Aalto University
School of Electrical
Engineering

Sami Keronen

APPROACHING HUMAN PERFORMANCE IN NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Thesis for the degree of Licentiate of Science in Technology submitted for inspection, Espoo, 27 January, 2014.

Supervising professor Professor Mikko Kurimo

Thesis advisor Kalle Palomäki, D.Sc. (Tech)

Aalto University

AALTO UNIVERSITY SCHOOLS OF TECHNOLOGY PO Box 11000, FI-00076 AALTO http://www.aalto.fi	ABSTRACT OF THE LICENTIATE THESIS

Author: Sami Keronen

Title: Approaching human performance in noise robust automatic speech recognition

School: Electrical Engineering | Department: Signal Processing and Acoustics

Research field: Speech and Language Technology | Code: ELECSCI005Z

Supervisor: Prof. Mikko Kurimo | Instructor(s): Kalle Palomäki, D.Sc. (Tech)

Examiner(s): Prof. Erkki Oja

Abstract:

Modern automatic speech recognition systems are able to achieve human-like performance on read speech in relatively noise-free environments. However, in the presence of heavily deteriorating noise, the gap between human and machine recognition remains large. The work presented in the thesis is aimed to enhance the speech recognition performance in varying noise and low signal-to-noise ratio conditions by improving the short-time spectral analysis of the speech signal and the spectrographic mask estimation in the missing data framework.

In the thesis, the fast Fourier transformation based spectrum estimation of Mel-frequency cepstral coefficients is substituted with extended weighted linear prediction. Temporal weighting in linear predictive analysis emphasizes the high amplitude samples that are assumed less corrupted by noise and attenuates the others. Extending the weighting to separately apply to each lag in the prediction of each sample arguably offers more modeling power for deteriorated speech. The extended weighted linear prediction is shown to exceed the recognition performance of conventional linear prediction, weighted linear prediction and fast Fourier transformation based feature extraction.

Missing data methods assume that only part of the spectro-temporal components of the deteriorated signal are corrupted by noise while the speech-dominant components hold the reliable information that can be used in recognition. Two spectrographic mask estimation techniques based on binary classification of features are proposed in the thesis. The first method is founded on a comprehensive set of design features and the second on the Gaussian-Bernoulli restricted Boltzmann machine that learns the feature set automatically. Both mask estimation methods are shown to outperform their respective reference mask estimation methods in recognition accuracy.

All the proposed noise robust techniques are immediately applicable to automatic speech recognition. With further refinement, the mask estimation methods could also be applied to hearing aids since they are able to attenuate the background noise thus increasing the speech intelligibility.

Date: 27.1.2014 | Language: English | Number of pages: 124

Keywords: noise robust, speech recognition, mask estimation, linear prediction, GRBM

Aalto-yliopisto

AALTO-YLIOPISTO TEKNIIKAN KORKEAKOULUT PL 11000, 00076 AALTO http://www.aalto.fi	LISENSIAATINTUTKIMUKSEN TIIVISTELMÄ

Tekijä: Sami Keronen

Työn nimi: Kohti ihmiskykyjä melusietoisessa automaattisessa puheentunnistuksessa

Korkeakoulu: Sähkötekniikan korkeakoulu	Laitos: Signaalinkäsittelyn ja akustiikan laitos
Tutkimusala: Puhe- ja kieliteknologia	Koodi: ELECSCI005Z
Työn vastuuprofessori: Prof. Mikko Kurimo	Työn ohjaaja(t): Kalle Palomäki, TkT

Työn ulkopuolinen tarkastaja(t): Prof. Erkki Oja

Tiivistelmä:

Nykyaikaiset automaattiset puheentunnistusjärjestelmät pystyvät tunnistamaan luettua puhetta vähämeluisissa käyttöympäristöissä lähes yhtä tarkasti kuin ihmiset, mutta kovassa taustamelussa ihmisen tunnistuskyky on huomattavasti konetta tehokkaampi. Tutkimuksessa esitetään menetelmiä puhesignaalin melusietoiseen spektrianalyysiin ja puuttuvan datan maskiestimointiin automaattisen puheentunnistuksen parantamiseksi melutyypiltään vaihtelevissa ja alhaisen signaalikohinasuhteen käyttöympäristöissä.

Tutkimuksessa parannetaan spektrianalyysin melusietoisuutta korvaamalla Mel-kepstrikerrointen laskennassa perinteisesti käytetty nopea Fourier-muunnos laajennetulla ja painotetulla lineaariprediktiolla. Aikatason painotuksella korostetaan suuriamplitudisten näytteiden tärkeyttä lineaariprediktio-analyysissä, sillä niiden oletetaan olevan suhteellisesti vähemmän korruptoituneita kuin pieniamplitudisten näytteiden. Laajentamalla painotusta kaikkiin viiveisiin näytteiden prediktiossa, voidaan meluisaa puhetta mallintaa joustavammin. Laajennetun ja painotetun lineaariprediktion näytetään parantavan Mel-kepstrikertoimiin pohjautuvan piirreirrotuksen melusietoisuutta nopeaan Fourier-muunnokseen, lineaariprediktioon ja painotettuun lineaariprediktioon verrattuna.

Puuttuvan datan menetelmät perustuvat oletukseen, että melu vääristää ainoastaan osan puhesignaalin aikataajuustason komponenteista, loppujen komponenttien säilyttäessä luotettavan puheinformaation, jota voidaan käyttää tunnistuksessa. Tutkimuksessa esitetään kaksi binääriluokittelupohjaista menetelmää maskien melusietoiseen estimointiin. Ensimmäinen estimointimenetelmä perustuu kattavaan käsintehtyjen piirteiden yhdistelmään ja toinen piirteiden automaattiseen oppimiseen Gaussian-Bernoulli rajoitetun Boltzmann koneen avulla. Tutkimuksessa osoitetaan molempien menetelmien parantavan puheentunnistustarkkuutta vastaaviin referenssimenetelmiin verrattuna.

Kaikki tutkimuksessa esitetyt menetelmät ovat välittömästi hyödynnettävissä automaattisissa puheentunnistusjärjestelmissä. Kiinnittämällä huomiota maskiestimointimenetelmien laskennallisiin vaatimuksiin, pystyttäisiin niitä soveltamaan myös kuulokojeissa, sillä menetelmillä voidaan vaimentaa taustamelua, mikä lisää puheen ymmärrettävyyttä.

Päivämäärä: 27.1.2014	Kieli: Englanti	Sivumäärä: 124

Avainsanat: melusietoinen, puheentunnistus, maskiestimointi, lineaariprediktio, GRBM

Preface

The research presented in the thesis has been carried out at the Department of Information and Computer Science of Aalto University School of Science during 2009–2013. The thesis itself has been compiled at the Department of Signal Processing and Acoustics of Aalto University School of Electrical Engineering. The work has been funded by Langnet doctoral programme, Academy of Finland through Adaptive Informatics Research Centre, Centre of Excellence in Computational Inference Research, and Finnish Computational Science research programme. I am grateful for the personal grants awarded by the Nokia Foundation, KAUTE Foundation, and Aalto University during the research period.

I praise Prof. Mikko Kurimo for giving the possibility to work in the speech group. His efforts in organizing the funding and supervising the thesis and several publications presented in it are highly appreciated. I express my gratitude to Doc. Kalle Palomäki for his invaluable research ideas, and for supervising and coauthoring in several publications presented in the thesis. The role of Prof. Erkki Oja, the examiner of the thesis, will not be forgotten.

The current and former colleagues Dr. Teemu Hirsimäki, Dr. Janne Pylkkönen, Dr. Ville Turunen, Ulpu Remes, Heikki Kallasjoki, and Cho KyungHyun all deserve big credits for their help and expertise in the field, and for all the research and non-research related discussions. I also want to give a warm hug to my family for giving the motivation to work hard.

Espoo, January 27, 2014,

Sami Keronen

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 Sami Keronen, Ulpu Remes, Kalle J. Palomäki, Tuomas Virtanen and Mikko Kurimo. Comparison of Noise Robust Methods in Large Vocabulary Speech Recognition. In *EUSIPCO 2010 – The 18th European Signal Processing Conference*, Aalborg, Denmark, pp. 1973–1977, August 2010.

II Sami Keronen, Jouni Pohjalainen, Paavo Alku and Mikko Kurimo. Noise Robust Feature Extraction Based on Extended Weighted Linear Prediction in LVCSR. In *INTERSPEECH 2011 – Proceedings of the 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, pp. 1265–1268, August 2011.

III 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 *CHiME – International Workshop on Machine Listening in Multisource Environments*, Florence, Italy, pp. 58–63, September 2011.

IV Sami Keronen, Heikki Kallasjoki, Ulpu Remes, Guy J. Brown, Jort F. Gemmeke and Kalle J. Palomäki. Mask Estimation and Imputation Methods for Missing Data Speech Recognition in a Multisource Reverberant Environment. *Computer Speech and Language*, vol. 27 no. 3, pp. 798–819, February 2013.

V Sami Keronen, KyungHyun Cho, Tapani Raiko, Alexander Ilin and Kalle J. Palomäki. Gaussian-Bernoulli Restricted Boltzmann Machines and Automatic Feature Extraction for Noise Robust Missing Data Mask Estimation. In *ICASSP 2013 – The 38th International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, pp. 6729–6733, May 2013.

VI Sami Keronen, Ulpu Remes, Heikki Kallasjoki and Kalle J. Palomäki. Noise Robust Missing Data Mask Estimation Based on Automatically Learned Features. In *CHiME 2013 – The 2nd International Workshop on Machine Listening in Multisource Environments*, Vancouver, Canada, pp. 77–78, June 2013.

Author's Contribution

Publication I: "Comparison of Noise Robust Methods in Large Vocabulary Speech Recognition"

In the article, three noise robust approaches are implemented and evaluated on a large vocabulary ASR task. The author implemented the data-driven parallel model combination and conducted the experiments except for the missing data related tests. The author wrote the first draft of the paper except for section 2.5.

Publication II: "Noise Robust Feature Extraction Based on Extended Weighted Linear Prediction in LVCSR"

A linear prediction based method for noise robust feature extraction is described in the paper. The author implemented the extended weighted linear prediction in the automatic speech recognition system, conducted all the experiments and wrote the first draft of the paper. The paper has also been accepted as an assignment to pass S-89.3690 special course in speech processing.

Publication III: "Mask Estimation and Sparse Imputation for Missing Data Speech Recognition in Multisource Reverberant Environments"

The paper describes a missing data mask estimation approach for enhanced noise robustness in adverse environments. The author implemented the binaural mask estimation and conducted the respective experiments. The author was responsible for writing sections 1. and 5. of the paper.

Publication IV: "Mask Estimation and Imputation Methods for Missing Data Speech Recognition in a Multisource Reverberant Environment"

The article describes a noise robust missing data method for low signal-to-noise ratio environments, proposes retraining acoustic models on imputed data, and extensively evaluates several design features. The author was responsible for the multifeature based mask estimation, conducted the statistical analyses except for analysis of the individual features, and conducted all the experiments related to cluster-based imputation. The author also carried out the experiments with binaural masks and sparsely imputed data provided by coauthors. The author was responsible for writing sections 1., 2.2.1., 2.2.3., 2.3., 3., 4.2., 5., appendix A, and parts of 2.2.2. of the paper.

Publication V: "Gaussian-Bernoulli Restricted Boltzmann Machines and Automatic Feature Extraction for Noise Robust Missing Data Mask Estimation"

The paper describes a method to automatically learn a set of features for missing data mask estimation. The author refined the concept of the work, assisted in merging the automatic speech recognition system with the neural network system, conducted all the experiments, and was responsible for writing the paper except for section 2.2.

Publication VI: "Noise Robust Missing Data Mask Estimation Based on Automatically Learned Features"

The article is an extension to Publication V. The work was mainly carried out by the author except for the initial preparation of data and writing parts of section 2.4. of the paper.

List of Abbreviations

ASR	Automatic speech recognition
CASA	Computational auditory scene analysis
CI	Cluster-based imputation
CMS	Cepstral mean subtraction
DPMC	Data-driven parallel model combination
FFT	Fast Fourier transform
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
MAP	Maximum a posteriori
MFCC	Mel-frequency cepstral coefficient
MLLR	Maximum likelihood linear regression
MLP	Multilayer perceptron
PLP	Perceptual linear prediction
PNCC	Power-normalized cepstral coefficient
RASTA	Relative spectral
SI	Sparse imputation
SNR	Signal-to-noise ratio
STE	Short-time energy
SVM	Support vector machine
TF	Time-frequency
WER	Word error rate
WLP	Weighted linear prediction
XLP	Extended weighted linear prediction

List of Symbols

a	Prediction coefficient
b_i	Bias of visible unit v_i
c_j	Bias of hidden unit h_j
C	GMM model scale factor
d	Index of frequency bin
h_j	Index of hidden unit
m	Effective length of the moving average memory
n	Sample index in time domain
n_v	Total number of visible units
n_h	Total number of hidden units
N	Total number of samples in a signal
$N(\tau, d)$	Log-Mel-spectral component of noise
$\boldsymbol{o}(\tau, d)$	Observed feature vector
p	Prediction order
$S(\tau, d)$	Log-Mel-spectral component of clean speech
T	Number of consecutive frames in imputation
v_i	Index of visible unit
w_{ij}	Weight connecting v_i to h_j
W	Weighting function
x	Time domain signal
X	Bandpass filtered signal
$Y(\tau, d)$	Log-Mel-spectral component of noisy speech
$\boldsymbol{Z}$	Normalizing constant
η	Learning rate
θ	Oracle mask reliability threshold
$\boldsymbol{\Theta}$	Set of GRBM parameters
σ	Standard deviation
τ	Index of time frame

1. Introduction

1.1 Background of automatic speech recognition

The automatic speech recognition (ASR) systems are designed to recognize the content of human speech and transform it into text form. ASR has been successfully adopted e.g. in health care dictation, speech-controlled user interfaces such as military, telephone exchange and disabled user applications, and in information retrieval.

Modern ASR has been based on solving the state sequences of statistically derived hidden Markov models (HMM) with a Gaussian mixture model (GMM) (Juang et al., 1986) structure but that is now changing. Applications of neural networks in ASR were an extensively researched area two decades ago (Bourlard and Wellekens, 1990; Kurimo and Torkkola, 1992; Robinson et al., 1993) and some success was achieved with relatively simple network architectures. The work by Hinton (2002) induced the comeback of neural networks and set the current ASR research trend towards hybrid multilayer-perceptron (MLP) and HMM structures (Hinton et al., 2012).

Solid recognition performance has been possible for years in controlled conditions where the noise levels are low and words are articulated clearly (Lippman, 1997; Siniscalchi et al., 2013). The continuously increasing computational power and workload parallelization has enabled the study of highly complex speech recognition systems trained on thousands of hours of speech data (Jaitly et al., 2012). Nevertheless, the performance of the most complex systems still degrade in the presence of rapidly changing noise (Delcroix et al., 2013) and the human performance remains far above the machine recognition (Barker et al., 2013).

The question why ASR is such a difficult task in noisy environments can be answered by the high number of acoustical parameters affecting the recognition process. While the speaker dependent factors such as accent, health and emotional status do not affect human recognition, the machine tolerance is much lower. However, the highest detrimental factors originate from external sources (Lippman, 1997). Unanticipated and high-level background noise such as a large truck bypassing the speaker may completely overpower the speech signal, and channel distortion such as the restricted signal bandwidth may lose important high frequency information. Therefore, noise robustness is a crucial feature in any practical, especially mobile, ASR system.

To be more precise, the typical errors made by the ASR systems are sound unit, or *phoneme*, classification errors caused by the mismatch between the parameters of the estimated acoustical model and the true data distribution. In practice, this is always the case to some degree since the true distribution remains unknown and can only be estimated. On the other hand, a perfect model for the data distribution does not guarantee optimal performance in every real life scenario since the signal is also influenced by external factors that are unrelated to the content of speech.

Several conceptually separate methods exist for minimizing the effect of external factors. Training the acoustic model on one type of noise, denoted as the *matched model*, yields high recognition accuracy on that particular noise. Training the ASR system on speech data that contains varying noise conditions can produce a more versatile model, denoted as the *multicondition model*, that may also reduce classification errors in cases in which a new background noise is encountered (Pearce and Hirsch, 2000). To reduce the overall effect of mismatch between the true data distribution and its model, discriminative training can be used (Pylkkönen and Kurimo, 2012). In discriminative training, the model parameters are optimized to minimize the recognition error instead of estimating the parameters of individual phoneme distributions optimally (Bahl et al., 1986).

Noise may also be addressed by deriving *features*, the low-dimensional representations of the speech signal, that are inherently robust to noise. While the majority of ASR systems extract features based on Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980), they are relatively prone to noise. Methods taking better advantage of auditory processing have been shown to achieve higher recognition performance

in poor conditions (Kim and Stern, 2012). These methods include, for example, perceptual linear prediction analysis (PLP) (Hermansky, 1990), minimum variance distortionless response (Murthi and Rao, 2000; Dharanipragada and Rao, 2001), and power-normalized cepstral coefficients (PNCC) (Kim and Stern, 2012).

In model compensation methods, the presence of noise is allowed in the recognition process itself by adapting the acoustic model to match the prevailing noisy environment. One example of these methods is the parallel model combination (Gales and Young, 1996) where an acoustical model is also constructed for the interfering noise. The noise model is then used to update the parameters of the clean-speech model through a mismatch function.

Missing data methods (Cooke et al., 1994) take a different approach on handling the aural contamination by assuming that only part of the spectro-temporal, or *time-frequency* (TF), components of the signal are corrupted by noise and the other part preserves the speech information. In other words, the speech-dominant components are considered to hold the reliable information that can be used in recognition. The unreliable noise-dominant components, on the other hand, can be either marginalized or reconstructed with their respective clean speech estimates obtained from a statistical model (Raj et al., 2004) or from a dictionary of speech exemplars (Gemmeke and Cranen, 2008). The reconstructed spectro-temporal representation can now be used to recognize speech. The key to success of the approach lies in estimating spectrographic *masks* that pinpoint the reliable components.

In the thesis, the process of dividing the audio signal into reliable and unreliable TF regions, or into several target and noise sources in general, originates from auditory scene analysis proposed by Bregman (1990) and extended to computational auditory scene analysis (CASA) by Brown and Cooke (1994). Inspired by the mechanisms of the human auditory system, the CASA-based ASR systems try to imitate the way the humans separate the sound sources in monaural and binaural audio streams. In CASA-based systems, the sound source separation can be realized, for example, by computing the mask that weights each TF component of the signal representation so that the target is emphasized and the undesired sources are damped (Lyon, 1984). From the ASR point of view, the missing data approach is closely related to the CASA source separation.

The speech segregation can be simplified to a binary classification problem, as done e.g. in (Han and Wang, 2011). Such procedure can be carried out by extracting acoustic cues that are important for the auditory organization of speech such as pitch-based features and on-/offset (Wang et al., 2012), and by bypassing the CASA stages of segmentation and grouping of TF units. The concept of an oracle mask, the highest recognition accuracy providing mask, can be seen as the goal of the feature-based mask estimation methods proposed in the thesis. The CASA approach has inspired several alternative aspects for mask estimation, for instance, Seltzer et al. (2004) proposed a method based on a Bayesian classifier to determine the reliability of spectrographic elements, Harding et al. (2006) utilized a histogram-based binaural mask estimation and in the study of Healy et al. (2013), subband MLP processing is used for mask estimation.

1.2 Background of the thesis

The history of ASR research in Aalto University began as isolated word recognition based on various types of neural networks such as self-organizing maps. Since the adoption of the HMM structure (Kurimo, 1997) to ASR, the research has shifted towards Finnish large vocabulary continuous speech recognition. The basis of the automatic speech recognizer was introduced in 2002 (Hirsimäki et al., 2006) and the following work of e.g. Siivola (2007), Hirsimäki (2009) and Pylkkönen (2013) led to the creation of the Aalto ASR tools (Speech Group of Aalto University, 2013) that are extensively used by the current speech group researchers.

The noise robust ASR has been a growing research area since the work made by Palomäki et al. (2004a,b) in missing data mask estimation and handling reverberant speech. Recent advanced in noise robustness have been made in discriminative training (Pylkkönen and Kurimo, 2012), uncertainty decoding of features (Kallasjoki et al., 2011, 2014) and missing data imputation techniques (Remes et al., 2011; Remes, 2013).

1.3 Contributions of the thesis

The work presented in the thesis is aimed to enhance the ASR performance in varying noise and low signal-to-noise ratio (SNR) conditions by improving the underlying methods and algorithms. All the proposed

methods are experimentally evaluated on speech data containing varying noise types and levels in the automatic speech recognition scheme.

The conventional linear prediction (LP) based spectral analysis is known to be relatively sensitive to channel distortion and additive noise. In the thesis, extended weighted linear prediction (XLP), introduced in Publication II, is applied to the feature extraction of ASR to enhance the noise robustness of spectrum estimation.

In missing data speech recognition, the quality of the spectrographic mask has a significant influence on the overall recognition accuracy. The missing data mask estimation is compared to two conceptually separate noise robust approaches in Publication I. In Publication III, a mask estimation method based on binaural modeling is introduced. The mask estimation methods relying either on local SNR estimates or simple binaural information, utilized in Publication I and Publication III, respectively, provide good results in high SNR environments but if the noise level is increased, their performances are radically decreased.

In the thesis, new methods for mask estimation are proposed; the first method, introduced in Publication IV, takes advantage of classifiers trained on a comprehensive set of acoustic design features to tolerate the severely deteriorated speech signal. In Publication IV, the design features are also analyzed for their power to discriminate the reliability information. The second method, introduced in Publication V, utilizes neural networks, more specifically Gaussian-Bernoulli restricted Boltzmann machines, to learn the set of features automatically. In Publication VI, the automatic feature learning is further improved and used in conjunction with discriminative training.

1.4 Contents of the thesis

The thesis consists of an introduction and a collection of six publications. Chapter 2 describes methods for transforming speech signals into features commonly used in ASR and mask estimation. Chapter 3 describes the concept of the missing data speech recognition, the proposed methods for mask estimation and two reference mask estimation techniques. In Chapter 4, the procedures of the experimental evaluations of the proposed methods are described and a summary of the results presented in the six publications is given. In Chapter 5, the results are discussed and reflected on the related work in the field and Chapter 6 presents the conclusions.

2. Feature extraction

As other pattern recognition tasks, ASR usually requires processing of raw audio waveform into a feature representation for classifying the speech into phonemes, for instance. Typically, the speech stream is split into sequences or short time *frames* of low-dimensional feature vectors that aim to represent only the relevant information on the speech content. Such processing is denoted as *feature extraction* or front-end processing. However, the relevance of information depends on the task — in speaker recognition, knowledge of the speaker's gender and speaking rate are highly valuable, whereas in ASR, the importance lies in preserving the information needed for phonetic classification. Regardless, the feature extraction method must be chosen carefully since after, typically nonlinear, feature extraction none of the discarded information can be recovered.

So far, the most common method for extracting features has been based on MFCCs (Davis and Mermelstein, 1980) that try to mimic the principles of auditory processing on a crude level. MFCCs possess two significant properties: the feature components are virtually uncorrelated and the dimensionality of the feature vectors are low.

The MFCC processing is commonly done in frames consisting of 16–20 ms of speech in a way that adjacent frames overlap half a frame. The first step in processing is to pre-emphasize the higher frequencies in order to smooth the decaying energy spectrum of natural speech. Second, short-time spectral analysis is done either with discrete Fourier transformation or with linear prediction (LP) from Hamming windowed frames. Next, the short-time magnitude spectrum is weighted by triangular Mel-scale filters whose center frequencies are equally spaced in Mel-scale. The triangular weighting operates as an approximation to auditory filtering. The energies of the Mel-spectral coefficients are compressed with log-function and the number of compressed coefficients are reduced with

discrete cosine transform. The log-function models the level intensity transfer function of human hearing and serves as mapping from the broad intensity range of audible sounds.

Common post-processing techniques to MFCCs are mean and variance normalizations, and computation of coefficient time derivatives and second-order time derivatives, which capture the temporal information from a longer time context than a single frame. Maximum likelihood linear transformation (Gales, 1999) tries to improve the separability of acoustic classes in the feature space and can be seen as a discriminative method.

However, MFCCs are not particularly robust if a mismatch occurs between the training and run-time environments. Such occasions demand more advanced methods such as perceptual linear prediction (PLP) analysis (Hermansky, 1990) based on the explicit modeling of the main phenomena of peripheral auditory processing. In PLP, the auditory spectrum is estimated on a critical-band spectral resolution followed by pre-emphasis by an equal-loudness curve and nonlinear compression with cubic-root intensity-loudness power law.

A good rule of thumb is that feature extraction methods based on the characteristics of the human auditory system are more likely to provide higher classification performance than heuristic or design techniques crafted for one particular speech related task. The explanation for the rule is the hand in hand evolution of human speech production and speech receiving mechanisms perfectly adjusted for each other.

More recent examples are the power-normalized cepstral coefficients (PNCC), which are another set of physiologically motivated features but with computational efficiency kept in mind, introduced by Kim and Stern (2012). The main differences between MFCCs and PNCCs are gammatone filtering, 50–150 ms medium-time nonlinear processing that suppresses the effects of additive noise and room reverberation, and a power function nonlinearity. The recognition accuracies achieved with PNCC features are significantly better than with MFCCs in the presence of additive noise and especially in reverberant environments (Kim and Stern, 2012).

2.1 Linear predictive models

The FFT based short-time spectral analysis in MFCC computation can be substituted with linear predictive methods. The key concept of LP

in speech processing is to model the output of the speech production mechanism as a linear combination of p past samples

$$\hat{x}_n = \sum_{k=1}^{p} a_k x_{n-k}, \tag{2.1}$$

where x_n denotes the time domain speech sample and a_k the prediction coefficient (Makhoul, 1975). The LP is used to estimate the smoothed spectral envelope that preserves the formant structure of speech. More specifically, LP finds the coefficients a_k defining the filter that characterizes the parameters of the vocal tract. The energies of the prediction errors, defined by $e_n = x_n - \hat{x}_n$, are minimized by setting the partial derivatives of $E_{\text{LP}} = \sum_n e_n^2$ with respect to each coefficient a_k to zero. This results to normal equations

$$\sum_{k=1}^{p} a_k \sum_n x_{n-k} x_{n-i} = \sum_n x_n x_{n-i}, \qquad 1 \leq i \leq p. \tag{2.2}$$

In the thesis, an all-pole model is only considered as the normal equations derived from it are straight-forward to solve by autocorrelation method (Makhoul, 1975) that produces a stable model. However, neither FFT- nor LP-based spectral analysis is designed to model the magnitude spectrum of a heavily deteriorated audio signal.

Several improvements to LP and PLP analyses have been proposed, for example, the relative spectral (RASTA) (Hermansky and Morgan, 1994) approach makes the PLP technique robuster to linear spectral distortions, and an adaptively weighted version of the LP cepstrum is introduced to speaker identification by Assaleh and Mammone (1994).

The spectrum estimation itself can be made more tolerant to noise by adding temporal weighting W_n in the linear predictive analysis; the regions of speech that are relatively less corrupted by noise are emphasized. Weighted linear prediction (WLP), introduced by Ma et al. (1993), can be seen as a generalization of LP. The energy of the prediction error $E_{\text{WLP}} = \sum_n e_n^2 W_n$ is minimized by solving the normal equations

$$\sum_{k=1}^{p} a_k \sum_n W_n x_{n-k} x_{n-i} = \sum_n W_n x_n x_{n-i}, \qquad 1 \leq i \leq p. \tag{2.3}$$

The weighting function can be chosen as the short-time energy (STE) of the adjacent samples $W_n = \sum_{i=1}^{M} x_{n-i}^2$, where M is typically selected close to the value of p (Kallasjoki et al., 2009). WLP used in conjunction with STE weighting in the MFCC feature extraction has been shown to improve the noise robustness of continuous speech recognition compared

to the standard FFT-based spectral approximation in MFCC feature extraction (Pohjalainen et al., 2009).

Extended weighted linear prediction (XLP) further generalizes WLP; the weighting is separately applied to each lag in the prediction of each sample, which arguably offers more flexibility for modeling the data (Pohjalainen et al., 2010). The energy of the XLP prediction error is defined by

$$E_{\mathrm{XLP}} = \sum_n \left(x_n W_{n,0} - \sum_{k=1}^{p} a_k x_{n-k} W_{n,k} \right)^2, \tag{2.4}$$

which is minimized by solving the normal equations

$$\sum_{k=1}^{p} a_k \sum_n W_{n,k} x_{n-k} W_{n,i} x_{n-i} = \sum_n W_{n,0} x_n W_{n,i} x_{n-i}, \qquad 1 \leq i \leq p. \tag{2.5}$$

In the thesis, the weighting function of XLP is defined as a recursive equation

$$W_{n,i} = \frac{m-1}{m} W_{n-1,i} + \frac{1}{m} \big(|x_n| + |x_{n-i}| \big), \tag{2.6}$$

where m denotes the effective length of the moving average memory and $W_{n,i}$ is assumed zero for all i's before the beginning of the prediction frame. This weighting assumes, similarly to the STE weighting, that high amplitude samples are relatively less corrupted than low amplitude samples. Typically the length of the average memory m is set close to the value of p. In the thesis, $p = M = m = 20$ is chosen. However, the WLP and XLP methods do not guarantee a stable all-pole filter.

An application of XLP to large vocabulary continuous speech recognition has been studied in Publication II, in which XLP is evaluated against FFT, traditional LP, and WLP by replacing the short-time Fourier transform of MFCCs with the variations of linear predictive models.

2.2 Design features

In the thesis, a design feature denotes a measure of one particular attribute of the speech or the environment. The main focus of each feature can be categorized into robustness to reverberation, robustness to additive noise, and sound source localization (or target detection). Some of the features have been developed from psychoacoustical perspective, for example, interaural time difference (ITD) and interaural level difference (ILD) measure the differences in arrival time and intensity of a sound signal between two ears for the azimuth and elevation, respectively (Blauert, 1996). This provides cues to the relative direction of the source. More

specifically, the localization of the azimuth is based on a cross-correlation type of processing. Furthermore, the design features can be categorized whether they are based on monaural or binaural signals.

The design features are not directly used in phoneme classification but in classifying the reliable and unreliable TF units in missing data mask estimation. Such features include, amongst others, pitch and pitch-based features (Wang et al., 2012), amplitude modulation spectrogram (Moritz et al., 2011; Han and Wang, 2011), and on-/offset (Hu and Wang, 2007). The design features could also be used in phoneme classification with caution since they have not been engineered to capture the characteristics of individual phonemes. In the thesis, the features providing directional cues are used for target detection since the speaker position is assumed known.

Next, the design criteria of the features, used for TF unit classification in Publication III and Publication IV, are briefly described. For mathematical descriptions of the features, please see Publication IV, pp. 801–805 and (Ramírez et al., 2006). All the features are computed in 21 Mel-frequency bins and provide a value for each TF unit except for 'mean-to-peak-ratio of temporal envelope', which provides a single scalar for the whole signal. Here, the binaural features are denoted by (B), monaural features by (M), features used in Publication III by (PubIII), and features used in Publication IV by (PubIV).

Features focused on target detection:

- Interaural time difference (B, PubIV): The time difference of the signal measured between two ears provides information of the source azimuth.

- Interaural level difference (B, PubIV): The intensity difference of the signal measured between two ears provides information of the source elevation.

- Harmonic energy (M, PubIII, PubIV) (van Hamme, 2004): The speech signal is divided into harmonic, or voiced, and inharmonic segments. Speech contains voiced segments composed of components harmonically related to speaker's pitch. The harmonic part is assumed to be dominated by speech. The drawback of the feature is that during voiceless speech, the outcome may be unsatisfactory.

- Peak ITD (B, PubIV): The ratio between the height of the highest peak and height at zero delay in the cross-correlation function between the ears. Peak ITD should be close to unity for sources at zero degrees azimuth.

- Voice activity detection (M, PubIII) (Ramírez et al., 2006): A binary feature motivated by the integrated bi-spectrum, which is defined as a cross-spectrum between the signal and its square.

Features focused on general noise robustness:

- Inharmonic energy (M, PubIII, PubIV): The residual energy of the spectral decomposition of harmonic energy that will mostly consist of noise in voiced speech segments.

- Noise estimate from long-term inharmonic energy (M, PubIII, PubIV): A noise estimate based on the inharmonic energy.

- Noise gain (M, PubIII, PubIV) (van Hamme, 2004): A rough SNR estimate based on the harmonic energy and noise estimate from long-term inharmonic energy.

- Spectral flatness (M, PubIII, PubIV) (Seltzer et al., 2004): Additive noise flattens the valleys of voiced speech segments which can be defined by measuring the local variance of the subband energy within a neighborhood around each TF unit in the log-Mel spectrogram.

- Subband energy to subband noise floor ratio (M, PubIII, PubIV) (Seltzer et al., 2004): An estimate for the noise floor of stationary noise computed from the distribution of subband energy across all frames of the signal.

- MFCC features (M, PubIII): The 21-dimensional MFCC features computed from the noise contaminated speech that has been converted to a single channel signal.

- Noise estimate from channel difference (B, PubIV): A simple noise and reverberation estimate derived from the subtraction of left- and right-ear signals of stereophonic speech. The target speaker's distance from

the microphones is assumed equal and stationary, which removes the direct sound component. Thus, the residual signal contains only noise and reverberation.

Features focused on robustness to reverberation:

- Modulation filtered spectrogram (M, PubIV) (Kingsbury et al., 1998; Hall et al., 2002; Palomäki et al., 2004b): The temporal modulation of speech signals depend on the syllabic rate. The modulation depth is decreased in the presence of reverberation and the goal of the filtering is to emphasize the syllabic modulations and to locate the reverberation-free syllable onsets. Here, the spectrograms are filtered against the time trajectory to amplify the modulation.

- Gradient of temporal envelope (M, PubIV) (Watkins and Makin, 2007): Reverberant regions of the temporal envelope of the speech are frequently characterized by descending tails. Detecting these tails indicates which parts of the speech are reverberated.

- Mean-to-peak-ratio of temporal envelope (M, PubIV): While the reverberation has its highest impact on the valleys of the temporal envelope, the peaks can be assumed to remain unaffected. Computing the ratio between mean and peak values of the temporal envelope over the whole speech signal, the overall "blurredness" of the signal can be measured.

- Interaural coherence (B, PubIV) (Faller and Merimaa, 2004): The direct sound received from a high intensity source is coherent at two ears but diffuse in reverberant sound field. Interaural coherence can be seen as a measurement of the reliability of ITD and ILD cues, and it can also be categorized into general noise robustness.

2.3 Automatically learned features

An alternative to basing the mask estimation on design features is to produce a set of features automatically. In the thesis, the automatic feature learning denotes the unsupervised training of acoustical patterns with neural networks. The automatic features are argued to enhance the mask estimation by capturing information that the design features are

not able to do. While the number of automatic features is higher than the design features, their individual expression power can be considered lower, which can be compensated by the freely adjustable parameter — the number of features.

A neural network based machine learning model capable of extracting discrete classes out of continuous valued input features was introduced in (Bertelsen, 1994) and recently, neural networks retrieving bottleneck features for a GMM/HMM acoustic model have been deployed (Plahl et al., 2010; Yu and Seltzer, 2011; Gehring et al., 2013).

In Publication V and Publication VI, Gaussian-Bernoulli restricted Boltzmann machines (GRBM) were implemented for automatic learning of features. GRBM is a two layer neural network that models the probability density of real-valued data using binary latent variables. The first layer of visible units, that correspond to the components of data vectors, are assumed Gaussian distributed and the second layer consists of binary hidden units. There are no lateral connections in GRBM i.e. each unit of one layer is only connected to all the units in the other layer. It has been shown e.g. in (Krizhevsky, 2009) and (Jaitly and Hinton, 2011) that the latent variables of a learnt GRBM can be used as meaningful unsupervised features.

A GRBM joint configuration $(\mathbf{v}, \mathbf{h})$ of the visible units v_i and hidden units h_j has an energy

$$E(\mathbf{v}, \mathbf{h} \mid \mathbf{\Theta}) = \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n_v} \sum_{j=1}^{n_h} w_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{n_h} c_j h_j, \tag{2.7}$$

where n_v and n_h are the numbers of visible and hidden units, and the parameter $\mathbf{\Theta}$ includes weights w_{ij} connecting the visible and hidden units, σ_i is the standard deviation of v_i, and biases b_i and c_j for each unit (Cho et al., 2011a).

A probability to every possible pair of a visible and a hidden vector is assigned via energy function

$$p(\mathbf{v}, \mathbf{h} \mid \mathbf{\Theta}) = \frac{1}{Z(\mathbf{\Theta})} \exp\big(-E(\mathbf{v}, \mathbf{h} \mid \mathbf{\Theta})\big), \tag{2.8}$$

where $Z(\mathbf{\Theta})$ is a normalizing constant, given by the sum over all possible pairs of visible and hidden vectors

$$Z(\mathbf{\Theta}) = \sum_{\mathbf{v}, \mathbf{h}} \exp\big(-E(\mathbf{v}, \mathbf{h} \mid \mathbf{\Theta})\big). \tag{2.9}$$

An efficient learning procedure called contrastive divergence for GRBMs was proposed by Hinton (2002). First, the states of the visible units are

set to a randomly selected training vector. Second, the binary states of all the hidden units are computed in parallel using Eq. 2.10

$$p(h_j|\mathbf{v}) = \text{sigmoid}\left(c_j + \sum_i w_{ij}\frac{v_i}{\sigma_i}\right), \tag{2.10}$$

where sigmoid denotes function $1/(1+\exp(-x))$. After the binary states have been chosen for the hidden units, a "reconstruction" is produced by setting each v_i to 1 with a probability computed by Eq. 2.11

$$p(v_i|\mathbf{h}) = \mathcal{N}\left(b_i + \sigma_i \sum_j h_j w_{ij}, \quad \sigma_i^2\right), \tag{2.11}$$

where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian. Finally, the states of the hidden units are updated again. The changes in the weights are then given by

$$\Delta w_{ij} = \eta(< v_i h_j >_{\text{data}} - < v_i h_j >_{\text{recon}}), \tag{2.12}$$

where the angle brackets denote the expectations under the distribution specified by the respective subscript. The learning procedure for the biases is essentially the same but the states of individual units are used instead of pairwise products.

Although the number of tunable parameters in GRBM training is quite large, the task can be simplified by normalizing the data components to have zero mean and unit variance so that the standard deviations are set to 1 when computing $p(v_i|\boldsymbol{h})$ in Eq. 2.11.

In the thesis, the weight parameters w_{ij} are updated with an enhanced gradient and an adaptive learning rate η proposed in (Cho et al., 2011b). The enhanced gradient was shown to outperform the conventional gradient update direction in terms of learning speed and invariance to data representation. Additionally, the binary units were replaced by noisy rectified linear units, which were shown in (Nair and Hinton, 2010) to learn more discriminative filters. Thus, the approximate mean activation of the hidden units (the corresponding Eq. 2.12) are given by

$$h_j = \max(0, r_j), \tag{2.13}$$

where $r_j = \sum_i w_{ij} v_i/\sigma_i^2 + c_j$ is the input to the jth hidden unit.

For generating the automatic features, cross-correlation vectors derived from bandpass filtered stereophonic speech signals were used as input to GRBMs according to Figure 2.1. This processing is thoroughly described in Publication V and briefly in here. First, the left- $x_l(n)$ and right-ear $x_r(n)$ time domain signals are filtered into 21 bandpass signals $X_l(n,d)$ and $X_r(n,d)$, where n denotes the sample index and d the frequency bin.

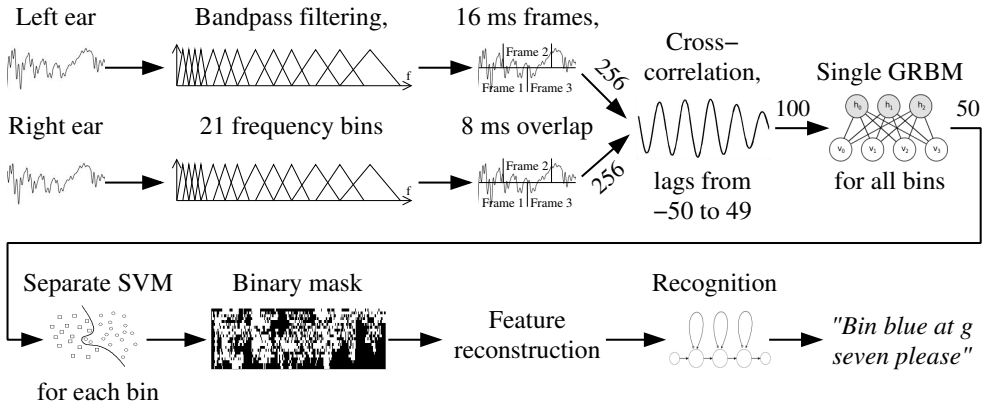

Figure 2.1. Overview of the ASR system, used in Publication V, based on automatic feature mask estimation and the signal processing pathway from the time domain into GRBM input. The numbers next to the arrows denote the number of components/TF unit.

The center frequencies of the bandpass filters conform to the centers of the triangular Mel-scale filters used in the MFCC computation. The cross-correlations between the two ears of frames of bandpass filtered signals are computed within lags ranging from -50 to 40 samples. In GRBM training, the starting points of the 256 sample long frames are chosen by random, whereas in evaluation, the starting points match the framing used in the MFCC computation.

3. Missing data methods

Missing data methods consider noise robust speech recognition as a problem of recognizing speech from incomplete spectrograms (Cooke et al., 2001). Missing data methods assume that the spectral features representing noisy speech can be divided into speech-dominant and noise-dominant regions since the energies of both the speech and noise sources are additive in all TF regions. In practice, if the noise level exceeds the speech level in local TF regions, subtracting the noise is usually not possible, thus the underlying speech information in those regions can be considered missing. Labels dividing the observations are referred to as a spectrographic mask, or simply a mask, as denoted in the thesis.

In the thesis, the TF components of spectral features are denoted as $Y(\tau, d)$, where τ and d represent the time frame and frequency bin, respectively. The speech-dominant and reliable spectral features are represented as $Y_r(\tau, d) \approx S(\tau, d)$, where $S(\tau, d)$ are the features of the underlying clean speech feature that would have been observed without the interfering noise. For the noise-dominant TF components, it holds that $Y_u(\tau, d) \geq S(\tau, d)$ which states that the unreliable feature observations provide an upper bound to the corresponding clean speech feature estimates denoted by $\hat{S}$.

The ASR system can either disregard the unreliable data or *impute*, i.e. reconstruct, the missing data by estimating their corresponding clean speech values $\hat{S}(\tau, d)$. Two imputation methods, cluster-based imputation (CI) and sparse imputation (SI), are briefly described in Section 3.3. CI has been applied in Publication I, Publication III, Publication IV, Publication V and Publication VI, and SI in Publication III and Publication IV. Both imputation methods estimate the clean features using a clean speech model and the reliable and unreliable information.

In the thesis, the speech and noise separation is simplified to a binary classification, or mask estimation, problem by extracting either design or automatically learned features, described in Sections 2.2 and 2.3, respectively. The mask estimation and the related processes are presented in the chapter as follows: mask estimation and the concept of oracle masks in Section 3.1, classifiers in Section 3.2, imputation methods in Section 3.3, and two reference mask estimation methods in Section 3.4.

3.1 Mask estimation and oracle masks

The performance of missing data methods stem from the quality of the estimated masks measured by the ratio of correctly estimated reliable TF units, for instance. A very high ASR performance can be achieved by the missing data techniques if *oracle mask*, or the conceptually similar ideal binary mask, information is used, which has been shown e.g. in (Seltzer et al., 2004), Publication III and Table 4.4 in Chapter 4. Oracle masks can be seen as the goal of mask estimation since they provide the upper limit to the recognition performance in noisy speech. Table 4.4 contains recognition results exclusive for the thesis obtained with oracle mask information. Figure 3.1 presents the Mel-scale spectrograms of clean and noisy speech signals, and the corresponding estimated and oracle masks.

Construction of the oracle masks exploits the fact that two speech signals, noisy and clean, contain the same speech information. This allows to exactly calculate the local SNR in each TF unit. Thus, the oracle mask values are defined as

$$Y(\tau,d) = \begin{cases} 1 & \text{if } 10\cdot\Big(\log_{10}\big(\exp S(\tau,d)\big) - \log_{10}\big(\exp N(\tau,d)\big)\Big) > \theta \\ 0 & \text{otherwise} \end{cases}, \quad (3.1)$$

where $S(\tau,d)$ denotes the clean log-Mel-spectral feature component, $N(\tau,d)$ denotes the log-Mel-spectral noise component, and θ denotes a threshold parameter in decibels. Online ASR applications, on the other hand, do not have access to stereo data sets so the masks are estimated on fly. In the thesis, oracle masks are mainly used in classifier training and for experimental simplicity. The requirement of oracle masks in classifier training for unknown background noises can be avoided e.g. by employing colored-noise to the training data of the mask classifier (Kim and Stern, 2011).

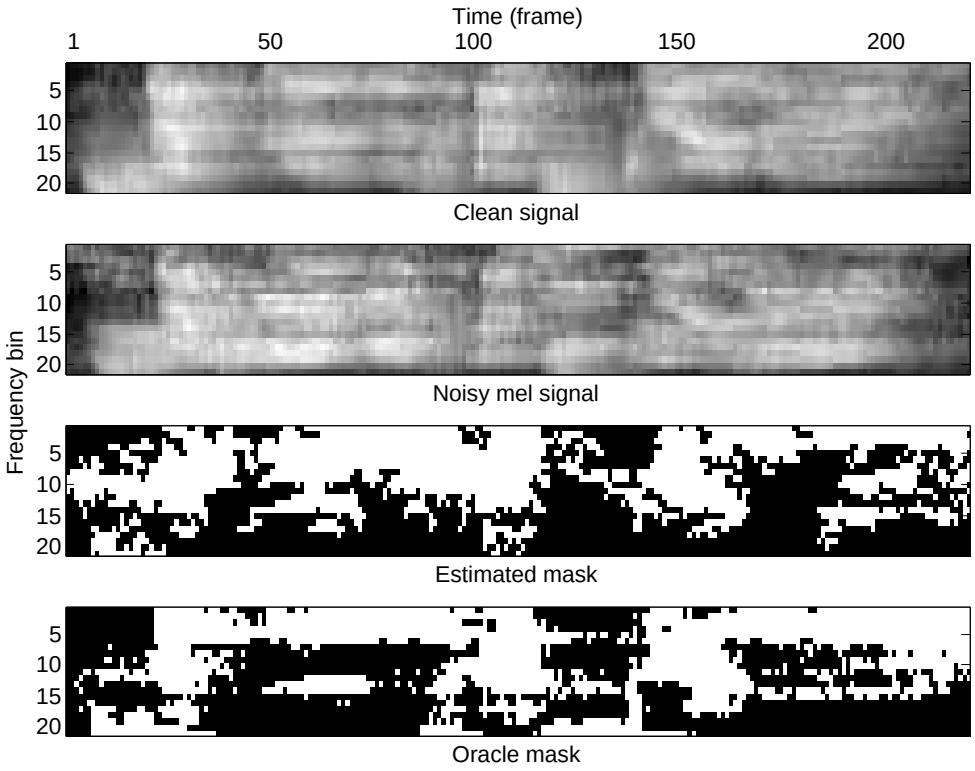

Figure 3.1. The two topmost subfigures present the Mel-scale spectrograms of a clean and a noisy speech signal with SNR of -3 decibels. The two undermost subfigures present the mask estimated from the noisy signal and the respective oracle mask with a reliability threshold θ set to zero decibels. The white/black areas denote high/low energy in the spectrograms and reliable/unreliable TF units in the masks, respectively.

The missing data approach has inspired several research branches for mask estimation. Seltzer et al. (2004) proposed a mask estimation method that uses a Bayesian classification strategy to determine the reliability of each TF unit. Classification is performed using a set of features, some of which are also used in the thesis, representing only the characteristics of speech with no explicit reference to the noise. Separate classifiers were constructed for voiced and unvoiced speech, and for each frequency bin. The two-class GMM model, presented in the thesis, has been motived by this approach. An alternative technique by Ma and Barker (2012) utilizes the CASA source grouping stage by localizing segments in the mask and applying a fragment decoder driven by a clean speech spectral model for improved source segregation.

In the thesis, only binary masks are considered. Rather than making discrete decisions whether a TF unit is either reliable or unreliable, real-valued or fuzzy masks can be used instead by estimating the probability of reliability of the TF unit. For an application of real-valued masks, please see (Barker et al., 2000).

3.2 Classifiers

Subsequent to extracting the features, a classifier makes the decision whether a TF unit is speech- or noise-dominant based on the observed features and a computed model distribution. In the thesis, two alternatives for classification are considered; a two-class GMM model and a support vector machine (SVM) (Cortes and Vapnik, 1995).

Since both the GMM and SVM classifiers process each TF unit independently, the estimated mask can contain isolated reliable components that are unlikely to contain usable information (Cooke, 2006). In missing data reconstruction, even a single isolated reliable component can result in reconstruction errors and notably degrade the system performance. Small scale testing conducted in Publication IV and Publication V has shown that false reliables have a larger impact on the speech recognition accuracy than false unreliables. Overall, the recognition systems usually gain from the removal of groups of reliable features containing less than 5–20 connected reliable elements as experimentally observed in Publication IV.

3.2.1 Two-class GMM model

GMMs have been found to generalize well on speech related data. Most of the 14 design features utilized in Publication IV have smooth probability distributions and therefore they are suitable for GMMs. The two-class GMM model, fully described in Publication IV, is based on using GMMs for the reliable and unreliable classes as a maximum likelihood classifier. First, the likelihoods $P_{d,r}(\boldsymbol{o}(\tau,d))$ and $P_{d,u}(\boldsymbol{o}(\tau,d))$ of N-dimensional feature vector $\boldsymbol{o}$ of τth frame in dth frequency bin are computed. Here, $P_{d,r}(\boldsymbol{o}(\tau,d))$ and $P_{d,u}(\boldsymbol{o}(\tau,d))$ denote the probabilities of the feature vector evaluated on the GMMs that represent the classes of reliable and unreliable features, respectively. The classification of the observation vector, which represents the TF unit, is based on the likelihood scores. However, in practice, the results are usually improved if the likelihood scores are scaled (Seltzer et al., 2004). Hence, by denoting the scale factor by C, the TF units $Y(\tau,d)$ are classified as reliable if $C \cdot P_{d,r}(\boldsymbol{o}(\tau,d)) > P_{d,u}(\boldsymbol{o}(\tau,d))$, and unreliable otherwise. In the thesis, oracle masks, described in Section 3.1, were used to distinguish the classes while training the GMM classifiers.

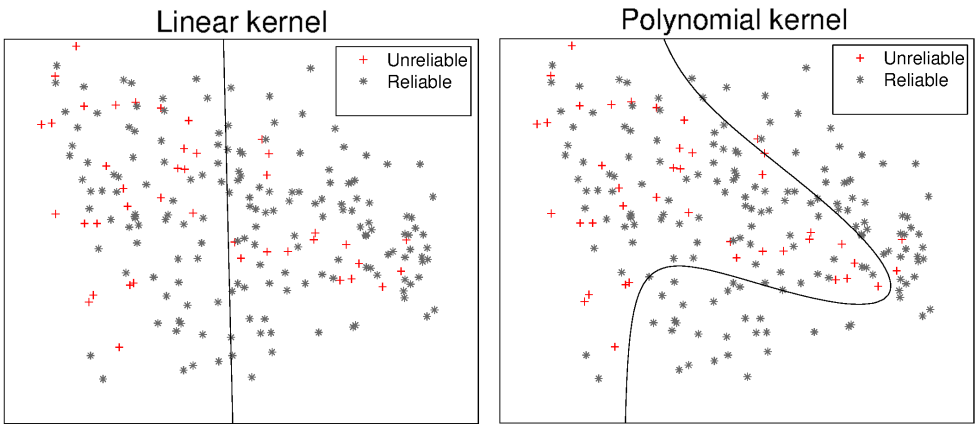

Figure 3.2. The kernel trick; the feature classification becomes more accurate as the number of false reliables is decreased by substituting the SVM linear kernel e.g. with a third order polynomial kernel.

3.2.2 Support vector machine

The SVM training algorithm builds a model that assigns example vectors into one of the two categories. The model represents the examples as points in space in a way that the hyperplane separating the categories has as wide margin as possible. The data points lying closest to the decision border are the support vectors, which define the decision function. New example vectors are mapped into the same space and the decision of belonging to either category is based on which side of the hyperplane they are mapped.

While the original algorithm implemented by SVMs was a linear classifier based on the generalized portrait algorithm introduced in (Vapnik and Lerner, 1963), the current SVMs utilize the kernel trick for a nonlinear classification of patterns that are not linearly separable (Boser et al., 1992). The kernel trick allows the SVM algorithm to fit the maximum-margin hyperplane in a transformed feature space. Thus, the classifier still remains a linear hyperplane in the high-dimensional feature space through the nonlinear transformations. Common transformations include polynomial kernel and radial basis function (RBF). An example of the use of kernel trick is presented in Figure 3.2. SVMs have been proven to be one of the most powerful classifiers in various tasks ranging from cancer diagnostics to credit approval (Meyera et al., 2003). Benefits of SVMs are also that they are not affected by the local minima and do not suffer from the curse of dimensionality.

In Publication III, Publication V and Publication VI, separate SVMs with RBF kernels were trained for each frequency bin d with oracle masks as targets for classifying the TF units into reliable and unreliable. In Publication V and Publication VI, the mean hidden activations of the GRBM given in Eq. 2.13 were taken as input features, and in Publication III, the input features were a subset of the 14 design features, described in Section 2.2. In Publication V, the cross-correlation vectors and the set of 14 design features were also used as SVM inputs.

3.3 Feature reconstruction

Two imputation approaches are utilized in the thesis: the cluster-based imputation (CI) and the sparse imputation (SI) algorithm. Both methods work on vectors formed by the concatenation of T consecutive Mel-spectral feature vectors, in order to include some amount of time context in the reconstruction of a single observation.

3.3.1 Cluster-based imputation

In cluster-based imputation (Raj et al., 2004), a GMM is constructed to model the feature distributions of clean speech. The unreliable values in each feature vector are substituted with estimates that are derived from the clean speech prior and the reliable observations of the vector. More specifically, the missing features are computed as a weighted sum of cluster-conditional bounded maximum a posteriori (MAP) estimates. For more details, please see Publication IV and (Remes et al., 2011).

In Publication I, Publication IV, Publication V and Publication VI, CI was applied in the log-Mel-spectral domain and mapped to the acoustic model feature domain, described in Section 4.2, after the reconstruction.

3.3.2 Sparse imputation

Sparse imputation (Gemmeke and Cranen, 2008) does not rely on the statistical description of clean speech, like the previously described CI, as it takes an alternative approach to reconstruction by modeling speech segments with a linear combination of real speech samples or *exemplars*. In SI, the first step is to construct a dictionary of exemplars from fixed-length clean speech samples that typically span across several frames. Second, using only the reliable speech features, a sparse linear combina-

tion (i.e. many of the coefficients are zero) of exemplars is searched. Then, if a bounding criterion is met, the unreliable features are replaced with the respective clean speech features of the linear combination, otherwise the unreliable features remain as observed. For the precise mathematical formulation of SI, please see (Gemmeke et al., 2011). In the thesis, SI was applied in Publication III and Publication IV.

3.4 Reference mask estimation methods

3.4.1 ITD – ILD pair mask estimation

For evaluating the two proposed mask estimation methods, they are compared in the noisy ASR scheme to the mask estimation approach found to perform the best in Publication III. The reference mask estimation method is based on the approach originally proposed in (Harding et al., 2006). The reference method assumes that the target location, more specifically the azimuth, is known a priori and then it creates a histogram for each frequency bin to model the joint distribution of ITD and ILD for target dominant TF units. From the observed ITD – ILD pair, the histograms are used to determine the probability of the TF unit being speech- or noise-dominant.

Harding et al. (2006) used the estimated probabilities directly as soft masks, whereas in Publication III and Publication IV, binary masks were utilized. A more detailed description of the method can be found in Publication IV, pp. 805–806.

3.4.2 Cross-correlation-based mask estimation

In Publication V, a mask estimation method based on the classification of cross-correlation vectors was used as a reference to the mask estimation based on automatic features. In cross-correlation-based mask estimation, the cross-correlation vectors, used as the input to the GRBM, were directly fed to the frequency bin specific mask classifiers instead.

4. Experiments

The proposed methods presented in the thesis are experimentally evaluated. The evaluation processes are described in the chapter as follows: the data sets and evaluation metrics are described in Section 4.1, the baseline ASR system and the publication specific recognizer configurations are concisely detailed in Section 4.2, and a summary of the essential results obtained in Publications I – VI is presented in Section 4.3.

4.1 Data sets and evaluation methods

The noise robust methods presented in the thesis are experimentally evaluated on ASR tasks which are carried out with three noise containing speech corpora described in this section. Each corpus is divided into a training, a development and an evaluation set. The development sets are used for tuning the ASR system parameters and the final recognition results, with the optimized parameters, are obtained with the evaluation sets.

The Finnish SPEECON corpus (Iskra et al., 2002) is a large vocabulary database gathered for developing speech interfaces for consumer devices and it contains both read and spontaneous speech. SPEECON corpus consists of *real noisy* recordings made in e.g. offices, public places and cars. The close-distance office recordings are considered noise-free with an average SNR of 26 decibels while the public place and car recordings are considered noisy. The public and car noise sets have been recorded simultaneously in three microphone positions; the close-distance recordings have been made with a headset and the mid-distance recordings with a lavaliere microphone. The far-distance position in the car noise set has been recorded by a microphone attached to the rear-view mirror and in the public noise set, the microphone has been placed 0.5–1 meter away from

Table 4.1. Summary of the SPEECON, the first and the second CHiME corpora used in the experimental evaluation of the proposed noise robust methods. The main difference between the two CHiME corpora is the small movement of the speaker in the second CHiME corpus.

Corpus		Lang.	Vocab. size	Mono/ stereo	Train. length (min)	Devel. length (min)	Eval. length (min)	SNRs (dB)
SPEECON		Fin.	large	mono				
Office	(clean)				1170	72	113	26
Public	(noisy)				293	60	94	24, 14, 9
Car	(noisy)				293	29	57	13, 5, 8
1st	CHiME,	Eng.	small	stereo	283	60	60	9, 6, 3,
2nd	CHiME							0, -3, -6

the speaker. The training, development and evaluation sets are speaker exclusive i.e. each speaker appears only in one set. The lengths and SNRs of the training, development and evaluation sets of each speech corpus is summarized in Table 4.1.

Typically, the speech recognition performance is measured in word error rates (WER), especially in English recognition tasks. WER is defined as 100% times the sum of word insertion, deletion, and substitution errors divided by the total number of words in the reference transcription. For languages such as Finnish where long and compound words are typical, it is common to measure the recognition performance in letter error rates (LER) since WERs are considered to over-weight misrecognized word breaks. In the thesis, the speech recognition performance is measured in LERs on SPEECON data.

The first CHiME corpus (Barker et al., 2013) is a small vocabulary English database constructed for the first Computational Hearing in Multisource Environment challenge. The *artificially noisy* CHiME utterances are made by convolving clean speech with a set of binaural room impulse responses simulating reverberation. Each utterance has then been mixed with real domestic noises without rescaling either speech or noise. The utterances contain simple spoken commands such as ”bin blue at c 4 please“. Instead of computing the LER or WER values for each utterance, the evaluation metric in both CHiME corpora is based on recognizing the keywords, which in this case are the alphabet c and the digit 4. The challenge provided training set contains 17 000 short

utterances of reverberated but noise-free speech. The development and evaluation sets both consist of 600 utterances, each mixed in six SNRs. For multicondition training, the reverberated training set was mixed with the challenge provided noise data with SNRs identical to the development and evaluation sets. 34 speakers are shared on all data sets.

The second CHiME challenge consisted of small (Track 1) and medium (Track 2) vocabulary speech recognition tasks (Vincent et al., 2013). In the thesis, only the small vocabulary Track 1 is considered. The differences between the first CHiME and the Track 1 of the second CHiME corpora are the simulated speaker movement within a square zone of +/- 10 cm around the position of two meter distance in front of the dummy head and that the clean, reverberated and noisy training sets were already provided by the challenge.

The obtained results typically contain an error margin due to the limited size of the evaluation material. Statistical analysis of the results must be conducted in order to resolve whether there is a statistically significant difference between the results of two methods. The analyses conducted within Publications I – VI have been carried out by the Wilcoxon signed-rank test (Wilcoxon, 1945) with a 95% confidence level for each system pair. In the thesis, however, a simplification is made by stating that any difference between the average result of each system pair within a result table is significant.

4.2 Recognition systems

The speech recognition system used in the thesis is a large vocabulary continuous speech recognizer originally developed in the Aalto University department of Information and Computer Science and currently maintained in the Aalto University department of Signal Processing and Acoustics. The acoustic modeling toolkit and the decoder of the recognition system are publicly available as Aalto ASR tools (Speech Group of Aalto University, 2013). The recognizer is based on left-right HMM structure with a GMM state likelihood modeling.

In the thesis, the acoustic models are speaker independent state-tied triphones constructed with a decision-tree method. The state modeling is limited to at most 100 Gaussians and the state durations are modeled with gamma probability functions (Pylkkönen and Kurimo, 2004). The speech signal is represented with a power and 12 MFCC features

computed in 16 ms frames that overlap half a frame. The features are concatenated with their first and second order differentials, as described in Section 2, totaling 39 features/frame. Cepstral mean subtraction (CMS) is applied prior to scaling and mapping with maximum likelihood linear transform (Gales, 1999) optimized in training. Finally, the covariance matrices of the Gaussians are diagonalized.

On large vocabulary tasks conducted in Publication I and Publication II, the recognition system utilizes a statistical morph-based variable length n-gram language model to represent a vocabulary of an unlimited size (Hirsimäki et al., 2006). The statistical morphs and the language model were learned from a 145 million word corpus containing text from Finnish books and newspapers in an unsupervised manner. The decoder, which utilizes a beam-pruned Viterbi token-pass system (Pylkkönen, 2005), combines the language and acoustic models by scaling the language model log-probability.

On small vocabulary tasks conducted in Publication II, Publication IV, Publication V, and Publication VI, the n-gram model is substituted with a simple no-backoff bigram model with uniform probabilities for all valid bigrams. This restricts the possible output word combinations to conform to the grammars of CHiME corpora.

The missing data imputation, applied in Publication I, Publication II, Publication IV, Publication V, and Publication VI, takes place in the 21-dimensional log-power Mel-frequency spectrum to obtain the estimates of clean speech features. Subsequent to imputation, the reconstructed spectra were transformed to 12 regular MFCC features and a log-energy feature concatenated with their first and second order differentials. In CI and SI, the features were processed in $T = 5$ and $T = 15$ frames, respectively, except for Publication I, in which CI was applied in $T = 1$ frame.

4.2.1 Publication specific ASR settings

In Publication I, the computation of MFCC feature differentials was simplified and CMS was omitted for implementing the data-driven parallel model combination (DPMC) as described in (Gales and Young, 1996). Voice activity detection was used for segregating the noise segments of the signal and for estimating the parameters of the noise model.

In Publication III, observation uncertainties (Arrowood and Clements, 2002) were used in computing the acoustic model likelihoods of the recon-

structed features. The ASR performance can be improved if the accuracy of the feature reconstruction is taken into account by emphasizing the reconstructed features that have a higher probability of being the correct estimate of clean speech.

In Publication IV, additional speaker adaptation tests with unsupervised maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) were conducted. The adaptation data for each speaker is obtained from the first-pass recognition hypotheses of the respective system using all the SNRs of the evaluation set.

In Publication VI, discriminative training with minimum phone frame error criterion (Zheng and Stolcke, 2005) and extended Baum-Welch optimization (Gopalakrishnan et al., 1989) was used for improved acoustic modeling. The approximate bounded MAP estimates of CI were varied to approximate bounded minimum-mean square error estimates, which was found to give a slight improvement to the speech recognition performance. Additional speaker adaptation tests with the unsupervised MLLR were conducted as in Publication IV.

4.3 Results

Selected results from Publications I – VI are gathered into this section. First, linear prediction based methods are evaluated as the spectral analyzers of MFCC for inherently robust features. Second, three conceptually separate noise robust approaches are evaluated. Finally, methods for improving the mask estimation are evaluated in the missing data ASR framework.

4.3.1 Spectral analyses

The LER results obtained with the noisy SPEECON data sets by substituting the baseline fast Fourier transform (FFT) short-time spectral analysis of MFCC with conventional linear prediction (LP), weighted linear prediction (WLP), and extended weighted linear prediction (XLP) are presented in Table 4.2.

On average, all LP-based methods are able to improve the noise robustness of the short-time spectral analysis of MFCC feature extraction although the improvements are relatively small. The baseline FFT spectral analysis does, however, remain unbeaten on the close recordings

Table 4.2. Letter error rates for the noisy SPEECON public place and car evaluation sets. The averages are computed over both noise sets with 0.62 weight for the public place noise and 0.38 weight for the car noise. The results are taken from Publication II.

	Public place			Car			
	Close	Mid	Far	Close	Mid	Far	Avg.
FFT	**3.3**	23.9	40.8	**4.0**	**29.8**	66.6	26.8
LP	4.6	19.6	34.0	5.9	36.8	**60.0**	25.0
WLP	4.6	19.9	34.3	6.2	32.7	60.1	24.7
XLP	4.5	**18.8**	**32.5**	5.9	33.4	62.7	**24.5**

of both noise sets. The lowest average LER is achieved by XLP-based spectral analysis although its performance on car noise leaves room for improvement. The relative average LER improvement of XLP over the FFT is 8.6%.

4.3.2 Noise robust methods

Table 4.3 presents the letter error rates obtained with three noise robust approaches evaluated on the noisy SPEECON data sets. In the upper part of Table 4.3, the baseline recognizer (BL-CMS), data-driven parallel model combination (DPMC), multicondition training (MC-CMS) and missing data mask estimation (MD-CMS) based methods are compared. Here, "-CMS" denotes the lack of CMS.

For comparison, results from other sources are also gathered in Table 4.3. In the lower part, the BL-CMS, MC-CMS and MD-CMS systems from the upper part of the table have been re-evaluated without the DPMC related simplifications. The corresponding systems are denoted by BL, MC and MD, respectively. The BL and MD results are taken from (Remes, 2013) and MC from (Kallasjoki et al., 2014). The baseline recognizer BL and the results obtained with it in (Remes, 2013) are essentially the same as in (Kallasjoki et al., 2014) with the exception of how the power feature is computed. The baseline system results from (Kallasjoki et al., 2014) are not shown here to avoid confusion.

Even though it is demonstrated that the MD approach is a noise robust method, it still is quite far from the performance of multicondition training. Focusing on improving the mask estimation will arguably lead to significant gains in recognition accuracy. The channel distortion

Table 4.3. Letter error rates for the noisy SPEECON public place and car evaluation sets. The averages are computed over both noise sets with 0.62 weight for the public place noise and 0.38 weight for the car noise. The results in the upper part are taken from Publication I, BL, MD results in the lower part from (Remes, 2013), and MC in the lower part from (Kallasjoki et al., 2014).

	Public place			Car			
	Close	Mid	Far	Close	Mid	Far	Avg.
BL-CMS	5.7	38.4	54.6	6.7	64.3	87.6	40.4
MD-CMS	4.5	24.9	38.0	**4.9**	32.6	75.7	28.2
DPMC	**4.3**	15.7	28.3	5.2	29.2	79.3	24.4
MC-CMS	7.5	**9.4**	**17.7**	5.9	**13.0**	**42.5**	**15.0**
BL	3.4	22.2	38.3	4.2	33.7	67.3	26.5
MD	3.6	14.3	23.1	3.9	19.9	39.6	16.5
MC	3.6	6.5	12.1	4.2	6.7	18.9	8.4

suppressing properties of CMS have a major impact on the recognition accuracies since by disabling it, the average LERs are almost doubled which renders DPMC, in its current form, an impractical method. Here, the MD approach yields 30.1% and 37.7%, and MC approach 62.9% and 68.3% relative LER improvements over the respective baselines.

The results of the FFT system, presented in Table 4.2, and the BL system, presented in Table 4.3, are practically comparable since the systems are identical except for a slightly inferior language model in the FFT system.

4.3.3 Mask estimation methods

Table 4.4 presents the keyword accuracy rates for the first CHiME corpus evaluation set. The baseline system trained on reverberant data is denoted by BL and the multicondition trained system by MC. The reference mask estimation method based on the ITD – ILD pairs, described in Section 3.4.1, is denoted by IIME. The mask estimation systems based on the 14-component design feature set united with GMM classifiers and with either sparse imputation or cluster-based imputation are denoted by 14C+GMM+SI and 14C+GMM+CI, respectively. The 14-component mask estimation with SVM classifiers and CI is denoted by 14C+SVM+CI. The mask estimation method based on direct SVM classification of cross-correlation representation, described in Section 3.4.2, and mask estimation based on automatic features learned by a single GRBM, described in

Table 4.4. Keyword accuracy rates for the first CHiME corpus evaluation set. The average keyword accuracies are computed over all SNRs without weighting. The results in the upper part are taken from Publication IV, in the lower part from Publication V, and the human results from (Barker et al., 2013).

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
BL	86.3	78.3	68.5	53.9	44.3	41.9	62.2
14C+GMM+SI	84.3	78.3	67.3	57.1	44.2	42.7	62.3
IIME+GMM+CI	88.5	83.2	73.5	63.6	54.9	48.6	68.7
14C+GMM+CI	90.3	84.3	76.9	68.2	58.2	56.3	72.3
MC	88.4	84.3	78.8	71.3	61.3	53.9	73.0
iMC	89.0	**86.3**	**82.4**	**74.3**	**65.2**	**61.8**	**76.5**
14C+SVM+CI	**91.0**	85.3	79.4	68.8	56.2	53.7	72.4
XCOR+SVM+CI	90.5	86.1	80.0	69.2	57.4	55.8	73.1
AFME+SVM+CI	90.7	85.8	81.0	69.8	61.4	58.9	74.6
Human	98.8	96.8	95.3	93.8	93.0	90.3	94.7

Section 2.3, are denoted by XCOR+SVM+CI and AFME+SVM+CI, respectively. iMC denotes a multicondition system trained on 14C+GMM+CI system imputed data. The human keyword accuracies are taken from (Barker et al., 2013).

The 14C+GMM+SI system barely outperforms the baseline system which suggests that either the SI does not succeed in reconstructing the missing data in the estimated masks or there is an issue with the implementation of SI. Compared to BL, the system based on the IIME reference mask estimation improves the average keyword accuracy by 10.5%. The 14C+GMM+CI system has a noticeable margin to the IIME in average accuracy and its performance is close to the MC system. Substituting the GMM classifiers to SVMs results in nearly identical average performance. The XCOR+SVM+CI reference system, on the other hand, performs even better than the respective 14C+SVM+CI system and the MC system. From all the mask estimation methods, the highest average accuracy is obtained with the AFME+SVM+CI system, proposed in Publication V. From all the systems, the highest average accuracy is obtained with iMC which indicates that is beneficial to train with imputed data in order to the recognizer to learn or adapt to the errors made in reconstruction.

Table 4.5 presents the keyword accuracy rates obtained with the second CHiME corpus evaluation set. The baseline system trained on reverberant data is denoted by BL and a multicondition trained system trained on

Table 4.5. Keyword accuracy rates for the Track 1 of the second CHiME corpus evaluation set. The average keyword accuracies are computed over all SNRs without weighting. The results are taken from Publication VI except for the human results which are taken from (Barker et al., 2013). The development set oracle mask results are exclusive for the thesis.

	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	Avg.
BL	87.3	80.3	70.2	57.3	45.6	42.0	63.8
MC	86.3	83.3	78.9	71.8	64.1	54.3	73.1
dMC	88.6	86.8	80.8	74.6	66.3	57.9	75.8
dMC+AFSC	88.0	85.8	82.9	77.4	68.8	63.5	77.7
dMC+AFSC+SA	**89.8**	**87.4**	**84.2**	**77.8**	**71.3**	**65.6**	**79.4**
Oracle+CI (devel)	93.3	92.7	91.7	90.8	87.3	86.4	90.4
Human	98.8	96.8	95.3	93.8	93.0	90.3	94.7

noisy data by MC. dMC denotes a discriminatively trained multicondition system. AFSC denotes mask estimation based on automatic features learned by 21 GRBMs and 21 SVM classifiers united with CI. MLLR-based speaker adaptation is denoted by SA and the adaptation data is obtained from the first-pass recognition hypotheses of the dMC+AFSC system. The human results are taken from the first CHiME corpus (Barker et al., 2013) and assumed valid also for the second CHiME corpus. The development set results for CI imputed oracle masks are also shown.

The oracle mask information in missing feature reconstruction provides superior recognition accuracy in all SNR cases compared to the real ASR systems. The oracle results even come relatively close to the human performance, especially in 0 dB case. From the real ASR systems, the best performing dMC+AFSC+SA system improves the keyword accuracies 2.9–56.2% compared to the baseline.

The results in Tables 4.4 and 4.5 are not fully comparable due to the changes made in the corpus during the second CHiME challenge. The BL systems in both CHiME corpora are functionally identical but with the second CHiME data, the BL system gives slightly higher accuracies in all SNRs.

5. Discussion

The methods presented in the thesis have been shown to improve the speech recognition accuracy in noisy environments. The improvements, however, are not always significant and the reasons for that are not always self-evident. The following general discussion provides insights and explanations for the results, which are also reflected on the related work in the field.

As demonstrated in Publication I and Table 4.3, the missing data approach is a robust method compared to the baseline recognizer but still quite far from the performance of multicondition training. One of the main factors in the missing data approach is the quality of the estimated spectrographic mask. This suggests that the mask estimation based on local SNR estimates, presented in Publication I and Table 4.3, produces excessively inaccurate masks since the mask estimation methods utilizing the variety of features are shown in Table 4.4 to achieve performance close to the multicondition trained model, thus producing more accurate masks. The system utilizing the proposed mask estimation method based on automatic features is the best performing system and produces arguably the highest quality masks.

Further improvement is obtained if the recognizer is trained in multicondition style with imputed data (Table 4.4). In general, the best possible robustness is not achieved with a single method but the combination of methods usually yields a significant leap in noise robustness as observed e.g. in (Tüske et al., 2013; Ma and Barker, 2013). For example, the robustness of MFCC features is increased if the FFT short-time spectral analysis is replaced by XLP but the improvement is not significant enough for practical benefit. However, XLP can be applied in conjunction with mask estimation and feature reconstruction for higher recognition performance. Furthermore, optimizing the effective length

of the moving average memory, as done in (Kallasjoki et al., 2009), may provide additional performance improvement to XLP.

In (Gemmeke et al., 2011), SI was shown to give higher performance than CI in reconstructing the oracle masks and estimated masks in low SNR conditions. In Publication IV and Table 4.4, however, SI was able to outperform CI on oracle masks but on the estimated masks, CI was found to give significantly higher recognition accuracy in all SNR cases. This provides more evidence that SI is relatively more sensitive to mask estimation errors than CI, which was also concluded in (Gemmeke et al., 2011).

The reconstructed features are successfully recognized with a discriminatively trained acoustic model subsequent to multicondition training although the optimum order would be to apply discriminative training to the multicondition acoustic model trained on imputed data (Table 4.5). Despite combining five noise robust methods, the relative keyword recognition accuracy drop on the best performing system is approximately 30% while the human accuracy drops only less than 9% when the SNR is decreased from 9 to -6 decibels. Examining the error distribution of letters and digits reveals that both the human- (Barker et al., 2013) and machine-made errors mostly originate from confusable letters such as m and n. The difference, however, between the machine and human recognition is that the humans performed better even in the worst case at -6 dB than the machine in the best case at 9 dB.

A closer implementation to the CASA framework taking advantage e.g. of fragment decoding for the core of denoising the cepstral features (Ma and Barker, 2013), achieved an average keyword accuracy of 79.7% on the first CHiME corpus, whereas 74.6% accuracy was obtained by the comparable automatic feature based system in the thesis. This indicates that the simplification of mask estimation to a binary classification problem is, perhaps, more harmful than expected since the CASA stage of source grouping seems a highly advantageous operation. With the second CHiME corpus, a system based on the ITD$-$ILD histogram mask estimation, described in Section 3.4.1, was used as a binaural front-end processor achieving an average accuracy of 81.9% (Meutzner et al., 2013). As opposed to the binary-valued histogram masks utilized in the thesis, Meutzner et al. (2013) applied real-valued histograms, which has provided significant advantages over the binary-valued histograms, even the different evaluation corpora considered. Real-valued front-end

processing, and real-valued masks in general, can be seen to share some of the principles of uncertainty decoding and offering its benefits.

Raw bandpass filtered speech signals were also initially used as input in GRBM and MLP training providing only a small improvement in recognition accuracy over the "do-nothing" baseline. The reason for such a modest gain might be that insufficient amount of training data was used as the GRBMs, and MLPs in general, require relatively large amount of training data in speech recognition related tasks for good acoustic modeling power and reduced overfitting (Hinton et al., 2012). On the contrary, training the GRBMs with the cross-correlation sample vectors, only a relatively small amount of data was necessary for the training algorithm to converge and for improvement in the mask quality. It is also possible that the optimal network architectures were not used.

One result worth noticing (Table 4.4) is that the mask estimation method based on direct classification of the cross-correlation vectors outperformed the method based on 14 design features which include features derived from the cross-correlation representation. The average keyword accuracy of XCOR mask estimation even exceeded the average accuracy of the MC system. One explanation for this could be the dimensionality of the features; the XCOR features contain 100 components/TF unit as opposed to the design feature vectors comprising of 14 components. This suggests that some information on the target location is lost in deriving the binaural features. Another explanation could be the strong a priori knowledge of the fixed speaker location that benefits the XCOR method. This implies that the method based on 14 design features would offer more flexibility if the target were moving, as in the second CHiME corpus. Furthermore, the design feature mask estimation has more potential to also work on single-channel speech signal.

The GMM and SVM classifiers offer similar recognition performance on classifying the set of 14 design features (Table 4.4). However, SVM was favored in the thesis because of the practical issues regarding the computational speed and the fact that the input values can be discrete. In Publication V, the values of oracle mask thresholds were originally optimized for GMMs thus giving the GMM classifiers advantage over SVMs. In Publication III, the results obtained with the SVM classifiers were relatively poor most likely due to the selection of features which captured only a few essential characteristics of the multisource and reverberant noise.

The recent work by Wang and Wang (2012) proposed a method for mask estimation by classifying a combination of amplitude modulated spectrogram, RASTA-PLP and MFCC features with multilayer perceptrons trained as deep belief nets. In their study, the proposed classifier provided higher classification accuracy and HIT−FA than the SVM classifier trained with the same data. HIT−FA, defined as the difference of the ratio of correctly classified speech-dominant TF units and the ratio of TF units misclassified as noise-dominant, has been shown to correlate with speech intelligibility scores (Kim et al., 2009). However, the mask accuracy as a metric does not automatically translate into good ASR recognition accuracy, which was witnessed during the classifier parameter optimization in Publication V. Wang and Wang (2012) did not incorporate any speech recognition results in their work to demonstrate whether their method would ultimately be beneficial to ASR.

6. Conclusions

The thesis proposes methods for noise robust automatic speech recognition by improving the spectrographic mask estimation in the missing data approach and by investigating the robustness of linear prediction based spectral analysis in the computation of MFCC features.

Substituting the fast Fourier transform with extended linear prediction in the short-time spectral analysis of MFCC feature extraction provides a modest gain in noise robustness. XLP is easy to implement since there is no need to modify other parts of the ASR system and further robustness could be obtained in conjunction with other noise robust methods.

The proposed mask estimation methods, on the other hand, are much more effective than XLP and are able to match or surpass the performance of the multicondition trained system. The mask estimation method based on automatically learned features by the GRBM network, especially, presents a significant improvement over the reference mask estimation method based on ITD – ILD histogram pairs. However, both proposed mask estimation methods rely on stereophonic speech signals, which restricts their online applicability. The high robustness also comes with a cost as extra steps are needed to train and tune the classifiers and imputation systems. Nevertheless, the process up to the imputation can be seen as feature enhancement so there is usually no need to make modifications to the ASR system.

The proposed noise robust mask estimation methods are not only applicable to ASR but also, computational aspects taken aside, to hearing aids. From the signal enhancement point of view, the spectrographic masks can be used to attenuate the background noise in noisy environments thus increasing the speech intelligibility for people with hearing loss.

The overall recognition performance of the missing data approach is heavily influenced by the correctness of the estimated spectrographic

mask. Whether the estimation is founded on classifying a set of perceptually motivated features or learning acoustical patterns automatically, the machines still base all their knowledge on tens to a couple of thousand hours of speech training, whereas humans have started training their recognition systems before they were even born. Despite the gap between human and machine performance, the machine is slowly but steadily reaching the human capabilities.

Broadening the study of incorporating neural networks in mask estimation would be a natural step since the research on the potential applications of the refound MLPs is still scarce. One future direction could be to investigate whether MLPs can automatically learn robust features from the autocorrelation representation of a single channel speech data, for instance, and evaluating the features on a monaural large vocabulary corpus such as SPEECON.

Bibliography

Arrowood, J. A., Clements, M. A., September 2002. Using observation uncertainty in HMM decoding. In: Proc. ICSLP. Denver, Colorado, USA, pp. 1561–1564.

Assaleh, K. T., Mammone, R. J., 1994. New LP-derived features for speaker identification. IEEE Trans. SAP 2 (4), 630–638.

Bahl, L., Brown, P., de Souza, P., Mercer, R., April 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proc. ICASSP. Tokyo, Japan, pp. 49–52.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP. Beijing, China, pp. 373–376.

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech & Language 27 (3), 621–633.

Bertelsen, R., 1994. Automatic feature extraction in machine learning. Master's thesis, Brigham Young University, Provo, UT, USA.

Blauert, J., 1996. Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press.

Boser, B., Guyon, I., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. In: COLT '92: Proc. 5th Annu. Workshop on Computational Learning Theory. ACM Press, New York, NY, USA, pp. 144–152.

Bourlard, H., Wellekens, C., 1990. Links between Markov models and multilayer perceptrons. IEEE Trans. PAMI 12 (12), 1167–1178.

Bregman, A. S., 1990. Auditory Scene Analysis. MIT Press.

Brown, G. J., Cooke, M., 1994. Computational auditory scene analysis. Computer Speech & Language 8, 297–336.

Cho, K., Ilin, A., Raiko, T., June 2011a. Improved learning of Gaussian-Bernoulli restricted Boltzmann machines. In: Proc. ICANN. Vol. 6791 of Lecture Notes in Computer Science. Espoo, Finland, pp. 10–17.

Cho, K., Raiko, T., Ilin, A., June 2011b. Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In: Proc. ICML. Bellevue, WA, USA, pp. 105–112.

Cooke, M., 2006. A glimpsing model of speech perception in noise. Journal of the Acoustical Society of America 119 (3), 1562–1573.

Cooke, M., Green, P., Crawford, M., September 1994. Handling missing data in speech recognition. In: Proc. ICSLP. Yokohama, Japan, pp. 1555–1558.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34 (3), 267–285.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. ASSP 28 (4), 357–366.

Delcroix, M., Kubo, Y., Nakatani, T., Nakamura, A., August 2013. Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling. In: Proc. Interspeech. Lyon, France, pp. 2992–2996.

Dharanipragada, S., Rao, B. D., May 2001. MVDR based feature extraction for robust speech recognition. In: Proc. ICASSP. Salt Lake City, UT, USA, pp. 309–312.

Faller, C., Merimaa, J., 2004. Source localization in complex listening situations: selection of binaural cues based on interaural coherence. Journal of the Acoustical Society of America 116 (5), 3075–3089.

Gales, M. J., 1999. Semi-tied covariance matrices for hidden Markov models. IEEE Trans. on SAP 7 (3), 272–281.

Gales, M. J., Young, S., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. SAP 4 (5), 352–359.

Gehring, J., Miao, Y., Waibel, F. M. A., May 2013. Extracting deep bottleneck features using stacked auto-encoders. In: Proc. ICASSP. Vancouver, Canada, pp. 3377–3381.

Gemmeke, J. F., Cranen, B., August 2008. Using sparse representations for missing data imputation in noise robust speech recognition. In: Proc. EUSIPCO. Lausanne, Switzerland.

Gemmeke, J. F., Cranen, B., Remes, U., 2011. Sparse imputation for large vocabulary noise robust ASR. Computer Speech & Language 25 (2), 462–479.

Gopalakrishnan, P., Kanevsky, D., Nadas, A., Nahamoo, D., May 1989. A generalization of the Baum algorithm to rational objective functions. In: Proc. ICASSP. Glasgow, Scotland, pp. 634–634.

Hall, D. A., Johnsrude, I., Haggard, M. P., Palmer, A. R., Akeroyd, M. A., Summerfield, A. Q., 2002. Spectral and temporal processing in human auditory cortex. Cereb Cortex 12 (2), 140–149.

Han, K., Wang, D. L., May 2011. An SVM based classification approach to speech separation. In: Proc. ICASSP. Prague, Czech Republic, pp. 4632–4635.

Harding, S., Barker, J., Brown, G. J., 2006. Mask estimation for missing data speech recognition based on statistics of binaural interaction. IEEE Trans. ASLP 14 (1), 58–67.

Healy, E. W., Yoho, S. E., Wang, Y., Wang, D. L., 2013. An algorithm to improve speech recognition in noise for hearing-impaired listeners. Journal of the Acoustical Society of America 134 (4), 3029–3038.

Hermansky, H., 1990. Perceptual linear predictive PLP analysis of speech. Journal of the Acoustical Society of America 87 (4), 1738–1752.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. SAP 2 (4), 630–638.

Hinton, G., 2002. Training products of experts by minimizing contrastive divergence. Neural Computation 14 (8), 1771–1800.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Sainath, T., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine 29 (6), 82–97.

Hirsimäki, T., 2009. Advances in unlimited-vocabulary speech recognition for morphologically rich languages. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

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

Hu, G., Wang, D. L., 2007. Auditory segmentation based on onset and offset analysis. IEEE Trans. ASLP 15 (2), 396–405.

Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Kiessling, A., May 2002. SPEECON– speech databases for consumer devices: database specification and validation. In: Proc. LREC. Las Palmas, Canary Islands, Spain, pp. 329–333.

Jaitly, N., Hinton, G., May 2011. Learning a better representation of speech soundwaves using restricted Boltzmann machines. In: Proc. ICASSP. Prague, Czech Republic, pp. 5884–5887.

Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V., September 2012. Application of pretrained deep neural networks to large vocabulary speech recognition. In: Proc. Interspeech. Portland, OR, USA, pp. 2578–2581.

Juang, B. H., Levinson, S., Sondhi, M., 1986. Maximum likelihood estimation for multivariate mixture observations of Markov chains. IEEE Trans. Information Theory 32 (2), 307–309.

Kallasjoki, H., Gemmeke, J. F., Palomäki, K. J., 2014. Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition. IEEE Trans. ASLP 22 (2), 368–380.

Kallasjoki, H., Palomäki, K., Magi, C., Alku, P., Kurimo, M., June 2009. Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. In: Proc. SPECOM 2009. St. Petersburg, Russia, pp. 221–225.

Kallasjoki, H., Remes, U., Gemmeke, J. F., Virtanen, T., Palomäki, K., August 2011. Uncertainty measures for improving exemplar-based source separation. In: Proc. Interspeech. Florence, Italy, pp. 469–472.

Kim, C., Stern, R., March 2012. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proc. ICASSP. Kyoto, Japan, pp. 4101–4104.

Kim, G., Lu, Y., Hu, Y., Loizou, P. C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. JASA 126 (3), 1486–1494.

Kim, W., Stern, R. M., 2011. Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise. Speech Communication 53 (1), 1–11.

Kingsbury, B., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 25 (1–3), 117–132.

Krizhevsky, A., 2009. Learning multiple layers of features from tiny images. Tech. rep., Computer Science Department, University of Toronto.

Kurimo, M., 1997. Using self-organizing maps and learning vector quantization for mixture density hidden Markov models. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Kurimo, M., Torkkola, K., October 1992. Application of SOMs and LVQ in training continuous density hidden Markov models. In: Proc. ICSLP. Banff, Canada, pp. 543–546.

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

Lippman, R. P., 1997. Speech recognition by machines and humans. Speech Communication 22 (1), 1–15.

Lyon, R. F., March 1984. A computational model of binaural localization and separation. In: Proc. ICASSP. San Diego, CA, USA, pp. 41–44.

Ma, C., Kamp, Y., Willems, L. F., 1993. Robust signal selection for linear prediction analysis of voiced speech. Speech Communication 12 (2), 69–81.

Ma, N., Barker, J., September 2012. Coupling identification and reconstruction of missing features for noise-robust automatic speech recognition. In: Proc. Interspeech. Portland, OR, USA, pp. 2638–2641.

Ma, N., Barker, J., June 2013. A fragment-decoding plus missing-data imputation ASR system evaluated on the 2nd CHiME challenge. In: Proc. 2nd CHiME Workshop on Machine Listening in Multisource Environments. Vancouver, Canada, pp. 53–58.

Makhoul, J., 1975. Linear prediction: A tutorial review. Proc. of the IEEE 63 (4), 561–580.

Meutzner, H., Schlesinger, A., Zeiler, S., Kolossa, D., June 2013. Binaural processing for enhanced speech recognition robustness in complex listening environments. In: Proc. 2nd CHiME Workshop on Machine Listening in Multisource Environments. Vancouver, Canada, pp. 7–12.

Meyera, D., Leischa, F., Hornik, K., 2003. The support vector machine under test. Neurocomputing 55 (1–2), 169–186.

Moritz, N., Anemuller, J., Kollmeier, B., May 2011. Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments. In: Proc. ICASSP. Prague, Czech Republic, pp. 5492–5495.

Murthi, M. N., Rao, B. D., 2000. All-pole modeling of speech based on the minimum variance distortionless response spectrum. IEEE Trans. SAP 8 (3), 221–239.

Nair, V., Hinton, G., June 2010. Rectified linear units improve restricted Boltzmann machines. In: ICML. Haifa, Israel, pp. 807–814.

Palomäki, K. J., Brown, G. J., Barker, J., 2004a. Techniques for handling convolutional distortion with missing data automatic speech recognition. Speech Communication 43 (1–2), 123–142.

Palomäki, K. J., Brown, G. J., Wang, D. L., 2004b. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Communication 43 (4), 361–378.

Pearce, D., Hirsch, H.-G., October 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP. Beijing, China, pp. 29–32.

Plahl, C., Schlüter, R., Ney, H., September 2010. Hierarchical bottle neck features for LVCSR. In: Proc. Interspeech. Makuhari, Japan, pp. 1197–1200.

Pohjalainen, J., Kallasjoki, H., Palomäki, K. J., Kurimo, M., Alku, P., May 2009. Weighted linear prediction for speech analysis in noisy conditions. In: Proc. Interspeech. Brighton, UK, pp. 1315–1318.

Pohjalainen, J., Saeidi, R., Kinnunen, T., Alku, P., September 2010. Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions. In: Proc. Interspeech. Makuhari, Japan, pp. 1477–1480.

Pylkkönen, J., April 2005. An efficient one-pass decoder for finnish large vocabulary continuous speech recognition. In: Proc. 2nd Baltic Conf. on Human Language Technologies. Tallinn, Estonia, pp. 167–172.

Pylkkönen, J., 2013. Towards efficient and robust automatic speech recognition: decoding techniques and discriminative training. Ph.D. thesis, Aalto University, Espoo, Finland.

Pylkkönen, J., Kurimo, M., October 2004. Duration modeling techniques for continuous speech recognition. In: Proc. Interspeech. Jaju Island, Korea, pp. 385–388.

Pylkkönen, J., Kurimo, M., September 2012. Improving discriminative training for robust acoustic models in large vocabulary continuous speech recognition. In: Proc. Interspeech. Portland, OR, USA, pp. 1211–1214.

Raj, B., Seltzer, M., Stern, R. M., 2004. Reconstruction of missing features for robust speech recognition. Speech Communication 43 (4), 275–296.

Ramírez, J., Górriz, J., Segura, J., Puntonet, C., Rubio, A., 2006. Speech/non-speech discrimination based on contextual information integrated bispectrum LRT. IEEE Signal Processing Letters 13 (8), 497–500.

Remes, U., August 2013. Bounded conditional mean imputation with an approximate posterior. In: Proc. Interspeech. Lyon, France, pp. 3007–3011.

Remes, U., Nankaku, Y., Tokuda, K., August 2011. GMM-based missing feature reconstruction on multi-frame windows. In: Proc. Interspeech. Florence, Italy, pp. 2407–2410.

Robinson, A. J., Almeida, L., Boite, J.-M., Bourlard, H., Fallside, F., Hochberg, M., Kershaw, D., Kohn, P., Konig, Y., Morgan, N., Neto, J. P., Renals, S., Saerens, M., Wooters, C., September 1993. A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The Wernicke project. In: Proc. Eurospeech. Berlin, Germany, pp. 1941–1944.

Seltzer, M., Raj, B., Stern, R. M., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Communication 43 (4), 379–393.

Siivola, V., 2007. Language models for automatic speech recognition: construction and complexity control. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Siniscalchi, S. M., Yu, D., Deng, L., Lee, C.-H., 2013. Exploiting deep neural networks for detection-based speech recognition. Neurocomputing 106 (1), 148–157.

Speech Group of Aalto University, 2013. Aalto ASR tools. Available: `https://github.com/aalto-speech/AaltoASR`, cited December 2013.

Tüske, Z., Schlüter, R., Ney, H., May 2013. Deep hierarchical bottleneck MRASTA features for LVCSR. In: Proc. ICASSP. Vancouver, Canada, pp. 6970–6974.

van Hamme, H., May 2004. Robust speech recognition using cepstral domain missing data techniques and noisy masks. In: Proc. ICASSP. Quebec, Canada, pp. 213–216.

Vapnik, V., Lerner, A., 1963. Pattern recognition using generalized portrait method. Automation and Remote Control 24 (6), 774–780.

Vincent, E., Barker, J., Watanabe, S., Roux, J. L., Nesta, F., Matassoni, M., May 2013. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In: Proc. ICASSP. Vancouver, Canada, pp. 126–130.

Wang, Y., Han, K., Wang, D. L., September 2012. Acoustic features for classification based speech separation. In: Proc. Interspeech. Portland, OR, USA, pp. 1532–1535.

Wang, Y., Wang, D. L., September 2012. Boosting classification based speech separation using temporal dynamics. In: Proc. Interspeech. Portland, OR, USA, pp. 1528–1531.

Watkins, A., Makin, S., 2007. Perceptual compensation for reverberation in speech identification: effect of single-band, multiple-band and wideband noise contexts. Acta Acustica United with Acustica 93 (3), 403–410.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1 (6), 80–83.

Yu, D., Seltzer, M., August 2011. Improved bottleneck features using pretrained deep neural networks. In: Proc. Interspeech. Florence, Italy, pp. 237–240.

Zheng, J., Stolcke, A., September 2005. Improved discriminative training using phone lattices. In: Proc. Interspeech. Lisbon, Portugal, pp. 2125–2128.