

Takiguchi, T., Nishimura, M. and Ariki, Y. (2006). Acoustic model adaptation using first-order linear prediction for reverberant speech, *IEICE Transactions on Information and Systems* **E89-D**(3): 908–914.

Turunen, V. T. and Kurimo, M. (2011). Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition and retrieval, *ACM Transactions on Speech and Language Processing* **8**(1): 1–25. doi: 10.1145/2036916.2036917

Vincent, E., Barker, J., Watanabe, S., Roux, J. L., Nesta, F. and Matassoni, M. (2013). The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines, *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada.

Virtanen, T., Raj, B., Gemmeke, J. F. and Van hamme, H. (2014). Active-set newton algorithm for non-negative sparse coding of audio, *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, Florence, Italy, pp. 3092–3096. doi: 10.1109/ICASSP.2014.6854169

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory* **13**(2): 260–269. doi: 10.1109/TIT.1967.1054010

Wong, D. Y., Markel, J. D. and Gray, Jr., A. H. (1979). Least squares glottal inverse filtering from the acoustic speech waveform, *IEEE Transactions on Acoustics, Speech and Signal Processing* **27**(4): 350–355. doi: 10.1109/TASSP.1979.1163260

Wölfel, M. (2009). Enhanced speech features by single-channel joint compensation of noise and reverberation, *IEEE Transactions on Audio, Speech, and Language Processing* **17**(2): 312–323. doi: 10.1109/TASL.2008.2009161

Yapanel, U. H. and Hansen, J. H. (2008). A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition, *Speech Communication* **50**(2): 142–152. doi: 10.1016/j.specom.2007.07.006

Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T. and Kellermann, W. (2012). Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition, *IEEE Signal Processing Magazine* **29**(6): 114–126. doi: 10.1109/MSP.2012.2205029

Young, S. (1996). A review of large-vocabulary continuous-speech recognition, *IEEE Signal Processing Magazine* **13**(5): 45–57. doi: 10.1109/79.536824

Young, S. J. (1993). The HTK hidden Markov model toolkit: Design and philosophy, *Technical report*, University of Cambridge, Department of Engineering.

Young, S. J., Odell, J. J. and Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling, *Proceedings of the Workshop on Human Language Technology (HLT '94)*, Plainsboro, NJ, USA, pp. 307–312. doi: 10.3115/1075812.1075885

Young, S. J., Russell, N. H. and Thornton, J. H. S. (1989). Token passing: a simple conceptual model for connected speech recognition systems, *Technical report CUED/FINFENG/TR38*, Cambridge University.

Yu, D., Seltzer, M., Li, J., Huang, J.-T. and Seide, F. (2013). Feature learning in deep neural networks - studies on speech recognition, *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, Scottsdale, AZ, USA.

Lack of robustness is a major issue for the use of automatic speech recognition in many applications. While human listeners are capable of understanding speech even in challenging environments, conventional systems quickly degrade in accuracy when presented with a signal distorted by noise and reflected sounds. This thesis proposes several systems that enhance the sequence of input features, with the aim of making them more invariant to changes in the recording environment. Complementing these systems, methods to extract and use information about the varying uncertainty of the enhanced features are also investigated. The positive impact of the proposed approaches on the accuracy of speech recognition is confirmed by experimental evaluation on realistic large vocabulary continuous speech recognition tasks.

Aalto-DD 31/2016

ISBN 978-952-60-6665-3 (printed)
ISBN 978-952-60-6666-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi