KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Non-negative Sparse Representations for Speech Enhancement and Recognition

Deepak Baby

Supervisor:
Prof. dr. ir. Hugo Van hamme

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

November 2016

Non-negative Sparse Representations for Speech Enhancement and Recognition

Deepak BABY

Supervisor:
Prof. dr. ir. Hugo Van hamme

Members of the
Examination Committee:
Prof. dr. ir. Paul Van Houtte, Chair
Prof. dr. ir. Dirk Van Compernolle
Prof. dr. ir. Marc Moonen
Prof. dr. Tuomas Virtanen
(Tampere University of Technology, Finland)
Dr. Jort Florent Gemmeke
(Google, Inc., Mountain View, USA)

Dissertation presented in partial
fulfillment of the requirements
for the degree of Doctor of
Engineering Science (PhD):
Electrical Engineering

November 2016

Uitgegeven in eigen beheer, Deepak Baby, Kasteelpark Arenberg 10, B-3001 Heverlee (Belgium)

Preface

Gratitude is not only the greatest of virtues, but the parent of all the others.

Marcus Tullius Cicer

This thesis describes the main research work I have done during the past four years at KU Leuven. Several people have contributed immensely to making the roller-coaster journey towards my PhD a memorable one. It is with great pleasure that I take this opportunity to thank everyone who have directly or indirectly contributed to my PhD.

First and foremost, I would like to thank my supervisor Prof. Hugo Van hamme for giving me an opportunity to carry out this research work, and for his constant support and guidance throughout this work. I have always admired your enthusiasm, attention to the scientific details, professionalism, and sense of commitment towards your work and your thesis students. I have learnt a lot from you in and out of work and am thankful for all your help and the long discussions we had on tackling various research problems.

Many thanks to Dr. Tuomas Virtanen for giving me an opportunity to visit his lab, the stimulating discussions and valuable comments on the papers we have co-authored. Special thanks to Dr. Jort Gemmeke for his guidance and support during the beginning phase of my PhD and also being a member of my jury. I am also grateful to the assessors of my PhD, Prof. Dirk Van Compernolle and Prof. Marc Moonen and also for being part of the jury. Their remarks made this thesis more readable and informative. Thanks are also due to my collaborators, Laura Seynaeve, Prof. Wim Van Paesschen and Prof. Patrick Dupont at the University Hospital of KU Leuven for giving me an opportunity to contribute towards the area of clinical neuroscience. Also Rudi Vuerinckx at Nuance, Merelbeke, whose help and advice during my visit at his company have been extremely helpful and highly appreciated.

I would also like to thank my office mates Meng, Emre, Vincent and Jeroen for the valuable technical and non-technical discussions and being extremely helpful. Many thanks to past and present members of the speech research group, Joris, Xeuru, Reza, Kseniya, Hasan, Bart, Lyan and Prof. Patrick for making the life at KU Leuven quite a memorable and fun-filled one. I will miss our coffee breaks, ping-pong, alma walks and speech group activities. I also thank our secretaries Annitta and Patricia for their support on administrative matters. I would also like to thank my Indian friends at ESAT, Sreeraj, Neetha, Minta, Dileep, Fradaric, Shruti, Bismi and Archana for our coffee-break talks and your support during my PhD.

My thanks also go to the European Commission for graciously funding the Marie Curie ITN project INSPIRE I have been part of. I would also like to express my gratitude towards all the members of the project and my fellow PhD students for the technical discussions, collaborations, and the fun-filled evenings and fellow's days during our project meetings.

My stay in Leuven was also supported by my Indian friends, especially Sreeraj and Radhika 'bro' who made my otherwise lonely life in Leuven a lively one with their unconditional love and friendship. I will never forget our cooking experiments, technical and non technical discussions, late-night card games and all the fun we had together. I am grateful to the priests at Heilige Geestcollege, especially Fr. Thomas, Fr. Veenus and Fr. Shinto for their support.

I owe a lot to my masters thesis supervisor Prof. Sibi Raj B. Pillai for his constant motivation and support throughout my study at IIT Bombay and igniting my passion towards research. I should also mention my dear friends in India, Johns, Geethu, Appu, Deepa, Aiswarya, Divya and Sreelakshmi for being always there for me during all the personal and professional ups and downs.

This note on gratitude would be incomplete without mentioning my dear family and relatives for supporting me throughout my life and studies. Words fail to express my gratitude towards my parents, and siblings Dinu and Dyna for their love and support. Many hugs to Reena aunty for her motherly care and love. I should also mention Sumam, my affianced, and her family for their love and support during the final stages of my PhD. My heartfelt thanks to all my family members, cousins, teachers, and other friends in Belgium and India who enriched my life with their presence.

To my dear mother, Silvi K. Paul
&
my father, Baby P. E.

Abstract

Speech recordings taken from real-world environments often contain background noises which degrade the speech signal and reduce its intelligibility. Such degradations also adversely affect the performance of digital systems that operate on the recorded speech such as mobile communication, automatic voice assistance, automatic speech recognition (ASR), hearing aids and speaker identification systems. Robustness of such systems to the various acoustic background noises is still a challenging problem despite decades of research and myriad of different approaches. Speech enhancement schemes aim at recovering the original speech signal by suppressing the noises to improve the speech intelligibility and are popularly used as a front-end for most of the applications.

This dissertation concentrates on speech enhancement schemes for single-channel recordings of noisy speech. Traditional single-channel speech enhancement schemes such as Wiener filtering do not work satisfactorily well in the presence of background noises that vary over time. Alternatively, composite models that approximate the noisy speech as a linear combination of long-context atoms that model the spectro-temporal behaviour of speech and noise signals can be effectively used for suppressing such non-stationary noises. The main goals of this thesis are to develop novel composite models for a better speech enhancement quality and to investigate the application of these settings as a front-end for the various state-of-the-art ASR systems. In particular, composite models derived from the family of non-negative matrix factorisation (NMF) algorithms, that have been successfully used for separating individual signals from a noisy mixture, are proposed. This thesis describes a set of new algorithms based on this theory that can be broadly subdivided into three main (overlapping) sections.

First, we propose a coupled dictionary based approach to the family of NMF-based speech enhancement systems. Typical NMF-based systems use the decomposition in lower dimensional spectro-temporal feature representations. Such feature spaces are preferred over the full-resolution frequency domain for their reduced computational complexity and the ability to better generalise to unseen noise cases. But the resulting noise suppression may be sub-optimal because a low-rank approximation is used to map the estimated speech and noise features to the full-resolution frequency domain. The proposed approach provides an efficient way to directly compute the full-resolution frequency estimates of speech and noise using coupled dictionaries: an input dictionary containing atoms from the desired feature space to obtain the decomposition and a coupled output dictionary composed of atoms from the full-resolution frequency domain. We also introduce the perceptually motivated modulation spectrogram features for the NMF-based tasks. The idea of using coupled dictionaries is then extended to define hybrid exemplar-spaces that are obtained by concatenating different spectro-temporal representations for a better speech and noise separation. The proposed systems were evaluated for various choices of input representations and yielded improved speech enhancement performances on the AURORA-2 and AURORA-4 databases. We further show that the proposed approaches result in improved speech recognition accuracies on the AURORA-2, AURORA-4 and the CHiME-3 challenge databases.

Next, we propose a novel approach to address the difficult problem of single-channel speech enhancement under noisy and reverberant environments. Such recordings are comprised of the original speech, its reflections from various surfaces and the background noise. Thus, the speech enhancement schemes for such scenarios should be able to remove these reflections as well, apart from separating background noise from the target speech. The effect of such reflections are mathematically modelled as a convolution of the original speech with a room impulse response (RIR) that typically has a decaying nature over time. We propose a novel approximation of the noisy reverberant speech in the frequency domain and non-negative matrix deconvolution (NMD). In the proposed model, the RIR in the frequency domain is defined such that its decaying structure can also be estimated from the recording itself. The proposed model is evaluated on a synthetic dataset created by convolving the recordings from the TIMIT database with RIRs measured from different rooms and varying speaker-and-microphone locations, and adding background noises taken from the CHiME-1 corpus. Simulation results show that the proposed model results in a better RIR estimate over the existing model and improves various instrumental speech quality measures.

Finally, we present an application of one of the proposed speech enhancement schemes together with an ASR setting in the field of clinical neuroscience for the pre-operative planning on patients with brain tumor. During the pre-operative planning, a neurosurgeon has to decide if the affected brain region is essential for the major functions such as motor movement and language related processes. To identify the functional relevance of a brain region for language related processes, a picture naming task together with magnetic stimulation of the relevant brain region has been used. The magnetic stimulation equipment produces impulsive noises which are also captured by the microphone. Currently, the accuracy and the reaction times of the responses are found manually from the recordings which is prone to substantial intra- and inter-observer variabilities especially in the presence of the noise from the equipment. A novel automatic and objective evaluation routine for the picture naming task using ASR and the proposed speech enhancement schemes is developed and is tested against the manual annotations on responses collected from 8 subjects.

Beknopte samenvatting

Spraak opgenomen in reële omstandigheden bevat vaak achtergrondlawaai dat het spraaksignaal degradeert en de verstaanbaarheid vermindert. Dergelijke degradaties verminderen ook de performantie van digitale systemen die het opgenomen signaal verwerken, zols mobiele communicaties, automatische spraakassitentie, automatische spraakherkenning (ASH), hoorapparaten en sprekeridentificatiesystemen. Ondanks de decennia aan onderzoek en de vele gepubliceerde methoden is de robuustheid van dergelijke systemen tegen verschillende achtergrondgeluiden nog steeds een uitdagend probleem. Spraakverbeteringsalgoritmen proberen het originele spraaksignaal te herstellen door de ruis te onderdrukken. Ze worden dan ook toegepast als voorverwerking in de meeste toepassingen.

Dit proefschrift handelt over spraakverbeteringsalgoritmen voor éénkanaalsopnames van ruizige spraak. Traditionele éénkanaalsspraakverbetering zoals Wienerfiltering werken niet op een bevredigende manier in de aanwezigheid van achtergrondlawaai dat variëert over de tijd. Samengestelde modellen die ruizige spraak modelleren als een lineaire combinatie van lange contextuele atomen die het tijds-frequentiegedrag van spraak en ruis vatten vormen een effectief alternatief voor het onderdrukken van niet-stationaire ruis. De belangrijkste doelen van dit proefschrift zijn om nieuwe samengestelde modellen te ontwikkelen teneinde een betere spraakverbeteringskwaliteit te bieden en om de toepassing ervan als voorverwerking voor verschillende hedendaagse ASH-systemen te onderzoeken. In het bijzonder worden samengestelde modellen voorgesteld gebaseerd op niet-negatieve matrixfactorisatie (NMF), die al eerder met succes werden toegepast voor het scheiden van individuele signalen uit een ruizig mengsel van geluiden. Dit proefschrift beschrijft een aantal nieuwe algoritmen die op deze theorie gebaseerd zijn en die onderverdeeld kunnen worden in drie (overlappende) secties.

Ten eerste wordt in de familie van op NMF gebaseerde spraakverbeteringssysteemen een aanpak voorgesteld gebaseerd op gekoppelde woordenboeken.

Typische op NMF gebaseerde systemen gebruiken een ontbinding in laagdimensionale tijd-frequentievoorstellingen. Deze kenmerken worden verkozen boven voorstellingen die de volledige frequentieresolutie benutten omwille van de lagere complexiteit van de berekeningen en hun vermogen om te veralgemenen naar ongeziene ruissoorten. De resulterende ruisonderdrukking kan echter suboptimaal zijn omdat een benadering van lage rang gebruikt wordt om de geschatte spraak en ruis terug af te beelden op een spectrum van volle resolutie. In de voorgestelde aanpak wordt een spectrum van volle frequentieresolutie van zowel de spraak als de ruis berekend: een ingangswoordenboek bevat atomen in een geschikte ruimte om het ruizige signaal effectief en efficiënt te ontbinden terwijl een gekoppeld uitgangswoordenboek atomen bevat van volle freuentieresolutie voor de signaalreconstructie. Ook introduceren we de perceptueel gemotiveerde modulatiespectrogramkenmerken voor NMF-gebaseerde taken. Het concept van gekoppelde woordenboeken wordt dan verder uitgebreid naar hybride ruimten die bekomen worden door meerdere tijd-frequentievoorstellingen naast elkaar te plaatsen zodat een betere spraak/ruis-scheiding bekomen wordt. De voorgestelde systemen werden geëvalueerd voor verschillende keuzes van de ingangsvoorstelling en leidden tot een betere spraakverbetering op de AURORA-2 en AURORA-4 databanken, evenals tot betere spraakherkenning op de AURORA-2, AURORA-4 en CHiME-3 taken.

Vervolgens stellen we een nieuwe aanpak voor voor het moeilijke probleem van éénkanaals spraakverbetering onder ruizige omstandigheden met galm. Het opgenomen signaal bevat nu de originele spraak, de reflecties hiervan op verschillende oppervlakken en achtergrondlawaai. De spraakverbetering moet in een dergelijk scenario in staat zijn om deze reflecties te onderdrukken én het achtergrondlawaai van de doelspraak te scheiden. Het effect van dergelijke reflecties wordt wiskundig gemodelleerd door middel van een convolutie van de originele spraak met een kamerimpulsresponsie (KIR), die typisch een uitdijende structuur vertoont. We stellen een nieuwe benadering in het freuentiedomein voor van de ruizige spraak met galm, gebruik makend van niet-negatieve matrixdeconvolutie (NMD). De KIR wordt in het frequentiedomein voorgesteld zodat het uitdijend gedrag kan geschat worden uit de opname. Het voorgestelde model wordt geëvalueerd op synthetische data bekomen door opnames uit de TIMIT-databank te convolueren met KIRs die opgemeten werden in verschillende kamers en verschillende locaties van de sprekers en opnamemicrofoons, en het bekomen signaal met galm te verstoren met achtergrondlawaai uit het CHiME-1 corpus. De simulaties tonen aan dat het voorgestelde model resulteert in een betere KIR-schatting dan een bestaand model en bovendien verschillende instrumentele spraakkwaliteitsmetrieken verbetert.

Tenslotte presenteren we een toepassing van één van de voorgestelde spraakverbeteringsalgoritmen samen met spraakherkenningstechnologye, in het domein van de klinische neurowetenschappen voor pre-operatieve planning ten behoeve van patiënten met een hersentumor. Tijdens de pre-operatieve planning moet een neurochirurg nagaan of het aangetaste hersenweefsel essentieel is voor functies als motoriek of taal. Om de functionele relevantie van een hersengebied voor taalgerelateerde processen te bepalen wordt een beeldbenoemingstaak opgezet onder inhibiterende magnetische stimulatie. De magnetische stimulatie is zo sterk dat deze ook opgevangen wordt door de microfoon. Vóór ons werk werd de responstijd van de patiënt manueel gemeten, wat substantiële intra- en interwaarnemervariaties teweeg brengt, zeker in de aanwezigheid van de stoorpulsen. Een nieuw automatisch en objectief evaluatieprotocol voor de beeldbenoemingstaak gebruik makend van ASH en de voorgestelde spraakverbetering werd ontwikkeld en vergeleken met de manuele annotaties voor acht testpersonen.

Abbreviations

ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CD	Cepstral Distance
CNN	Convolutional Neural Network
DNN	Deep Neural Network
ERB	Equivalent Rectangular Bandwidth
fMLLR	feature-space Maximum Likelihood Linear Regression
LPF	Low-Pass Filter
GBFB	Gabor Filter-Bank
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IMCRA	Improved Minima Controlled Averaging
KLD	Kullback-Leibler Divergence
LDA	Linear Discriminant Analysis
LLR	Log-Likelihood Ratio
MLLT	Maximum-Likelihood Linear Transform
MS	Modulation (Envelope) Spectrogram

NN	Neural Network
NMF	Non-negative Matrix Factorisation
NMD	Non-negative Matrix Deconvolution
PER	Phone Error Rate
PESQ	Perceptual Evaluation of Speech Quality
PLP	Perceptual Linear Prediction
PSD	Power Spectral Density
RBM	Restricted Boltzmann Machine
RIR	Room Impulse Response
RNN	Recurrent Neural Network
sMBR	state-level Minimum Bayes Risk
SDR	Signal-to-Distortion Ratio
segSNR	segmental Signal-to-Noise Ratio
SRMR	Signal-to-Reverberation-Modulation Ratio
STFT	Short-Time Fourier Transform
TMS	Transcranial Magnetic Stimulation
VAD	Voice Activity Detector
WER	Word Error Rate
WSJ	Wall Street Journal

List of Symbols

$\mathbf{A}$	Dictionary containing speech and noise atoms
B	Number of frequency-bins
$\mathcal{C}$	Cost function
D	Dimension of the exemplars
D_{KLD}	Kullback-Leibler divergence
F	Number of frames in the spectrogram
$h[n]$	Room impulse response
$\mathcal{H}$	Short-time Fourier transform of $h[n]$
$\mathbf{H}$	Magnitude short-time Fourier transform of $h[n]$
J_s	Number of speech atoms
J_n	Number of noise atoms
M	Number of filters in a filter-bank
$\mathbf{N}$	Dictionary of noise atoms
$\mathbf{S}$	Dictionary of speech atoms
T	Number of consecutive frames used to create an exemplar
$w[n]$	Sampled noise signal
$\mathcal{W}$	Short-time Fourier transform of $w[n]$
$\mathbf{W}$	Magnitude short-time Fourier transform of $w[n]$
$y[n]$	Sampled clean speech signal
$\mathcal{Y}$	Short-time Fourier transform of $y[n]$
$\mathbf{Y}$	Magnitude short-time Fourier transform of $y[n]$
$z[n]$	Sampled noisy signal
$\mathcal{Z}$	Short-time Fourier transform of $z[n]$
$\mathbf{Z}$	Magnitude short-time Fourier transform of $z[n]$

λ_s	Speech sparsity penalty
λ_n	Noise sparsity penalty
$\boldsymbol{\Lambda}$	Sparsity weight matrix
μ	Mean vector
$\boldsymbol{\Sigma}$	Covariance matrix

$\odot$	Element-wise multiplication
$\oslash$	Element-wise division
$\mathbf{1}$	A vector/matrix with all elements equal to unity
$[\cdot]_p$	p-th column of a matrix
$\|\cdot\|_1$	ℓ_1 norm of a vector
$\|\cdot\|_2$	ℓ_2 norm of a vector

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is one of the most fundamental forms of human communication. Human speech has been researched for several years focussing on various speech technology applications such as speech transmission over mobile communication, background noise suppression, automatic speech recognition (ASR) and speaker identification. Most of these systems make use of a microphone to capture the speech signal and process it digitally to achieve the various tasks. However in the real world, the speech is recorded from environments with varying levels and types of background noises such as traffic noise, environmental noises and multi-talker babble noise as in a restaurant. Such background noises have a detrimental effect on the various speech technology applications such as poorer speech intelligibility over mobile communication and reduced performance of ASR and speaker identification systems.

Thus, the recorded signal fed to the various speech related applications is typically a mixture of the *clean* speech signal that we want to process and background noises. Robustness to such background noises is among the major limiting factors to the widespread deployment of speech related services. Even though the various state-of-the-art speech related applications show impressive performances under controlled environments, the performance of such applications degrades rapidly in real world scenarios. Suppression of the acoustic background noise is a relevant and challenging problem that can reduce listener fatigue by improving the intelligibility of speech and is crucial for a better and reliable performance of the various speech processing systems under such adverse conditions.

The family of speech enhancement algorithms include various approaches to recovering the clean speech signal from the noisy mixture by acoustic background noise reduction, dereverberation, separation of multiple speech signals from a mixture, bandwidth extension of narrow-band speech, correcting the distortions introduced by different recording equipments, etc. The term dereverberation denotes the process of removing the echos in a recorded signal when the recording is taken from an enclosed space. This thesis concentrates mostly on the acoustic background noise suppression and then proceeds to joint noise suppression and dereverberation, and this thesis uses the term speech enhancement to describe these topics. Such speech enhancement settings can be broadly classified into two categories: single-channel and multi-channel systems. Single-channel methods operate on recordings obtained using a single microphone whereas the multi-channel techniques operate on recordings that are obtained using an array of two or more microphones to exploit the spatial information.

This thesis concentrates on single-channel speech enhancement systems and introduces several novel approaches for noise suppression and dereverberation. This chapter aims at giving an introduction to the noise suppression schemes and the fundamental mathematical model used in this thesis that is later extended for dereverberation as well. The dereverberation problem is addressed in the later chapters of this thesis and a brief overview is given in Chapter 5. The fundamentals of ASR systems is also included in this chapter since the thesis also investigates the application of the proposed speech enhancement schemes as a pre-processing stage for ASR systems for an improved recognition performance.

1.1 Traditional single-channel noise suppression schemes

The problem of single-channel noise suppression aims at recovering the unknown clean speech signal from a noisy mixture. Such systems typically operate by exploiting the spectral diversity between the speech and noise signals. However, the frequency spectra of speech and noise often overlap and therefore such systems generally achieve noise reduction at the expense of speech distortion. Such methods require to estimate the noise statistics assuming the additive noise model,

$$z[n] = y[n] + w[n] \tag{1.1}$$

where, $z[n]$, $y[n]$ and $w[n]$ are respectively the sampled noisy mixture, speech signal and the additive background noise at index n. Thus the goal is to obtain an estimate of the clean speech signal from $z[n]$.

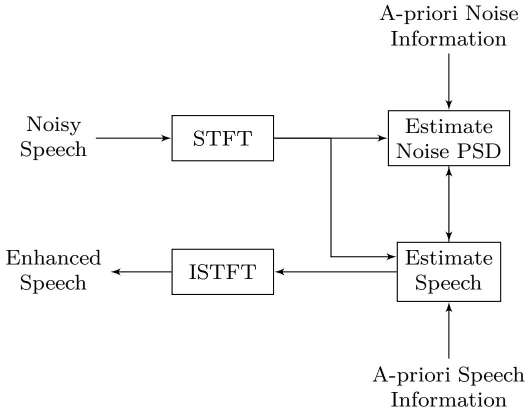


Figure 1.1: Block diagram of a traditional speech enhancement system.

In order to make use of the spectral diversity between speech and noise signals, we operate on the frequency domain over short segments of the recording, typically of the order of $20 - 30$ ms assuming the speech signal is wide-sense stationary over such small durations. The segmentation is done using a sliding window of appropriate shape to reduce the spectral leakage and with a window shift of typically 10 ms. Let F be the number of such sliding windows (or *frames*). The short-time Fourier transforms (STFT) over these windows are then obtained, which according to the additive model becomes:

$$\mathcal{Z}(b, f) = \mathcal{Y}(b, f) + \mathcal{W}(b, f) \tag{1.2}$$

where, $\mathcal{Z}$, $\mathcal{Y}$ and $\mathcal{W}$ are the STFTs of $z[n]$, $y[n]$ and $w[n]$, respectively. b and f denote the frequency-bin and frame indices respectively. Assuming that the speech and noise signals are statistically independent, the power spectral density (PSD) holds the following relation:

$$\mathbf{P}_z(f) = \mathbf{P}_y(f) + \mathbf{P}_w(f) \tag{1.3}$$

where the frequency-index b is omitted for brevity. Typical speech enhancement schemes attempt to estimate the PSD of noise and the spectral amplitudes of the noisy mixture are modified according to this estimate. Then the enhanced speech signal is obtained using the overlap-add method using the noisy phase [188].

The block diagram of a traditional speech enhancement system is given in Figure 1.1. An estimate of the noise PSD is obtained from the noisy mixture and using any a-priori information about noise if available. This PSD estimate is used to enhance the noisy mixture to suppress the noise, and the enhanced speech is obtained using the inverse STFT (overlap-add) operation. In addition, depending on the problem definition, any a-priori information available on the speech spectrum can also be used to get a better estimate of the clean speech.

As it can be seen, estimating the PSD of the additive noise is an essential component in traditional speech enhancement schemes. A common approach is to use a binary voice activity detector (VAD) to identify the segments of speech activity and estimate the noise statistics from segments where speech is absent [32,107,108]. Such settings assume that noise is contained in the speech-free regions of the signal. The noise statistics are updated for every segment of speech pause. There are soft-VAD versions as well where the VAD output is not binary. Such soft-VADs assign the probability of speech presence and allow estimating the noise PSD continuously during the speech activity as well [31,146,170]. However, the VAD becomes unreliable when the signal-to-noise ratios are low. Such models work reliably well when the noise is stationary which is not often the case in practice.

Another approach to adaptively estimate the noise PSD is based on the minimum statistics approach [120]. This approach exploits the fact that the PSD of the noisy signal often decays to that of the noise signal. For every frequency bin, the noise statistics are obtained by taking the minimum of a buffer containing the smoothed PSD of the noisy signal over frames. The minima-controlled recursive averaging technique [40,41] is a derivative of the minimum statistics method where the minima are taken from a recursively averaged power spectrum of noisy speech. The improved minima controlled averaging (IMCRA) presented in [40] uses two iterations of smoothing and minimum tracking.

The popular single-channel methods that make use of the estimated noise statistics include Wiener filtering, spectral subtraction, signal subspace models [79,91], missing data techniques [25,61] and Kalman filtering [28,56,99]. In the following subsections, the Wiener filtering and spectral subtraction -based techniques are briefly summarised since they appear in the later sections of the thesis. More information on the other mentioned techniques can be found in the references given above.

1.1.1 Wiener filtering

In the context of speech enhancement, the Wiener filter is obtained by minimising the mean-square error between the target clean speech $\mathbf{y}$ and its estimate $\hat{\mathbf{y}} = \mathbf{H}\mathbf{z} = \mathbf{H}\mathbf{y} + \mathbf{H}\mathbf{w}$, $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{w}$ are the vectorised versions (of length, say N) of the signals $y[n]$, $z[n]$ and $w[n]$, respectively. The filter $\mathbf{H}$ will thus be of size $N \times N$. The estimator error is $\epsilon = \mathbf{y} - \mathbf{H}\mathbf{y} - \mathbf{H}\mathbf{w}$. The minimum mean-square error (MMSE) estimate for $\mathbf{H}$ can be obtained as given below. The derivation assumes that the speech and noise signals are uncorrelated and stationary stochastic processes. In the frequency domain, the MMSE estimator is obtained as (derivations can be found in [122,174]):

$$\hat{\mathbf{H}} = \mathbf{P}_y(\mathbf{P}_y + \mathbf{P}_w)^{-1} = \mathbf{P}_y\mathbf{P}_z^{-1} \tag{1.4}$$

where, $\mathbf{P}_y$, $\mathbf{P}_z$ and $\mathbf{P}_z$ are diagonal matrices containing the PSDs of $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{w}$, respectively. However in practice, the PSD of clean speech is not known and an estimate is used instead. The estimate of the PSD of speech is obtained from (1.3) and the estimated noise PSD $\hat{\mathbf{P}}_w$ as $\hat{\mathbf{P}}_y = \max(\mathbf{P}_z - \hat{\mathbf{P}}_w, 0)$. The negative values are set to 0 since the PSD cannot be negative. $\mathbf{P}_z$ is typically obtained as a smoothed periodogram of the noisy signal. Thus the resulting enhanced speech spectra in the Wiener filtering approach is obtained as,

$$\hat{\mathcal{Y}}(f) = \frac{\max(\mathbf{P}_z(f) - \hat{\mathbf{P}}_w(f), 0)}{\mathbf{P}_z(f)} \cdot \mathcal{Z}(f). \tag{1.5}$$

1.1.2 Spectral subtraction

The spectral subtraction technique presented in [20] is based on a direct estimation of short-time spectral amplitude of clean speech. This technique makes use of an average estimate of the magnitude spectrum of the additive noise $|\hat{\mathcal{W}}(f)|$. The magnitude STFT of clean speech is directly obtained by subtracting this noise estimate from the magnitude STFT of the noisy speech signal. In order to obtain a non-negative magnitude spectrum, the negative values are set to 0. The resulting estimate of the STFT of clean speech is,

$$\hat{\mathcal{Y}}(f) = \max(|\mathcal{Z}(f)| - |\hat{\mathcal{W}}(f)|, 0) \cdot \frac{\mathcal{Z}(f)}{|\mathcal{Z}(f)|}. \tag{1.6}$$

One of the drawbacks of spectral subtraction schemes is that it suffers from musical noise arising from the randomly spaced peaks in the resulting enhanced spectrum that are caused by the random fluctuations in the periodogram. Algorithms that attempt to reduce such musical noise are presented in [24, 50,90,127]. Other variants of spectral subtraction can be found in [12,119,184].

1.2 Spectrogram factorisation

Another class of speech enhancement systems is based on decomposing the noisy speech spectrogram as a linear combination of speech and noise components. The fundamental assumption in such models is that the individual sound/noise events have their characteristic spectro-temporal patterns that can be represented using a *dictionary* of basis atoms. This thesis concentrates on such a class of enhancement schemes where the non-negative spectral representations of the noisy mixture are approximated as a linear combination of previously stored speech and noise spectro-temporal patches by means of non-negative matrix factorisation (NMF)-based approaches. Ever since its introduction [110], NMF has been successfully used for numerous source separation problems [157,167,185]. Given a dictionary containing atoms representing the sources, NMF-based algorithms decompose a noisy observation as a sparse non-negative linear combination of the atoms. In our framework, the atoms used are time-frequency representations of the training speech and of the noise data. The NMF-based decomposition thus yields estimates of speech and noise from the observation which can then be used to obtain a time-varying filter in the full-resolution frequency domain for speech enhancement.

Notice that the underlying concept in this model is strict additivity, where the magnitude of the sum of sound sources in the spectral domain is approximated as the sum of magnitudes of individual sound sources. Even though such a model holds only approximately, such models are shown to perform reasonably well in practice [62,185], especially when the signals are spectro-temporally sparse where the energy of each source is concentrated only over a limited amount of spectrogram bins.

One of the popular approaches in NMF-based algorithms is to use overcomplete dictionaries created using “exemplars” of speech and noise that are the directly sampled versions of the training speech and noise data itself [58,62, 89]. Another approach is to train the dictionary atoms from the training samples using the NMF updates [111], where generalisable models for speech and noise are learned as undercomplete dictionaries [132,156]. A study presented in [109] compares these two approaches and showed that the NMF-learned dictionaries outperform the exemplar-based dictionaries for speech enhancement in reverberant environments. However, the comparisons are done only with undercomplete dictionaries. It is also observed that, given enough training data to create overcomplete dictionaries, using exemplars from the training data as such leads to better separation performance than the NMF-learned dictionaries [130,168]. In this work we use overcomplete dictionaries where exemplars are expected to work better and we refer this approach to as “exemplar-based approach”.

Exemplar-based separation of speech and noise in a noisy recording makes use of a speech dictionary $\mathbf{S}$ containing J_s exemplars sampled from segments of clean speech and a noise dictionary $\mathbf{N}$ containing J_n exemplars sampled from segments of noise recordings. Exemplars are spectro-temporal representations of the recorded data, with the spectral axis referred to as *frequency bins* or *coefficients* and temporal axis as *frames*. The principle behind the approach is that the noisy speech, being an addition of speech and noise, can be approximated as a weighted sum of atoms in the speech and noise dictionaries. The exemplars may span multiple, say T, frames (which are reshaped to a vector) to capture the temporal dynamics [63]. Let D be the dimensionality of the resulting exemplars and $\mathbf{A} = [\mathbf{S}\ \mathbf{N}]$ be the dictionary of size $D \times (J_s + J_n)$ used for the decomposition. The various exemplar spaces used in this work are detailed in Section 2.3.

Given the non-negative representation of noisy speech $\mathbf{Z}$ and the exemplar dictionary $\mathbf{A}$, there are two popular models that decompose $\mathbf{Z}$ into its speech and noise components: one is called non-negative matrix factorisation where every T frames of the input test data is decomposed independently to get the mixing weights and the second one is called non-negative matrix deconvolution that models the frame-level speech and noise components as a convolution between the corresponding exemplars and mixing weights. Both of these paradigms are detailed below.

1.2.1 Non-negative matrix factorisation (NMF)

To obtain the NMF-based decomposition of noisy data, the noisy recording is first converted to the desired time-frequency representation used to create the dictionaries. Let this representation be denoted as $\mathbf{Z} \in \mathbb{R}_+^{B \times F}$, where B is the dimensionality of the spectral representation and F is the number of frames. A sliding window of length T frames is moved along its time axis at a hop size of 1 frame resulting in a total of $W = F - T + 1$ windows. The frames corresponding to each window are reshaped to a vector and are stacked as columns in the observation data matrix $\mathbf{\Psi}$ of size $D \times W$, where $D = B \cdot T$. This is then approximated as a weighted sum of the atoms in the speech and noise dictionaries to obtain the activations $\mathbf{X}$ (of size $(J_s + J_n) \times W$) as:

$$\mathbf{\Psi} \approx \tilde{\mathbf{\Psi}} = \begin{bmatrix}\mathbf{S} & \mathbf{N}\end{bmatrix} \begin{bmatrix}\mathbf{X_s} \\ \mathbf{X_n}\end{bmatrix} = \mathbf{AX} \quad s.t. \quad \mathbf{X} \geq 0 \tag{1.7}$$

where, $\mathbf{X_s}$ and $\mathbf{X_n}$ are the activations for the speech and noise dictionaries respectively and $\mathbf{X} = [\mathbf{X_s}^\intercal\ \mathbf{X_n}^\intercal]^\intercal$. Here, $\intercal$ denotes the matrix transpose. For the rest of the thesis, the subscripts $\mathbf{s}$ and $\mathbf{n}$ denote the speech and noise, respectively.

Given the dictionary of exemplars, the decomposition problem boils down to estimating the activations so that they minimise some cost metric between

$\boldsymbol{\Psi}$ and its approximation $\tilde{\boldsymbol{\Psi}}$. One of the common metrics used is the Eucledian distance which however is dominated by the largest differences between the elements in $\boldsymbol{\Psi}$ and $\tilde{\boldsymbol{\Psi}}$. In speech and audio related applications, the generalised Kullback-Leibler divergence (KLD) measure that emphasises differences between elements of smaller magnitude is found to perform better [185]. Other divergence measures are also proposed [37] and their performances are compared in various papers [36,38,52,53,112,172,197]. The generalised KLD between two vectors $\mathbf{x}$ and $\mathbf{y}$ of length B each is given as:

$$D_{KLD}(\mathbf{x}\|\mathbf{y}) = \sum_{b=1}^{B} \left(x_b \cdot log\frac{x_b}{y_b} - x_b + y_b \right) \tag{1.8}$$

where, x_b is the b-th element of a vector $\mathbf{x}$. The optimal values for the activations that minimise the KLD can be obtained using a gradient-descent technique that also ensures the non-negativity of the resulting solution (given in Section 1.2.3).

After this decomposition, we can obtain the windowed estimates of speech and noise as $\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{S}\mathbf{X}_{\mathbf{s}}$ and $\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{N}\mathbf{X}_{\mathbf{n}}$ respectively, each of size $D \times W$. Notice that there are multiple approximations of the same time-frequency frame appearing over multiple overlapping windows of these windowed estimates. To remove this windowing effect and to obtain the frame level estimates, we first append a zero matrix of size $D \times (T-1)$ to the windowed estimate, to get a matrix of size $D \times F$, and consider it as a block matrix having T block rows of size $(D/T) \times F$ each (notice that $D/T = B$). Let $\hat{\mathbf{s}}_{\mathbf{w},\tau}$ be the τ-th block matrix. The frame-level estimate of size $(D/T) \times F$ is then obtained, similar to an overlap-add method, as:

$$\hat{\mathbf{y}} = \sum_{\tau=1}^{T} \overset{\rightarrow(\tau-1)}{\hat{\mathbf{s}}_{\mathbf{w},\tau}} \tag{1.9}$$

where, $\overset{\rightarrow(\tau)}{(\cdot)}$ denotes right shifting a matrix by τ columns (prepending τ columns of zeros on the left and deleting τ columns on the right so as to maintain the original matrix size during addition). Averaging by the number of overlapping windows is omitted as it will be typically cancelled in the later processing stages. The frame-level noise estimate $\hat{\mathbf{w}}$ is obtained in the same manner.

From the frame level estimates $\hat{\mathbf{y}}$ and $\hat{\mathbf{w}}$ the corresponding time-varying filter is obtained by element-wise division as:

$$\mathbf{G} = \hat{\mathbf{y}} \oslash (\hat{\mathbf{y}} + \hat{\mathbf{w}}) \, . \tag{1.10}$$

This is then multiplied element-wise to the short-time Fourier transform (STFT) of the noisy speech $\mathcal{Z}$ of size $B \times L$, where B is the number of frequency bins used to obtain the STFT. The enhanced STFT, $\hat{\mathcal{Y}} = \mathcal{Z} \odot \mathbf{G}$, is converted to the time-domain by taking its inverse STFT using the overlap-add method to obtain the enhanced speech.

1.2.2 Non-negative matrix deconvolution

As mentioned before, NMF processes every window of T frames of the test data independently even if there is overlap between the consecutive processing windows. Non-negative matrix deconvolution [166] is proposed as an alternative to NMF where the approximation for $\mathbf{Z}$ is obtained convolutively over all the time (or frame) indices. Thus the resulting activations jointly generate a single approximation which is of the same size as the test data. Mathematically, NMD is formulated as:

$$\mathbf{Z} \approx \tilde{\mathbf{Z}} = \sum_{t=1}^{T} \mathbf{S_t} \overset{(t-1)\rightarrow}{\mathbf{X_s}} + \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\rightarrow}{\mathbf{X_n}} \tag{1.11}$$

$$= \sum_{t=1}^{T} \mathbf{A_t} \overset{(t-1)\rightarrow}{\mathbf{X}} \; . \tag{1.12}$$

The matrix $\mathbf{S_t}$ denotes the t-th block matrix obtained by partitioning $\mathbf{S}$ into T block rows each of size $B \times J_s$ [166]. $\mathbf{N_t}$ is also defined in the same manner from $\mathbf{N}$ and $\mathbf{A_t} = [\mathbf{S_t}\ \mathbf{N_t}]$. The approximation is obtained such that mixing weights or activations $\mathbf{X_s}$ and $\mathbf{X_n}$ are also non-negative. These activations can also be obtained by minimising some dissimilarity metric between $\mathbf{Z}$ and its approximation. Here also, we make use of KLD and the corresponding gradient-descent updates preserving the non-negativity of the activations can be found in Section 1.2.3. Notice that the NMD approximation using the optimal activations directly yields the frame-level estimates.

NMD has also been widely used for speech enhancement [85] and other speech related applications [59,88,183,194]. It is also observed that given a limited number of exemplars, NMD performs better than NMF [87].

1.2.3 Obtaining the activations

For both the NMF and NMD formulations, the approximations are done to obtain the activations $\mathbf{X}$ that minimise the generalised Kullback-Leibler divergence between $\boldsymbol{\Psi}$ or $\mathbf{Z}$ and its approximation with additional sparsity constraint on $\mathbf{X}$, which in matrix form is formulated as (for NMF):

$$\mathcal{C} = \sum_{d=1}^{D}\sum_{w=1}^{W}\left\{\boldsymbol{\Psi}_{d,w}\log\frac{\boldsymbol{\Psi}_{d,w}}{(\tilde{\boldsymbol{\Psi}})_{d,w}} - \boldsymbol{\Psi}_{d,w} + (\tilde{\boldsymbol{\Psi}})_{d,w}\right\} + \sum_{n=1}^{(J_s+J_n)}\sum_{w=1}^{W}(\boldsymbol{\Lambda}\odot\mathbf{X})_{n,w} \tag{1.13}$$

where $\boldsymbol{\Lambda}$ is a matrix of size $(J_s + J_n) \times W$ which, in effect, penalises the ℓ_1-norm of the activations and serves as a parameter to control the sparsity of $\mathbf{X}$. $\odot$ denotes element-wise multiplication. The cost function for NMD can be obtained by replacing $\boldsymbol{\Psi}$ by $\mathbf{Z}$. Notice that sparse activations are an integral part of such approximations especially when we use an overcomplete dictionary of exemplars. The sparsity forces the setting to use only a few best matching exemplars for approximation and yield a more plausible solution.

Notice that the sparsity penalty matrix $\boldsymbol{\Lambda}$ has a size equal to the number of atoms in the dictionary times the number of observation vectors. This matrix thus can be used to individually adjust the relative weight of any atom in the dictionary to approximate any column in the observation matrix $\boldsymbol{\Psi}$. However, in practise, the penalty is kept constant as λ_s for all speech atoms and λ_n for all noise atoms across all columns in the observation matrix, reducing the number of parameters to be tuned to two. $\boldsymbol{\Lambda}$ will thus have a structure comprised of an upper-block matrix of size $J_s \times W$ with all elements equal to λ_s and a lower block matrix of size $J_n \times W$ with all elements set as λ_n.

The cost function (1.13) can be minimised by iteratively applying the multiplicative-update rule on activations [64,111] using the method of positive and negative gradients [109,185]:

$$\mathbf{X} \longleftarrow \mathbf{X} \odot \frac{\nabla^{-}_{\mathbf{X}}\mathcal{C}}{\nabla^{+}_{\mathbf{X}}\mathcal{C}} \tag{1.14}$$

where, $\nabla^{-}_{\mathbf{X}}\mathcal{C}$ and $\nabla^{+}_{\mathbf{X}}\mathcal{C}$ are respectively the negative and the positive terms in the derivative $\nabla_{\mathbf{X}}\mathcal{C} = \partial\mathcal{C}/\partial\mathbf{X}$. It is shown that these multiplicative updates always result in a decreasing cost without affecting the non-negativity of the variable being updated [109,111]. The derivative of $\mathcal{C}$ with respect to the parameter $\mathbf{X}$

is :

$$\nabla_{\mathbf{X}}\mathcal{C} = \sum_{w=W}^{F}\sum_{d=1}^{D}\left(-\frac{\mathbf{\Psi}(d,w)}{\tilde{\mathbf{\Psi}}(d,w)}\frac{\partial\tilde{\mathbf{\Psi}}(d,w)}{\partial\mathbf{X}} + \frac{\partial\tilde{\mathbf{\Psi}}(d,w)}{\partial\mathbf{X}}\right) + \mathbf{\Lambda}. \tag{1.15}$$

For NMF, the resulting multiplicative update for obtaining the activations is:

$$\mathbf{X} \leftarrow \mathbf{X} \odot \frac{\mathbf{A}^{\intercal}\left(\frac{\mathbf{\Psi}}{\tilde{\mathbf{\Psi}}}\right)}{\mathbf{A}^{\intercal}\mathbf{1} + \mathbf{\Lambda}} \tag{1.16}$$

where all divisions are element-wise and $\mathbf{1}$ is a matrix of ones of size $D \times W$. This update rule is the bottleneck to the processing speed and computational complexity is linear in D, J_s, J_n and W.

Similarly for NMD, the multiplicative updates for the activations is:

$$\mathbf{X} \leftarrow \mathbf{X} \odot \frac{\sum_{t=1}^{T}\mathbf{A_t}^{\intercal} \quad \overset{\leftarrow(t-1)}{\frac{\mathbf{Z}}{\tilde{\mathbf{Z}}}}}{\sum_{t=1}^{T}\mathbf{A_t}^{\intercal} \quad \overset{\leftarrow(t-1)}{\mathbf{1}} \quad +\mathbf{\Lambda}}. \tag{1.17}$$

The derivation for the above updates can be found in Appendix A.

1.3 Automatic speech recognition

Apart from proposing novel speech enhancement schemes to improve the speech intelligibility, this thesis also investigates its applications to the various state-of-the-art automatic speech recognition (ASR) systems. One of the biggest issues the current ASR systems face is the degradation in performance due to added background noise. So in order to improve noise robustness, most of the ASR systems employ some mechanism which attempts to enhance the speech features by removing these artefacts [115]. Most of these mechanisms, like spectral subtraction [20], vector Taylor series [135], etc., work on spectro-temporal representations spanning a few tens of milli-seconds of the speech recording. An ASR system that makes use of longer contexts of data for recognising noisy speech that makes use of hidden-Markov model decomposition is presented in [179]. This work investigates the use of the proposed spectral factorisation-based speech enhancement schemes using exemplars which span hundreds of milli-seconds of the recorded data.

The main goal of an ASR system is to convert a speech recording into a word sequence that best fits the given recording. The basic architecture of an ASR

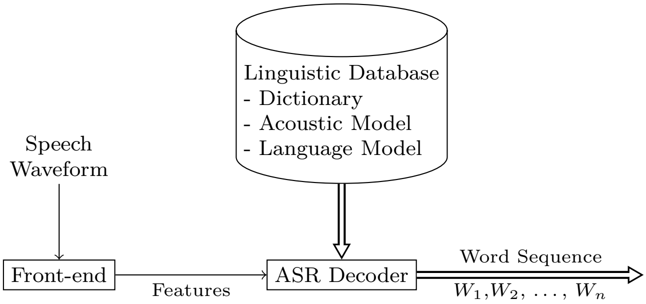


Figure 1.2: Architecture of an ASR system.

setting is given in Figure 1.2. An ASR system takes the recorded raw speech waveform as input and the front-end converts the raw signal into a sequence of feature vectors. Since the spectral characteristics of speech vary over time, these feature vectors are extracted over short intervals of time (also called as *frames*). Typically a time-window of length 25 ms and 100 windows per second are used, corresponding to a window-shift of 10 ms. These features are typically represented in the frequency domain and the goal of the ASR decoder is to uncover the underlying word sequence corresponding to these spectro-temporal patterns. The ASR decoder also makes use of language specific information such as a vocabulary (that defines the list of words to be recognised), pronunciation of words (or lexicon) and the grammar in order to arrive at the best matching word sequence.

For mathematical formulation, let the input speech signal be $y[n]$ and $\mathcal{O}$ be the corresponding sequence of feature vectors. The goal of the ASR decoder is to find the most probable word sequence $\mathbb{W} = \{W_1, W_2, \ldots, W_n\}$ obtained as:

$$\hat{\mathbb{W}} = \underset{\mathbb{W}}{\operatorname{argmax}}\ \mathbb{P}(\mathbb{W}|\mathcal{O}). \tag{1.18}$$

After applying Bayes' rule and omitting the normalisation, the above formulation becomes,

$$\hat{\mathbb{W}} = \underset{\mathbb{W}}{\operatorname{argmax}}\ \mathbb{P}(\mathcal{O}|\mathbb{W}) \cdot \mathbb{P}(\mathbb{W}) \tag{1.19}$$

where, $\mathbb{P}(\mathcal{O}|\mathbb{W})$ is the *likelihood* of the observed feature vectors $\mathcal{O}$ for a given word sequence $\mathbb{W}$ and $\mathbb{P}(\mathbb{W})$ is the prior probability of the word sequence $\mathbb{W}$ which is obtained using the language specific knowledge mentioned above.

The likelihood term is also known as the *acoustic model* in the ASR context since it yields the probability of an acoustic pattern to occur for a given word sequence, which is learned during the training phase of the ASR design using large corpora of speech recordings. Further, the words are decomposed into sub-word units such as phones or triphones in order to reduce the complexity of learning the likelihoods corresponding to every word in the vocabulary. Thus the acoustic model can be trained to yield likelihoods for an observed feature vector for a sequence of speech units defined by the ASR architecture. Notice that these speech units have to be decided before-hand depending on the complexity of the ASR task. For small vocabulary ASR task (e.g. digit recognition), a word level speech unit can be used whereas for larger vocabulary ASR tasks, phone level speech units are preferred. The prior probability $\mathbb{P}(\mathbb{W})$ is also referred to as the *language model* since it makes use of language specific information which maps the sequence of speech units to the corresponding word sequence and filters out grammatically unlikely sequences, thereby improving the recognition accuracy. These language models are typically learned from corpora of written text in the target language.

Since the ASR problem is the same as identifying the underlying sequence of speech units corresponding to the given acoustic feature pattern, typical ASR systems make use of hidden Markov models (HMM) that statistically model the observed features as the outputs of hidden state sequences. In this formulation, every frame of the input data is assumed to be *emitted* by a hidden state with some probability (or likelihood). These emission probabilities are to be learned during the training phase of the ASR setting. Thus, an HMM-based ASR decoder makes use of these emission probabilities as the acoustic model which yields the likelihood of the observed feature vectors given the HMM state sequence. Thus, HMMs can deal with the temporal variabilities in speech and the acoustic model determines how well an HMM state fits an observed feature vector. HMM state sequences are constrained by the lexicon allowing only speech unit sequences that correspond to valid words. State sequences can hence also be assigned a likelihood from the HMM's state transition model and the word sequence model, i.e. the language model. This thesis makes use of HMM-based ASR decoders for evaluation. Other ASR decoder variants include end-to-end speech recognition systems such as connectionist temporal classification [69] and attention-based neural networks [11,27,34] which will be briefly discussed in the later sections.

1.3.1 Acoustic modelling

As mentioned above, in an HMM-based ASR system the likelihoods for an observed feature vector given the HMM state sequence are computed from the emission probabilities. Thus the likelihoods are obtained for every frame in the observed feature vectors $\mathcal{O}$ for every HMM state. There are several approaches to obtaining these emission probabilities that include Gaussian mixture models (GMMs), template-matching followed by mapping to likelihoods [3,43,153] and neural networks that are trained to yield "pseudo-likelihoods" [81]. This thesis makes use of only GMM-based and neural network-based acoustic modelling which are discussed below.

Gaussian mixture models

Acoustic modelling using GMMs yields the emission probability of an HMM state q to emit a feature vector $\mathbf{o_t}$ at frame t as a weighted sum of multivariate Gaussian distributions as follows:

$$p(\mathbf{o_t}|q) \triangleq \sum_{k=1}^{N} w_k \cdot \mathcal{N}\left(\mathbf{o_t}; \mu_\mathbf{k}, \mathbf{\Sigma_k}\right). \tag{1.20}$$

Here, w_k are the mixing weights with $\sum_{k=1}^{N} w_k = 1$, N is the number of Gaussians in the GMM and $\mathcal{N}(\mathbf{o_t}; \mu_\mathbf{k}, \mathbf{\Sigma_k})$ is a multivariate Gaussian with mean $\mu_\mathbf{k}$ and covariance $\mathbf{\Sigma_k}$. Notice that these emission probabilities are to be obtained for every HMM state. Thus, the training phase of the ASR requires labelled training data that maps the feature vectors to different HMM states and the GMM tries to find the best fitting probability distribution using these feature vectors for every HMM state. Let the feature vectors be of length G. Acoustic modelling using GMMs makes use of the fact that it can approximate probability distributions to any required level of accuracy, provided that there are enough components.

Notice that every GMM is characterised using three sets of parameters, viz, N weights, N mean-vectors each of size $G \times 1$ and N covariance matrices of size $G \times G$ each which are to estimated during the acoustic model training. In order to reduce the number of parameters to be estimated, the covariance matrix is assumed to be diagonal which can be achieved by using decorrelated feature vectors. Popular choices of feature vectors include linear predictive coding (LPC) coefficients [4], Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) coefficients [77] and relative spectral transform - perceptual linear prediction (RASTA-PLP) features [78]. We refer the reader to [144] for further details on the training procedure. Other feature transform techniques for decorrelation, dimensionality reduction and speaker adaptation on top of these features have also been shown to improve the recognition accuracies. Popular feature transforms include principle component analysis (PCA) [95], linear discriminative analysis (LDA) [55], maximum-likelihood linear transform (MLLT) [154] and feature space maximum likelihood linear regression (fMLLR) [65].

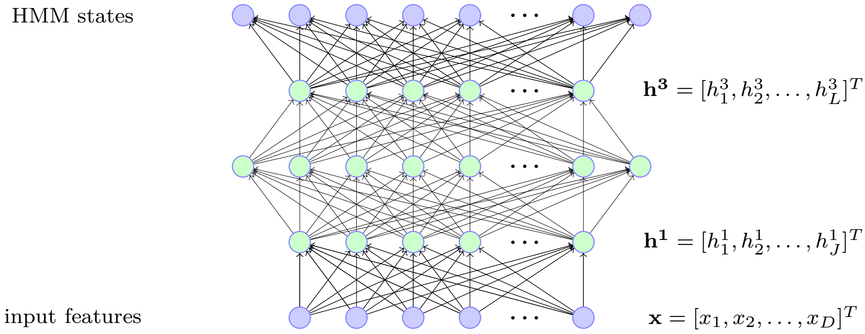

HMM states
$\mathbf{h^3} = [h^3_1, h^3_2, \ldots, h^3_L]^T$
$\mathbf{h^1} = [h^1_1, h^1_2, \ldots, h^1_J]^T$
input features
$\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$

Figure 1.3: Basic structure of a fully-connected DNN trained to yield HMM state posteriors.

Neural networks

Recently, neural network-based acoustic modelling has been successfully used to directly obtain the likelihoods at its output with significant performance improvements over the GMM-based likelihoods. In this framework, deep neural networks (DNNs) that have many hidden layers are trained to output the likelihoods over the HMM states with several frames of features as input [81]. The basic structure of a DNN that contains three hidden layers is shown in Figure 1.3. Let $\mathbf{x} = [x_1, x_2, \ldots, x_D]^\intercal$ be the input feature vector typically obtained by concatenating several frames of feature vectors. This vector is then fed-forward through the fully-connected hidden layers to finally yield the likelihoods.

In the figure every edge is associated with a multiplication with a weight and at the nodes the weighted values are summed and undergone a non-linear transformation. For mathematical formulation, let the first hidden layer contains J states and $\mathbf{h^1} = [h^1_1, h^1_2, \ldots, h^1_J]^\intercal$ be its output. Let w_{ij} be the weight associated with the connection between the i-th input state and j-th output state and $\mathbf{W^1}$ be the matrix formed by these weights where i and j are the row and column indices. Then the output of the first hidden layer can be written as $\mathbf{h^1} = f(\mathbf{W^1}^\intercal \mathbf{x} + \mathbf{b^1})$, where f denotes the element-wise non-linear transformation applied on the linear combinations of the input and $\mathbf{b^1}$ is the bias term. Popular choices of the non-linear functions include sigmoid non-linearity, rectified linear units (ReLU) and hyperbolic tangent (*tanh*). Thus at every hidden layer, the input data is multiplied by a matrix which acts like a feature extractor followed by a non-linear transformation.

As mentioned before, the output layer of a DNN is designed to yield the likelihoods corresponding to the HMM states. Thus the number of states in the output layer equals the number of HMM states N_h and the total input to the output layer $\mathbf{W^4}^\mathsf{T}\mathbf{h^3} + \mathbf{b^4}$ is converted into probabilities using the *softmax* function. Let $\mathbf{p} \in \mathbb{R}_+^{N_h \times 1}$ be the resulting “probabilities” which are obtained as

$$\mathbf{p} = \frac{\exp(\mathbf{W^4}^\mathsf{T}\mathbf{h^3} + \mathbf{b^4})}{\|\exp(\mathbf{W^4}^\mathsf{T}\mathbf{h^3} + \mathbf{b^4})\|_\mathbf{1}} \tag{1.21}$$

where, $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector. Since these are in fact not based on any probability distribution, these are also called as “pseudo-likelihoods”. All the weight matrices and the biases are learned during the training phase by minimising the cross-entropy between $\mathbf{p}$ and and the target probabilities $\mathbf{d}$,

$$\mathcal{C} = -\sum_{n=1}^{N_h} d_n \log p_n \tag{1.22}$$

where, d_n and p_n are the n-th component of the vectors $\mathbf{d}$ and $\mathbf{p}$, respectively. The target probability vector $\mathbf{d}$ is typically a one-hot vector obtained from the labelling information.

Intuitively, the first layers act as the feature detectors and the last layers act as the classifier that are jointly trained to get the optimal classification. Such a joint optimisation is one of the reasons why DNNs have a better discrimination ability than the GMMs where the features are typically engineered independent of the cost function that is being minimised during the GMM training. These pseudo-likelihoods generated by the DNNs have been successfully used for a wide variety of ASR applications [42,92,129,161]. Such systems are known as neural network-HMM (NN-HMM) hybrid systems.

Since the first few layers of a DNN architecture essentially act as feature detectors, several attempts have been made to replace the input part with other neural network architectures such as convolutional neural networks (CNN) [1,149], recurrent neural networks (RNN) or long short-term memory (LSTM) cells [152] in order to obtain better features that also take care of the spectral and temporal redundancies in the speech spectrum. Notice that such systems use a few DNN layers at the output to classify the features extracted using the CNN and LSTM layers [151].

1.3.2 Language modelling

The language model provides the prior probability of a word sequence $\mathbb{W}$ to occur in the target language. Typically, these probabilities are learned independently of the acoustics using large corpora of text from the target language. A language model can improve the recognition accuracy of an ASR setting by incorporating the semantic, syntactic and grammatic information of the target language. During the decoding phase, the language model also helps to reduce the possible number of hypotheses on $\mathbb{W}$ by suppressing grammatically unlikely word sequences. The popular language modelling techniques can be broadly divided into deterministic and probabilistic language models [83]. For most of the small vocabulary ASR tasks, deterministic language models such as context-free grammars would suffice. But for large vocabulary tasks, probabilistic language models such as N-grams are preferred.

The large vocabulary ASR evaluations presented in this thesis make use of N-grams which predicts the probability of a word based on the context, i.e. the preceding $N-1$ words. The probability for a word sequence $\mathbb{W} = \{W_1, W_2, \dots, W_n\}$ can be found using the chain rule as

$$\mathbb{P}(\mathbb{W}) = \prod_{j=1}^{n} \mathbb{P}(W_j | W_1, W_2, \dots, W_{j-1}). \tag{1.23}$$

An N-gram language model approximates the conditional probability of the word W_j to depend only on the preceding $N-1$ words, i.e. $\{W_{j-N+1}, W_{j-N+2}, \dots, W_{j-1}\}$. The prior probability of $\mathbb{W}$ is thus simplified to

$$\mathbb{P}(\mathbb{W}) \approx \prod_{j=1}^{n} \mathbb{P}(W_j | W_{j-N+1}, W_{j-N+2}, \dots, W_{j-1}). \tag{1.24}$$

The maximum likelihood estimator of the N-gram probability of a word is obtained by counting the instances of the context followed by the the target word appearing in the text and normalising it with the total instances of the context.

$$\mathbb{P}(W_j | W_{j-N+1}, W_{j-N+2}, \dots, W_{j-1}) = \frac{\text{count}(W_{j-N+1}, W_{j-N+2}, \dots, W_{j-1}, W_j)}{\text{count}(W_{j-N+1}, W_{j-N+2}, \dots, W_{j-1})} \tag{1.25}$$

Typically $N \leq 3$ is used to limit the complexity and cope with data sparsity problems since there will be fewer instances of the context word sequences for larger N resulting in unreliable estimates. Notice that, even with a small N some grammatically correct word sequences still might end up with very small, even zero, estimates simply because those do not appear in the chosen text corpora. In order to address these issues, several back-off and smoothing techniques have also been proposed in the literature [29,98,103]. Apart from the popular N-grams, recurrent neural network (RNN)-based language modelling is also becoming popular and several architectures have been proposed recently [123,124,195]. These topics are not discussed since those are beyond the scope of the brief introduction of this thesis.

1.3.3 End-to-end speech recognition

As described before, the traditional ASR systems that use an acoustic model and a language model to predict the most likely word sequence hypothesis have gained much success. The model still considers acoustic and language models as two different entities that are trained independent from each other. The field of end-to-end ASR approaches attempts to train both these models jointly using neural networks to directly yield character sequences at the output without using HMMs. These systems use acoustic feature vectors at its input and generates character level transcriptions of the underlying word sequence at its output.

One of such end-to-end trainable HMM-free neural models is called as connectionist temporal classification (CTC) [69] where the neural network predicts the posterior probability of a character (for e.g., English alphabets) for every frame in the input data. These are modelled as bag of characters that are later mapped to the corresponding words. Such CTC models combined with some word level language model (for rescoring) achieved promising results on various ASR benchmarks [2,70,75].

Another architecture used in end-to-end speech recognition makes use of a sequence-to-sequence model where the neural networks that learn to focus their "attention" to specific parts of their input which is named as *Listen, Attend and Spell* [26]. Such systems have an encoder-decoder structure [30,175]. The encoder part typically is comprised of a bi-directional RNN that converts the input speech to a suitable feature representation. These features are then fed to the decoder part that is an attention-based recurrent sequence generator to yield a sequence of characters [11]. Other ASR systems that make use of this architecture can be found in [10,33,34].

1.4 Scope of the thesis

The fundamental mathematical model used in this thesis is based on the spectrogram factorisation of noisy speech using exemplars stored in a dictionary. The common goal is to decompose the noisy speech into its speech and noise components for single-channel speech enhancement. Emphasis is given to improve the speech intelligibility in terms of various speech enhancement quality measures and to investigate how such settings can benefit when used as a front-end to the various state-of-the-art ASR systems. The work is motivated from earlier works on NMF-based decomposition of noisy speech [64,185] and the key factor that decides the performance of such settings is how well the constituent speech and noise can be differentiated in the chosen exemplar space. This thesis proposes several extensions to these models in order to achieve a better speech and noise separation for an improved speech enhancement quality.

The major part of thesis was done as part of the Marie-Curie ITN project INSPIRE (INvestigating Speech Processing In Realistic Environments) of which our contribution was aimed at incorporating our knowledge about human speech recognition into the ASR frameworks. This work introduces the perceptually motivated modulation envelope spectrogram (or modulation spectrogram) features, referred to as MS features, [8,71] into the field of spectrogram factorisation and proposes an efficient way to map these features back to the magnitude short-time Fourier Transform (STFT) for speech enhancement. The setting uses coupled dictionaries where the exemplars corresponding to the MS and STFT feature spaces are jointly extracted.

It was observed that the speech and noise separation capabilities of different exemplar spaces depends on the different noise types and the separation problem becomes difficult when the type of additive noise is not present in the noise dictionary. In order to achieve a better speech and noise separation especially in the presence of such unseen noises, hybrid exemplar spaces formed by combining different feature spaces are proposed. The technique also makes use of the previously proposed coupled dictionaries.

The thesis also evaluates and compares various features for training a state-of-the-art DNN-based acoustic model. In particular, the use of the perceptually motivated features such as MS features and Gabor filter-bank features [155] are investigated and are compared with the conventional features such as Mel, STFT and PLP features.

The thesis also addresses the difficult problem of single-channel speech enhancement in noisy and reverberant environments. An algorithm that incorporates the reverberation into the NMD-based spectrogram factorisation model is proposed and the multiplicative updates to jointly estimate the reverberation, anechoic speech and noise are provided.

Finally, the proposed speech enhancement schemes are applied to the field of clinical neuroscience for the pre-operative planning on patients with brain tumor. During the pre-operative planning, a neurosurgeon has to decide if the affected brain region is essential for the major functions such as motor movement and language related processes. To identify the functional relevance of a brain region for language related processes, a picture naming task together with magnetic stimulation of the relevant brain region (called transcranial magnetic stimulation or TMS) [17,74] has been effectively used. The methodology currently followed is to record the responses and to manually check the accuracy and the reaction time by listening to it. However, such a process is prone to substantial intra- and inter-observer variabilities [105,171]. A novel automatic and objective evaluation routine for the picture naming task using ASR and the proposed speech enhancement schemes is developed.

The algorithms proposed in this thesis are evaluated on various competitive benchmark databases which are described next.

1.4.1 Databases used

Aurora-2

AURORA-2 database [82] is a database based on the TI Digits corpus [114] containing utterances of digits from '0-9' and 'oh' sampled at 8 kHz. For training the acoustic models, a clean speech dataset and a noisy training dataset each containing 8 440 utterances are used. The noisy training set contains car, babble, subway and exhibition hall noises added artificially at signal-to-noise ratios (SNRs) of 5, 10, 15 and 20 dB.

For testing, test sets A and B of the database are used. Test set A contains one clean subset containing 1 001 recordings of clean speech and its noisy versions at varying SNRs -5, 0, 5, 10, 15 and 20 dB for every noise type present in the training set, summing to a total of 28 subsets. Test set B also has the same structure as in test set A but with four different noise types which are not present in the training data. The noise types in test set B are restaurant, train station, street and airport noises.

Aurora-4

AURORA-4 database is a large vocabulary continuous speech recognition database based on the Wall Street Journal-0 (WSJ0) corpus of read English speech. In order to study the effect of channel variations, the database contains two different sets of recordings: one recorded using a Sennheiser microphone (denoted as Mic1) and the second set recorded using multiple microphones (denoted as Mic2 or multicondition set). In order to train the acoustic models, there are four training conditions each containing 7 138 utterances which are listed below.

1. Clean Set : Noise-free speech recorded using the Sennheiser microphone.
2. Multi-clean Set : Noise-free speech recorded using multiple microphones.

3. Multinoise Set : Clean Set added with 6 additive noise types synthetically added at SNRs between 10 dB to 20 dB in steps of 1 dB.
4. Multicondition Set : Same as the multinoise set but created using the multi-clean training set.

The test set of the database contains 14 sets (test01-test14), each containing 330 utterances. Test01 (or test A) contains the clean utterances recorded with the single microphone (Sennheiser) and test02-test07 sets (or collectively test B) contain its noisy versions added with the six noise types at varying SNRs between 5 to 15 dB in steps of 1 dB. Test08 (or test C) contains the clean utterances recorded with multiple microphones and test09-test14 (or collectively test C) sets contain its noisy versions same as in test B. A development set of the same structure as of the test set is provided, but with a different set of 330 utterances, for parameter tuning and cross-validation. The six types of noise conditions used are car, babble, restaurant, street, airport and train station. Word error rates (WER) in % is used to compare the various ASR systems evaluated on this database.

TIMIT

TIMIT is a benchmark database for evaluating and comparing the phone recognition accuracy of various ASR systems in clean conditions [57]. The training set of the database contains 3 696 utterances recorded from 462 speakers with 8 utterances per speaker. For evaluation, the core test set is used, which contains 192 utterances with 8 sentences each from 24 speakers. The development set of the database contains 400 utterances from 50 speakers. The ASR setting used in this thesis is designed to recognise the underlying phone sequence that uses a phone set containing 39 symbols. The phone error rates (PER) in % are reported for ASR evaluations on the TIMIT database.

CHiME-3

The CHiME-3 challenge [14] targets the performance of an ASR setting in a real world, commercially motivated scenario where the recordings are obtained using a tablet fitted with a six-channel microphone array. It contains WSJ0 sentences recorded using an apparatus that yields six-channel recordings. There are real (REAL), and simulated (SIM) utterances in the database. The real utterances are recorded from four outdoor environments; bus, pedestrian street, cafe and street junction. Simulated data contains WSJ0 utterances that are filtered using an estimated impulse response and added noise from each of these environments. The development set contains 410 real and simulated utterances for each of the four environments. The test set also has the same structure, but with 330 different utterances. The test and the development sets contain a total of 2 840 and 3 280 utterances, respectively. The training data contains 7 138 utterances that are simulated from the WSJ0 training set and 1 600 real recordings taken from the four environments, adding to a total of 8 738 utterances. A detailed description of the CHiME-3 dataset can be found in [14].

Background noise from the CHiME challenge

The PASCAL-CHiME challenge [15] investigates ASR in the presence of background noises from a domestic environment. This thesis makes use of the background noise recordings provided with this database to simulate noisy speech for some experiments. These background noises are recorded from two rooms in a house: the lounge and the kitchen that capture multi-source noise from different locations including washing machine noise, kids running and playing, audio from the television and multiple speakers talking [35].

1.5 Thesis overview

This section provides a short overview of the thesis which proposes various novel algorithms for spectrogram factorisation -based speech enhancement models. A brief summary of the remaining chapters is provided below.

- **Chapter 2 :** A novel approach using coupled dictionaries for exemplar-based speech enhancement is proposed in this chapter in order to obtain better estimates of speech and noise for the time-varying filter. This chapter also introduces the perceptually motivated modulation spectrogram features to the field of exemplar-based techniques. The performance of the proposed scheme is evaluated using various speech intelligibility measures and by using it as a front-end of the state-of-the-art GMM and DNN-based ASR systems (with AURORA-2 and AURORA-4 databases).

  This chapter also includes our contribution to the CHiME-3 challenge where the online learning of coupled dictionary atoms from the test data is also proposed. The use of exemplar-based schemes as a front-end to a CNN-DNN-based ASR setting is also investigated.

This chapter is adapted from the following publications: [1] D. Baby, T. Virtanen, J. F. Gemmeke and H. Van hamme. *Coupled Dictionaries for Exemplar-based Speech Enhancement and Automatic Speech Recognition.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23 (11), pp 1788–1799, November 2015.

[2] D. Baby, T. Virtanen and H. Van hamme. *Coupled Dictionary-based Speech Enhancement for the CHiME-3 Challenge.* Technical report KUL/ESAT/PSI/1503, Leuven, Belgium, KU Leuven, ESAT. September 2015.

- **Chapter 3 :** This chapter introduces the aforementioned idea of hybrid-exemplar spaces obtained by concatenating two different feature spaces to combine the speech and noise discrimination capabilities of different feature spaces. The goal is to combine the speech and noise separation capabilities of different feature spaces to get a better speech enhancement quality. The proposed framework makes use of coupled dictionaries and investigates several combinations of features for a better speech and noise separation. The experiments were conducted on the AURORA-2 database.

  This chapter is adapted from: Deepak Baby and Hugo Van hamme. *Hybrid Input Spaces for Exemplar-based Noise Robust Speech Recognition using Coupled Dictionaries.* 23rd European Signal Processing Conference (EUSIPCO), pp. 1676-1680, September 2015.

- **Chapter 4 :** Motivated from the performance of the modulation spectrogram features in exemplar-based speech enhancement from the previous chapters, this chapter investigates the use of these features for DNN-based acoustic modelling. The chapter presents comparison between different features such as Mel, PLP, Gabor filter-bank features and the STFT features when used as an input to the DNN. Evaluations are provided on TIMIT and AURORA-4 databases.

  This chapter is adapted from: Deepak Baby and Hugo Van hamme. *Investigating Modulation Spectrogram Features for Deep Neural Network-based Automatic Speech Recognition.* Proc. INTERSPEECH, ISCA, pp. 2479–2483, September 2015.

- **Chapter 5 :** This chapter extends the spectrogram factorisation models to achieve joint denoising and dereverberation for enhancing speech recordings taken from a noisy enclosed room and when the speaker is far from the microphone. A novel algorithm to incorporate and estimate the room impulse response (RIR) with a decaying norm constraint is proposed. The chapter provides multiplicative updates to jointly estimate the RIR, its decay and the estimates of anechoic speech and noise. The updates are derived by using the NMD-based spectral factorisation model.

This chapter is adapted from: Deepak Baby and Hugo Van hamme. *Joint Denoising and Dereverberation using Exemplar-based Sparse Representations and Decaying Norm Criterion.* Submitted to IEEE/ACM Trans. on Audio, Speech and Language Processing, 2016.

- **Chapter 6 :** An application of the proposed speech enhancement settings as the front-end of an ASR setting for clinical neuroscience is presented in this chapter. A novel framework to automate the reaction time measurement on a picture naming task is proposed. The responses are recorded and the algorithm operates on these recordings to identify if the responses were indeed correct and it also yields the reaction times for the correct responses. The algorithm makes use of SPRAAK [44] for ASR and the evaluations are performed on the data collected from patients/volunteers.

  This chapter is adapted from: Deepak Baby, Laura Seynaeve, Patrick Dupont, Wim Van Paesschen and Hugo Van hamme. *An automatic evaluation routine for picture naming task with transcranial magnetic stimulation using machine speech recognition.* Submitted to the Journal of Neuroscience Methods, 2016.

- **Chapter 7 :** This chapter concludes the thesis by listing the original contributions and suggestions for future work.

Chapter 2

Coupled Dictionaries for Exemplar-based Speech Enhancement & ASR

For decomposing the noisy speech in traditional exemplar-based speech enhancement systems, exemplars sampled in lower dimensional spaces are preferred over the full-resolution frequency domain for their reduced computational complexity and the ability to better generalise to unseen cases. But the resulting filter may be sub-optimal as the mapping of the obtained speech and noise estimates to the full-resolution frequency domain involves a low-rank approximation. This chapter presents an efficient way to directly compute the full-resolution frequency estimates of speech and noise using coupled dictionaries: an input dictionary containing atoms from the desired exemplar space to obtain the decomposition and a coupled output dictionary containing exemplars from the full-resolution frequency domain. We also introduce modulation spectrogram features for the exemplar-based tasks using this approach. The proposed system was evaluated for various choices of input exemplars and yielded improved speech enhancement performances on the AURORA-2 and AURORA-4 databases. We further show that the proposed approach also results in improved word error rates (WERs) for the speech recognition tasks using GMM-HMM and DNN-HMM -based systems.

This chapter also includes our contribution to the CHiME-3 challenge where the coupled dictionary -based speech enhancement setting is used as a front-end to the various ASR decoders provided by the challenge organisers [14].

The algorithm is also extended to learn adaptive atoms to model unseen noise cases and the coupled atoms are also learned from the test data. We also introduced a CNN-DNN-based decoder for CHiME-3 evaluation and it is shown that the coupled dictionary-based speech enhancement together with adaptive noise dictionaries significantly improves the ASR performance.

This chapter is adapted from: D. Baby, T. Virtanen, J. F. Gemmeke and H. Van hamme. *Coupled Dictionaries for Exemplar-based Speech Enhancement and Automatic Speech Recognition.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23 (11), pp 1788–1799, November 2015.

D. Baby, T. Virtanen and H. Van hamme. *Coupled Dictionary-based Speech Enhancement for the CHiME-3 Challenge.* Technical report KUL/ESAT/PSI/1503, Leuven, Belgium, KU Leuven, ESAT. September 2015.

2.1 Introduction

Speech recordings taken from realistic environments typically contain degradations along with the required speech signal which reduce its intelligibility and also result in poor performance of speech related tasks like automatic speech recognition (ASR), automatic voice assistance, etc. Therefore, some speech enhancement mechanism is deployed as the first step in most of these applications to circumvent the degradations which are mainly introduced by the background noise and room reverberation.

In scenarios where a model for speech and noise is not known *a priori*, unsupervised techniques like spectral subtraction [20], Kalman filtering [68], using the periodic structure in speech [93], etc., have been successfully used for speech enhancement. But most of these approaches rely on stationarity assumptions on the noise, which are often invalid for realistic data. Alternatively, supervised techniques can yield improved performance using codebook based [173] or model based [49] approaches, since the models for speech and noise are known *a priori*.

In this work, we investigate speech enhancement on a single channel noisy recording in the presence of additive noise using non-negative matrix factorization (NMF) algorithms employing exemplars of speech and noise. The performance of an exemplar-based approach depends on two key factors: First, on how well the speech and noise can be differentiated in the chosen time-frequency representation or the “exemplar space”. Popular choices of exemplar spaces include Mel-integrated magnitude spectra [62], DFT (refers to the magnitude of the short-time Fourier transform in this work) [145] and Gabor filterbank coefficients [89]. Using DFT as the exemplar space has the advantage that the time-varying filter can be directly obtained in the full-resolution frequency (DFT) domain. However, such systems suffer from increased computational complexity, poor speech and noise separation especially in presence of babble noise [134] and inability to generalise well for unseen noise cases [7]. It is observed that using lower dimensional features like the Mel features can address most of these issues fairly well [7] and this introduces the second factor: how well we can map the resulting lower-dimensional estimates to the DFT space to obtain the time-varying filter? Most of the current approaches make use of a pseudo-inverse [58] to obtain the mapping which always yield a low-rank approximation of the estimates, resulting in a sub-optimal filter which cannot account for all the added noise content and results in poorer noise suppression.

In this work, we have three main goals. First, to effectively utilise the advantages of the low-dimensional features and to address the low-rank approximation, we propose to use coupled dictionaries, which has been used earlier to increase the spectro-temporal resolution [130,137], voice conversion [193] and dimensionality reduction for multi-label learning [67]. In this work, we make use of two (coupled) dictionaries: an input dictionary containing atoms sampled in the exemplar space where the NMF-based decomposition is to be done and a coupled output dictionary containing the corresponding DFT exemplars to directly reconstruct the estimates in the DFT domain. This approach thus can obtain a better decomposition at a reduced computational complexity, and make use of the resulting weights or *activations* of the input dictionary atoms to directly reconstruct the DFT estimates using the coupled output dictionary, which will be explained in Section 2.2.

Second, we introduce using modulation spectrogram (MS) features [71] for exemplar-based speech enhancement. The MS representation for speech was introduced as part of a computational model for human hearing and a better separation between speech and noise can be expected in the MS domain considering the fact that speech and noise often have different modulation frequency contents. However, obtaining the MS representation involves non-linear operations making it hard to invert to the frequency domain where the mixture signal is processed. In this work, we investigate the use of coupled dictionaries to reconstruct the underlying DFT features following the decomposition in the MS domain for exemplar-based speech enhancement and ASR tasks.

Finally, we investigate the performance of various state-of-the-art automatic speech recognition (ASR) tasks on these enhanced speech data. ASR evaluation serves two purposes in this work. First, the recognition performance acts as an additional evaluation measure to assess the utility of the enhanced speech data on small and large vocabulary speech recognition. Second, we investigate how much the HMM-GMM based and deep-neural network (DNN) based state-of-the-art ASR systems can benefit from making use of the enhanced data.

The rest of the chapter is organised as follows: Section 2.2 details the proposed exemplar-based speech enhancement technique using coupled dictionaries. The various choices of input exemplars investigated in this work are described in Section 2.3. The evaluation setup is explained in Section 2.4 followed by some results and observations made on the experiments done on the AURORA-2 database in Section 2.5. Section 2.6 details the results obtained for speech enhancement and ASR evaluations on the AURORA-4 database. Our contribution to the CHiME-3 challenge using the proposed speech enhancement setting is given in Section 2.7. Section 2.8 concludes the work along with some directions for future work.

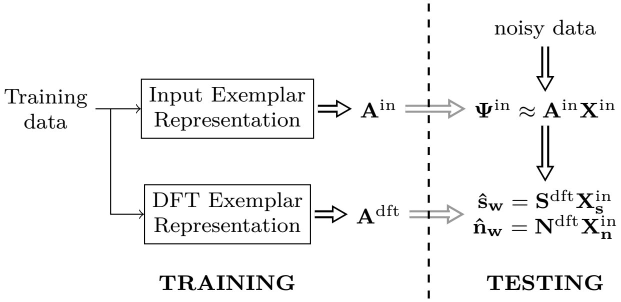

noisy data
Training data
Input Exemplar Representation
$\mathbf{A}^{\text{in}}$
$\boldsymbol{\Psi}^{\text{in}} \approx \mathbf{A}^{\text{in}}\mathbf{X}^{\text{in}}$
DFT Exemplar Representation
$\mathbf{A}^{\text{dft}}$
$\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{S}^{\text{dft}}\mathbf{X}^{\text{in}}_{\mathbf{s}}$
$\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{N}^{\text{dft}}\mathbf{X}^{\text{in}}_{\mathbf{n}}$
TRAINING
TESTING

Figure 2.1: Block diagram overview of the proposed system using modulation spectrogram features and coupled dictionaries.

2.2 Speech enhancement using coupled dictionaries

The proposed approach to obtain the DFT estimates using coupled dictionaries is summarised in Fig. 2.1. In this approach, the NMF-based decomposition is obtained in an additive and non-negative feature space of choice which serves as the front-end of the speech enhancement system. For simplicity, the front-end features are referred to as "input exemplars" and the dictionary used to obtain the NMF compositional model is denoted as $\mathbf{A}^{\text{in}} = [\mathbf{S}^{\text{in}}\ \mathbf{N}^{\text{in}}]$. This dictionary has a size $D^{\text{in}} \times (J_s + J_n)$, where D^{in} is the dimensionality of the input exemplar space. The observation data matrix in the input exemplar domain $\boldsymbol{\Psi}^{\text{in}}$ is decomposed using $\mathbf{A}^{\text{in}}$ as explained in section 1.2.1. The resulting activations $\mathbf{X}^{\text{in}}$ are then applied with the output DFT dictionary to directly obtain the windowed speech and noise estimates in the DFT domain as $\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{S}^{\text{dft}}\mathbf{X}^{\text{in}}_{\mathbf{s}}$ and $\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{N}^{\text{dft}}\mathbf{X}^{\text{in}}_{\mathbf{n}}$, respectively. For the remaining part of the thesis, the superscripts denote the type of exemplar space used.

To obtain a reliable reconstruction of the underlying DFT estimates, the mapping between the corresponding atoms in both the dictionaries should nearly be one-to-one. Such an approximation would work if the input and the output DFT exemplars are temporally aligned and scale alike with signal strength. Regarding the last criterion, signal representations that vary linearly with the input signal strength work best in conjunction with the considered cost function (1.13). These are achieved by properly choosing the input exemplars and extracting the corresponding DFT exemplars from the same piece of training data spanning T frames (ref. Fig. 2.1).

From the windowed estimates, the frame level estimates $\hat{\mathbf{y}}$ and $\hat{\mathbf{w}}$ are obtained by removing the windowing effect and the corresponding time-varying filter is obtained by element-wise division as:

$$\mathbf{G} = \hat{\mathbf{y}} \oslash (\hat{\mathbf{y}} + \hat{\mathbf{w}})\,. \tag{2.1}$$

This is then multiplied element-wise to the short-time Fourier transform (STFT) of the noisy speech $\mathcal{Z}$ of size $B \times L$, where B is the number of frequency bins used to obtain the STFT. The enhanced STFT, $\hat{\mathcal{Y}} = \mathcal{Z} \odot \mathbf{G}$, is converted to time-domain using overlap-add method to obtain the enhanced speech. Notice that the DFT dictionary is of size $D^{\mathrm{dft}} \times (J_s + J_n)$, where $D^{\mathrm{dft}} = B \cdot T$ and the time-varying filter has the same size as $\mathbf{Y}$. In short, the proposed method thus can exploit the speech and noise separation capabilities for various choices of input spaces and can generate a filter which has full-rank in the DFT space.

2.3 Choice of input representation

The various choices for input representation that are investigated in this work are explained in this section. Notice that the underlying assumption in the exemplar-based approach is that the speech and noise are approximately additive in the chosen exemplar spaces. The processing chains for obtaining the coupled exemplars are summarised in Fig. 2.2.

2.3.1 DFT exemplars

First, the DFT space is chosen as the input exemplar space to obtain the decomposition. To obtain DFT exemplars to create the DFT dictionary, a segment of length T frames (T_t seconds in time domain) of training data is chosen at random and its magnitude STFT is used for non-negativity. Let the STFT be obtained using a window length and hop size of t_w^{dft} and t_h^{dft}, respectively. This yields a spectro-temporal representation of size $2B \times T$, where $2B$ is the number of frequency bins used to obtain the STFT. Since the magnitude STFT is symmetric, only the positive half is considered. This is reshaped to a vector of size $(B \cdot T) \times 1$ to obtain the DFT exemplar. i.e., $D^{\mathrm{dft}} = B \cdot T$.

During evaluation, the NMF-based decomposition is done in the DFT space after converting the noisy observation into its equivalent DFT exemplar representation. The resulting activations are used to obtain the frame-level speech and noise estimates, and the enhanced speech is obtained as explained in Section 2.2. This setting is chosen as one of the baseline systems in this work and is denoted as *DFT-DFT* setting.

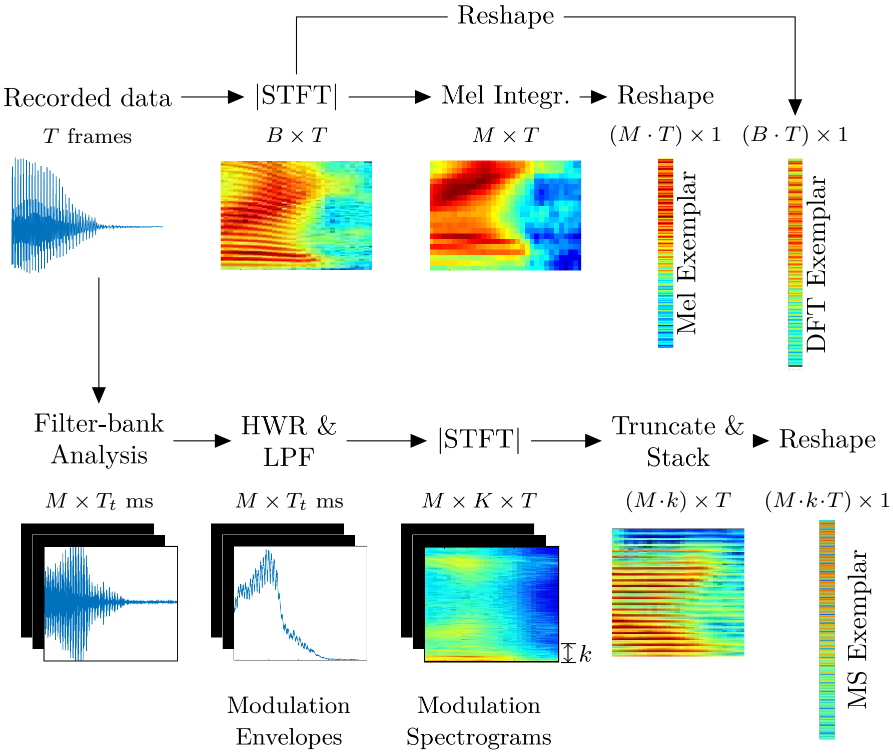


Figure 2.2: Block diagram overview of the processing chains to obtain various exemplars. All the coupled exemplars are extracted from the same piece of recorded data spanning T frames (T_t seconds in time-domain). The resulting representations along with their size are shown below in each of the steps. Figures are not shown at the same scale.

2.3.2 Mel exemplars

Mel exemplars are chosen for their lower dimensionality and robust speech and noise separation performance in the presence of a variety of noises. First, the Mel features for T frames of data are obtained after applying Mel-integration of the magnitude STFT as depicted in Fig. 2.2. This is done by multiplying the magnitude STFT by the DFT-to-Mel matrix $\mathbf{M}$ which contains the magnitude response of M Mel bands along its rows. The resulting representation of size $M \times T$ is reshaped to a vector to obtain the Mel exemplar of length $D^{\mathrm{mel}} = M \cdot T$. The Mel dictionaries for speech and noise are denoted as $\mathbf{S}^{\mathrm{mel}}$ and $\mathbf{N}^{\mathrm{mel}}$, respectively.

During the test phase, the noisy data represented in the Mel exemplar space is decomposed using the Mel dictionary $\mathbf{A}^{\mathrm{mel}} = [\mathbf{S}^{\mathrm{mel}}\ \mathbf{N}^{\mathrm{mel}}]$ and the corresponding activations $\mathbf{X}_{\mathbf{s}}^{\mathrm{mel}}$ and $\mathbf{X}_{\mathbf{n}}^{\mathrm{mel}}$ are obtained. Once these activations are obtained, we use it to evaluate two systems.

First, another baseline system is defined which is denoted as the *Mel-Mel* setting. In this setup, the windowed speech and noise estimates are obtained using the Mel dictionary as $\mathbf{S}^{\mathrm{mel}}\mathbf{X}_{\mathbf{s}}^{\mathrm{mel}}$ and $\mathbf{N}^{\mathrm{mel}}\mathbf{X}_{\mathbf{n}}^{\mathrm{mel}}$, respectively. The frame level Mel estimates, $\hat{\mathbf{y}}'$ and $\hat{\mathbf{w}}'$ are obtained as explained in Section 1.2.1. These are then mapped to the DFT domain using the pseudo-inverse of the DFT-to-Mel matrix, $\mathbf{M}^{\dagger} = \mathbf{M}^{\intercal}(\mathbf{M}\mathbf{M}^{\intercal})^{-1}$ to obtain the enhanced STFT as [58]:

$$\hat{\mathcal{Y}} = \mathcal{Z} \odot \left(\mathbf{M}^{\dagger}\left[\hat{\mathbf{y}}' \oslash (\hat{\mathbf{y}}' + \hat{\mathbf{w}}')\right]\right). \tag{2.2}$$

It is evident that this setting has a lower computational complexity as $M \ll B$ while performing the multiplicative updates. It is also observed that Mel features have a better speech and noise separation capability and generalise better for unseen noise cases when compared to the DFT exemplars [7]. However, the pseudo-inverse mapping in (2.2) will always fall in a subspace of rank M spanned by the rows of $\mathbf{M}$. The frequency response of the Mel filter-bank being triangular, such a mapping is equivalent to a piece-wise linear approximation of M points located at the central frequencies of the filter-bank. It is thus clear that such a transformation may not be able to model most of the speech and noise content in the full-resolution DFT space with $M \ll B$, which in turn may reduce the speech enhancement quality. This issue will be further explored in later sections.

For the second setting, we investigate the proposed approach using Mel exemplars as the input features to deal with the low-rank approximation in the Mel-Mel setting. Here, the underlying (windowed) DFT estimates for speech and noise are directly obtained using the Mel activations as $\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{S}^{\mathrm{dft}}\mathbf{X}_{\mathbf{s}}^{\mathrm{mel}}$ and $\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{N}^{\mathrm{dft}}\mathbf{X}_{\mathbf{n}}^{\mathrm{mel}}$, and are then used for speech enhancement (ref. Section 2.2). This is referred to as the *Mel-DFT* setting. Since in this setting, the output DFT dictionary is coupled to the Mel input dictionary and is overcomplete, a full-rank reconstruction of the estimates can be enforced and a better noise suppression could be achieved.

2.3.3 MS exemplars

The modulation spectrogram (MS) representation of speech was proposed as part of a computational model for human hearing which relies on low frequency amplitude modulation variations within frequency bands [141]. These variations play a key role in the higher level human auditory processing [21] and are computationally modelled as modulation envelopes. The bottom row in Fig. 2.2 summarises the processing chain to obtain the MS representation for speech.

To obtain the modulation envelopes, the acoustic data is first filtered using a filter bank containing M channels to model the frequency discrimination property of the basilar membrane. The resulting M bandlimited signals are half-wave rectified to model the non-negative nerve firings followed by low-pass filtering to obtain the modulation envelopes. The 3dB cut-off frequency of the low-pass filter used is around 20Hz as human speech contains modulations of very low frequency [158] and hence the spectrograms of these envelopes, called the modulation spectrograms, can yield a more effective representation [71] in comparison to the time domain. Let the window length and hop size used to obtain the MS representation be t_w^{MS} and t_h^{MS}, respectively. The MS representation is typically obtained over longer window lengths when compared to the DFT features (i.e., $t_w^{\mathrm{MS}} > t_w^{\mathrm{dft}}$), to capture the variation in modulation envelopes, which also allows larger choices for t_h^{MS} than t_h^{dft}. This representation of speech has successfully been used for blind source separation [16] and noise-robust ASR [100].

Notice that converting acoustic data into the MS space results in a three-dimensional representation of size $M \times K \times T$, where M, K and T are the number of input frequency channels, number of modulation frequency bins and number frames in the acoustic data, respectively. However, since the modulation envelopes are obtained after a low-pass filtering operation, only a few bins in the MS will contain significant energy and it is possible to truncate each of the MS to the lowest few, say k, bins. These truncated M modulation spectrograms, each of size $k \times T$, are stacked to get a two-dimensional representation of size $(Mk) \times T$, referred to as the *MS features*. This representation is then reshaped to a vector to obtain the MS exemplar. The dimensionality of an MS exemplar will thus be $D^{\mathrm{MS}} = M \cdot k \cdot T$. In our previous works [7,8] we showed that the approximate additivity assumption of speech and noise is valid in the MS exemplar space as well. In comparison to the established Mel exemplar-based approaches, the MS representation essentially retains the same information within each frequency band for each frame, but also more accurate information about the spectral distribution of different modulation frequencies.

In this work, the MS exemplars are used as input exemplars to obtain the NMF-based decomposition using the dictionary of MS exemplars $\mathbf{A}^{\mathrm{MS}} = [\mathbf{S}^{\mathrm{MS}}\ \mathbf{N}^{\mathrm{MS}}]$ to obtain the activations $\mathbf{X}^{\mathrm{MS}}$. However, since the processing chain to obtain the MS features involves non-linear operations, there is no direct way to make use of this decomposition to enhance the noisy speech as the inversion of the

MS features to the time domain is not unique. We propose using the coupled DFT dictionary extracted together with the MS dictionary to reconstruct the DFT estimates and to obtain speech enhancement, i.e., the speech and noise estimates are approximated as $\mathbf{S}^{\mathrm{dft}}\mathbf{X}_{\mathbf{s}}^{\mathrm{MS}}$ and $\mathbf{N}^{\mathrm{dft}}\mathbf{X}_{\mathbf{n}}^{\mathrm{MS}}$, respectively. The resulting frame-level estimates are used to enhance the noisy spectrogram. This system is denoted as the *MS-DFT* setting.

However, any circular temporal shift (modulo the window length) of the DFT spectrogram can yield the same MS representation and makes the mapping many-to-one. To address this, we make use of temporal oversampling, i.e., smaller t_h^{MS} while obtaining the MS, to reduce this ambiguity as pointed out in [72]. In our previous work, setting $t_h^{\mathrm{MS}} = t_h^{\mathrm{dft}}$ was found to be the best choice [8]. It is also to be noted that increasing the low-pass cut of frequency beyond 20 Hz should be useful for a better speech and noise separation when the data is corrupted by some noise having higher modulation frequencies. This on the other hand requires a higher value of k which increases the computational complexity and may lead to data overfitting. Hence, a compromise must be pursued.

2.4 Experimental setup

2.4.1 Databases

To evaluate and compare the various settings, two databases were used. Preliminary experiments were conducted on the AURORA-2 database which is a small-vocabulary task and are then extended to the large-vocabulary database AURORA-4 [82].

1) AURORA-2 Database: is a database based on the TI Digits corpus containing utterances of digits from '0-9' and 'oh' sampled at 8 kHz. For training the acoustic models, a clean speech dataset and a noisy training dataset each containing 8 440 utterances are used. For testing, test sets A and B are used. The description of the database can be found in Section 1.4.1.

2) AURORA-4 Database: is a large vocabulary continuous speech database based on the WSJ-0 corpus of read speech. In this work, only the single microphone test set with 16 kHz sampling frequency, which contains a noise free data set (test 01 or test A) with six noisy sets (test 02-07 or collectively test B) corrupted with car, babble, restaurant, airport, street and train noises added artificially at varying SNRs between 5 and 15 dB in steps of 1 dB, is used. The development set of the database was for validation and parameter tuning. For training the acoustic models and preparing the dictionaries, the clean and the multi-noise training sets containing 7 138 utterances each were used. The description of the database is given in Section 1.4.1.

2.4.2 Exemplars and dictionary preparation

The dictionaries used to obtain the decomposition were prepared from the training data. The noise data used to create the noise exemplars were obtained from the noisy training data using the two-step procedure described in [64]. The dictionaries were created using exemplars originating from random segments of length T frames taken from the clean and noise training sets. Throughout this work, the choice of T used was 30 and 15 frames for the AURORA-2 and AURORA-4 databases, respectively as these values were found to yield the best performance on similar tasks [58,62]. The value of T for AURORA-4 is chosen to be smaller than the AURORA-2 database as the former has a lot more variety of speech to be modelled as opposed to the latter and it demands a larger dictionary to reasonably model the large vocabulary speech data, which increases the computational complexity.

Every chosen random segment of length T_t seconds was first pre-processed by removing the DC component and applying a pre-emphasis filter (a single order high-pass filter of coefficient 0.97). The coupled exemplars were then extracted as follows (ref. Fig. 2.2):

1. The STFT of the samples were obtained using a Hamming window of length $t_w^{\text{dft}} = 25$ ms and a hop size $t_h^{\text{dft}} = 10$ ms. The magnitude of the STFT is then obtained yielding a representation of size $B \times T$. This is then reshaped to obtain the DFT exemplar of length $B \cdot T$.

2. The magnitude STFT obtained in the step above is pre-multiplied with the DFT-to-Mel matrix $\mathbf{M}$ of size $M \times B$ to obtain the Mel-integrated magnitude spectra of size $M \times T$. The Mel exemplar is then obtained by reshaping the Mel spectra.

3. To obtain the MS representation, the time-domain signal is first filtered into M band-limited signals using the equivalent rectangular bandwidth filter banks implemented using Slaney's toolbox [165]. Each of these signals is then half-wave rectified and low-pass filtered at a 3 dB cut-off frequency of 30 Hz (as used in [16]) to obtain the modulation envelopes. A cut-off frequency of 30 Hz was chosen as it was found to perform better when compared to other cut-off frequencies (20, 25 and 35 Hz) during the pilot experiments (not included in the thesis). The MS representation is then obtained by taking the magnitude STFT of these envelopes by

keeping the hop size $t_h^{\mathrm{MS}} = t_h^{\mathrm{dft}} = 10$ ms and using a window length $t_w^{\mathrm{MS}} = 64$ ms as in [8]. $K = 64\text{ms} \times f_s$ frequency bins are used to obtain the STFT, where f_s is the sampling frequency. i.e., the frequency resolution is ≈ 15 Hz resulting in approximately 3 frequency bins below 30 Hz cut-off frequency including the DC component. A value of $k = 5$ is chosen to capture the frequency leakage during low-pass filtering and windowing. The MS exemplar is then obtained as detailed in Section 2.3.3.

Notice that the number of channels in the filter bank is the same as the number of Mel filters used in the previous step. This choice is made to have a fair comparison between the performances of the Mel and the MS exemplars in separating speech and noise.

For the experiments on the AURORA-2 database, the parameters used were $B = 128$ and $M = 23$ whereas the AURORA-4 setting used were $B = 256$ and $M = 40$. Zero-padding was used while taking the STFT, whenever necessary. Then three coupled dictionaries each for speech and noise were created with the corresponding exemplars extracted from the same piece of training data.

To create the speech dictionary, $J_s = 10\ 000$ exemplars were extracted at random from the respective clean training data for experiments on the AURORA-2 and AURORA-4 databases as used in [7,8]. Evaluations on the AURORA-2 database used a noise dictionary containing $J_n = 10\,000$ exemplars, whilst for the AURORA-4 experiments, the noise dictionary used is comprised of two parts: a fixed noise dictionary containing 5 000 exemplars extracted from the noise training data and a small noise dictionary extracted from the noisy test data to be enhanced itself, which are the cyclicly shifted versions of its first $T = 15$ frames resulting in a total of $J_n = 5\ 015$ noise exemplars as in [8,58]. Making use of the first 15 frames to model the noise is termed *noise-sniffing* assuming the first 15 frames of the noisy test data contain noise only. Notice that the second noise dictionary is changed for every utterance and is concatenated with the fixed noise dictionary.

Extracting the fixed part of the coupled dictionaries was done only once per database and they are kept fixed for all the experiments. The noise dictionaries for the AURORA-2 database contain exemplars sampled from all the four noise types available in the training data and the fixed noise dictionary for AURORA-4 experiments contain all the six noise types in the training data. No supervision was done to avoid silences in the speech exemplars or adjusting the number of exemplars per noise type in the noise dictionary.

2.4.3 NMF based speech enhancement

For testing, the noisy utterance is converted to the input exemplar space to obtain the observation data matrix $\mathbf{\Psi}^{\text{in}}$ as explained in Section 2.2. $\mathbf{\Psi}^{\text{in}}$ is then decomposed using the respective input dictionary using 600 NMF multiplicative updates (1.16) with $\mathbf{X}^{\text{in}}$ initialised as $(\mathbf{A}^{\text{in}})^{\mathsf{T}}\,\mathbf{\Psi}^{\text{in}}$ and the corresponding filters are obtained as described in Section 2.3. The resulting enhanced STFT is inverted to the time-domain using the overlap-add method to obtain the enhanced speech.

For the AURORA-2 setting, the decomposition was obtained with speech and noise sparsity penalties as $\lambda_s = 1.5$ and $\lambda_n = 1$ for the Mel dictionary as used in [64] whilst for the decomposition using the MS and DFT dictionaries, the values used were $\lambda_s = 1.75$ and $\lambda_n = 0.75$ as in [7]. These values were obtained after doing a grid-search in the range $[0, 3]$ on a development set which is a subset of 100 files taken from the test set A.

For the AURORA-4 experiments, in contrast to the AURORA-2 setting, the noise sparsity penalty is fixed as 0.5 times the sparsity penalty of speech, i.e., $\lambda_n = \lambda_s/2$, to reduce the computational effort while doing the grid-search [58] on the development set. The decomposition using the Mel, MS and DFT settings used a λ_s equal to 1.2, 1.6 and 1.7 respectively.

Speech enhancement was implemented using MATLAB and GPUs were used for accelerating the NMF multiplicative updates using the parallel computing toolbox. To evaluate and compare the speech enhancement qualities, we used signal-to-distortion ratio (SDR), segmental SNR (SegSNR) and PESQ measurements. SDRs were obtained using the BSS evaluation toolkit [181], and the other two measurements were calculated using an implementation by Loizou [118]. The improvement of these quality measures over the noisy speech is reported as ΔSDR in dB, ΔPESQ in mean opinion score (MOS) and ΔSegSNR in dB.

2.4.4 ASR back-ends

1) HMM-GMM decoder for AURORA-2: For evaluating the ASR performance on the AURORA-2 database, a GMM-HMM-based recogniser using the Mel-frequency cepstral coefficients (MFCCs) was used. The HMM topology had a total of 179 states comprised of 16 states describing each digit with 3 states for silence $(16 \times 11 + 3)$. GMM models were trained on MFCCs with 13 static coefficients along with the delta and delta-delta coefficients leading to a 39 dimensional feature space. The emission probabilities of each of the HMM states were modelled using a GMM of 32 Gaussians with diagonal covariance. The decoding is done using the Viterbi decoder with a finite state language model as given in the AURORA-2 benchmark [82] with all digits having the same word entrance penalties.

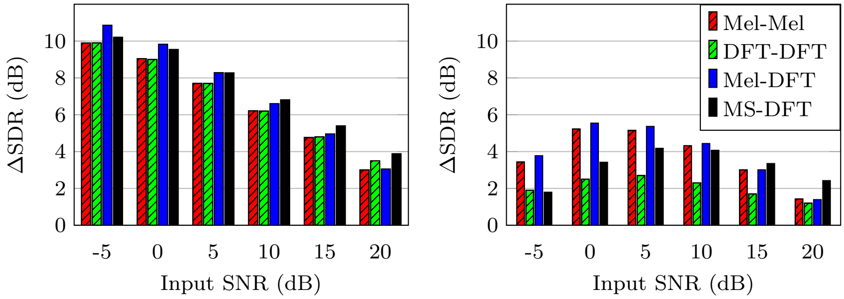
2) Hybrid setting for AURORA-2: Preliminary experiments on the AURORA-2 database revealed a complementarity in the number of insertions and deletions between the MS-DFT and the Mel-DFT systems. So a hybrid approach is proposed to combine the outcomes of these two recognisers to achieve a better ASR performance. There exist several ways to combine results from two systems like assuming independence and then balance the two streams [62], minimum error based approach [54], etc. In order to avoid extra parameters, we propose to combine the two streams by simply multiplying the likelihoods originating from the Mel-DFT and MS-DFT settings [62]. Equal weights are given to both the streams by raising both the resulting likelihoods by 0.5. i.e.,

$$p'(y_t|q_t) = \left(p_{\text{mel}}(y_t|q_t)\right)^{1/2} \left(p_{\text{MS}}(y_t|q_t)\right)^{1/2} \tag{2.3}$$

where, p_{mel} and p_{MS} are the likelihoods for the observation y_t given the HMM state q_t resulting from the Mel-DFT and MS-DFT streams, respectively. These are then fed to the Viterbi decoder to obtain the ASR results.

3) HMM-GMM decoder for AURORA-4: For the AURORA-4 experiments, the "recipe" recognisers in the Kaldi toolkit [143] are used. The HMM-GMM-based recipe decoder for AURORA-4 makes use of context dependent tied-state triphone models. Each model is comprised of three states and there are around 2000 distinct HMM states in total. GMM models are trained on 13 static MFCC features from 7 consecutive frames upon which feature decorrelation is applied using maximum-likelihood linear transform (MLLT) [154] and linear discriminant analysis (LDA) [73], reducing the 91-dimensional vector to 40 dimensions. To compensate for channel variations, cepstral mean and variance normalisation was also applied on the MFCC features.

4) DNN-HMM decoder for AURORA-4: In this work, we also evaluate the ASR performance using the DNN-HMM hybrid system, where the posterior probability estimates for the HMM states are provided by the trained DNNs [81]. DNNs are comprised of multiple hidden layers stacked on top of each other which allow them to learn higher-level information in the upper layers [138]. The recipe recogniser is based on the implementation described in [180] with 6 hidden layers comprised of 2048 sigmoid neurons per layer. The input layer used 40 Mel filterbank coefficients with a context size of 11 frames summing up to 440 input features in total.



(a) Test set A

(b) Test set B

Figure 2.3: Average SDR improvements in dB obtained on test sets A and B of the AURORA-2 database as a function of input SNRs in dB for various settings. Legends are same for both plots.

To train the DNN, pre-training based on restricted Boltzmann machines (RBMs) [80] is done first in order to avoid issues with random initialization of the layers resulting in poor local optima. Once the pre-training is done, a DNN which classifies the frames into triphone states is trained using the stochastic gradient descent technique. Finally, the DNN is trained to classify the whole sentence correctly. For the DNNs trained on clean training data, only the clean part of the development set was used for cross-validation.

Average word error rates (WERs) are used as the performance measure in all the ASR experiments. For training the acoustic models, the original clean training data (referred to as *clean training*) and the enhanced noisy training data processed by the corresponding NMF-based front-ends (referred to as *retraining*) are used. Retraining equips the GMMs and DNNs to learn the artefacts introduced by the enhancement stage and thus can improve the ASR performance on the enhanced noisy test data.

2.5 Pilot experiments on AURORA-2

This section details the speech enhancement and ASR evaluations performed on the AURORA-2 database. The results are reported on the entire test sets including the 100 files used for tuning the sparsity parameters. Some useful insights and discussions are also included in this section.

2.5.1 Results on speech enhancement

ΔSDR in dB averaged over the four noise types obtained for various systems on the AURORA-2 database are summarised in Fig. 2.3. The shaded bars denote the baseline systems and it can be seen that the proposed approach using coupled dictionaries results in better SDRs in all cases. Notice that, even though the Mel-Mel setting uses a pseudo-inverse, it yields almost the same SDRs as of the DFT-DFT setting on test set A. This can be attributed to the better speech and noise separation achieved by the Mel exemplars when compared to the DFT exemplars. It can also be seen that the Mel-DFT setting yields better SDRs than the Mel-Mel setting for both test sets, even though the decomposition in both the systems are done in the Mel exemplar space. It reveals the effectiveness of using the proposed coupled DFT dictionary approach to directly obtain the DFT estimates over the low-rank approximation using pseudo-inverse.

From the SDR evaluations on test set B which contains unseen noise cases, it can be seen that the speech enhancement obtained is poorer when compared to that of test set A, as the noise dictionary generalises poorly to the unseen noise cases. It can also be seen that the Mel feature space is able to better generalise to the unseen noise cases when compared to the DFT and MS exemplar spaces. Using the proposed Mel-DFT approach can further increase the SDR performance, which is a scenario where the proposed approach is highly beneficial. It can also be seen that the MS space can yield a better speech and noise separation at high SNRs when compared to the Mel features.

2.5.2 ASR evaluation

The average WERs obtained on the enhanced AURORA-2 data using the HMM-GMM based decoder and also using the hybrid setting described in Section 2.4.4 are summarised in Table 2.1 for GMMs trained on the clean training data (clean training) and the enhanced noisy training data (retrained). It can be seen that the method using coupled dictionaries yields improved WERs and retraining the GMMs using the enhanced training data can further improve the ASR performance. The Mel-DFT setting resulted only in a slight improvement when compared to the Mel-Mel setting, even though the former setting yielded a better speech enhancement in terms of SDRs. This can be attributed to the simplicity of the AURORA-2 recognition task as it has a limited vocabulary, and the digit classification is not affected by the deformation introduced during the pseudo-inverse step.

Table 2.1: Average WERs in % obtained for test sets A and B of the AURORA-2 database for various settings with GMMs trained on clean and enhanced noisy training data. Shaded rows denote the baseline settings. Best scores are highlighted in bold font.

		Test Set A							**Test Set B**						
Setting	**clean**	**-5**	**0**	**5**	**10**	**15**	**20**	**Avg. (20-0)**	**-5**	**0**	**5**	**10**	**15**	**20**	**Avg. (20-0)**
								GMM on clean training data							
No Enh.	0.3	76.9	48.7	22.4	9.2	3.6	1.6	17.1	77.2	46.9	20.7	7.7	2.8	1.2	15.9
Mel-Mel	0.4	31.2	12.4	6.1	3.6	2.3	1.4	5.2	58.2	30.3	**12.4**	5.8	2.7	0.9	10.4
DFT-DFT	**0.3**	34.7	17.5	7.8	3.1	1.7	0.9	6.2	70.8	40.1	16.9	6.1	2.3	1.0	13.3
Mel-DFT	0.4	31.1	12.4	6.0	3.5	2.1	1.2	5.0	**58.0**	**30.1**	**12.4**	5.7	2.7	0.8	**10.3**
MS-DFT	**0.3**	30.5	12.5	4.4	**2.1**	**1.3**	**0.7**	4.2	68.6	34.3	14.5	**5.1**	2.1	0.8	11.4
Hybrid	0.4	**27.2**	**11.4**	**3.7**	**2.1**	1.5	0.9	**3.9**	62.4	32.8	13.0	5.3	**2.0**	**0.6**	10.7
								GMM on noisy training data (Retrained)							
No Enh.	0.8	61.9	24.9	6.8	2.6	1.2	0.7	7.2	64.3	26.2	8.5	2.9	1.4	0.8	8.0
Mel-Mel	0.5	25.1	8.9	3.3	1.5	0.9	1.0	3.1	52.8	20.8	6.8	2.6	1.2	0.7	6.4
DFT-DFT	**0.4**	21.4	8.5	2.5	1.1	**0.7**	0.5	2.7	58.1	24.5	7.5	2.4	**1.0**	0.6	7.2
Mel-DFT	0.5	25.2	9.0	3.1	1.4	**0.7**	0.6	3.0	**52.6**	21.0	6.8	2.7	1.2	0.6	6.4
MS-DFT	**0.4**	21.1	7.7	**2.4**	1.0	**0.7**	**0.4**	**2.4**	62.4	26.3	7.6	2.2	**1.0**	**0.5**	7.5
Hybrid	**0.4**	**20.6**	**7.1**	**2.4**	**1.0**	**0.7**	0.7	**2.4**	54.2	**20.7**	**6.3**	**2.1**	**1.0**	**0.5**	**6.1**

It is also observed that the use of the MS representation can result in a WER improvement for test set A and poorer results for test set B as it generalises poorly for unseen noise cases. Nevertheless, it yielded complementary results in terms of insertions and deletions when compared to the Mel setting and the proposed hybrid setting was found to yield superior WER improvements on both test sets by exploiting this complementarity. To the best of our knowledge, average WERs of 20.6% (test A, SNR-5), 2.4% (test A, SNR(20-0)) and 6.1% (test B, SNR(20-0)) using the hybrid setting are among the best results ever reported on the AURORA-2 recognition task (reported in [62]). Overall, from SNR -5 dB to 20 dB, the hybrid setting yielded WERs of 5.4% and 14.1% on test set A and B, respectively.

Also notice that the method described in [64] directly makes use of enhanced Mel features for the ASR back-end rather than going back to the time-domain. Evaluations (not shown) revealed that this setting and the Mel-Mel setting are equivalent as the ASR back-end for the latter also goes back to the Mel domain by multiplication with the same Mel matrix $\mathbf{M}$ to obtain the MFCCs.

2.5.3 A qualitative analysis

A qualitative analysis on the observations made during the pilot experiments on the AURORA-2 database is discussed in this section. The outcomes of interest resulting from these evaluations are visualised in Fig. 2.4. The input noisy signal is an arbitrary signal from the AURORA-2 database containing the utterance "nine six zero" (transcribed as *96Z*) corrupted with babble noise at an SNR of 0 dB. The filter weights used for enhancing the noisy STFT arising from the various settings are shown in the middle row followed by the resulting enhanced speech in the bottom row. For comparison, the oracle binary mask is also included which yielded an output SDR of 10.9 dB. It is evident that the quality of enhanced speech depends on how well these filter weights model the constituent speech and noise contained in the noisy speech. The key aspects which decide the performance of various settings are detailed below (ref. Fig. 2.4):

1) Low-rank approximation in the Mel-Mel setting: It can be seen that the piece-wise linear approximation results in a set of filter weights that are smooth which in turn cannot model the underlying harmonic structure of the constituent speech signal and results in frequency smearing. This setting thus will always result in a sub-optimal set of filter weights. Also notice that this setting still yielded a reasonable SDR improvement which can be attributed to a better speech and noise separation achieved using the Mel exemplars.

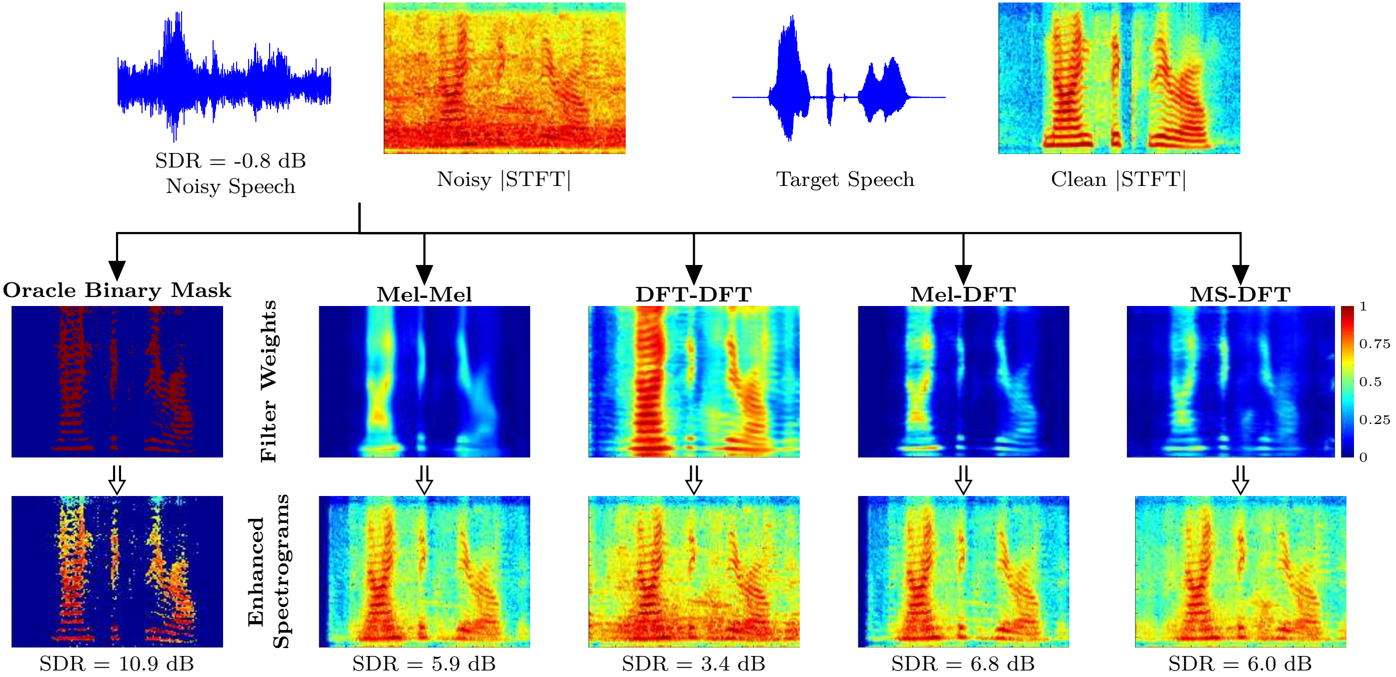


Figure 2.4: Comparison of filter weights with oracle binary mask and the resulting enhanced speech spectrograms obtained for various settings for an arbitrary noisy speech signal from the AURORA-2 database corrupted with babble noise at an SNR of 0 dB. Log spectrograms are shown for a better visualisation. All filter weight plots used a linear color mapping in $[0, 1]$. The resulting SDRs are also shown below each of the enhanced spectrograms.

2) Poorer speech and noise separation in the DFT exemplar space: It can be seen that the filter weights arising from the DFT-DFT setting are able to model the underlying harmonic structure of speech since this setting can directly obtain the estimates in the full-resolution frequency domain. However, a majority of these weights are close to 1 even though the true SNR of the underlying speech is 0 dB, which in turn retain most of the noise content and results in poorer SDRs. Also notice that the noise in the speech inactive regions are not properly suppressed. These happen because the speech exemplars are also activated to model the babble noise contained in the noisy input during the exemplar-based decomposition in the DFT space. Similar instances of speech exemplars modelling noise are observed for unseen noise cases also (not shown) [7]. This setting hence results in a poorer SDR improvement even though the detrimental mapping stage is absent.

3) Full-rank approximation in the Mel-DFT setting: The filter weights obtained for the Mel-Mel and Mel-DFT settings arise from the same set of activations obtained from the NMF-based decomposition in the Mel exemplar space. It can be seen that the Mel-DFT approach is able to better model the harmonic structure in speech and utilise the better speech and noise separation properties of the Mel exemplar space, yielding an SDR improvement of 0.9 dB over the setting where the pseudo-inverse is used. This approach thus can yield a better speech enhancement without any additional computational cost in the matrix factorisation part, which is the most time-consuming part of the method.

4) Coupled dictionaries as a reliable mapping from the MS space to the DFT/time domain: It is evident from the filter weights obtained for the MS-DFT setting that the MS exemplars can yield a good speech and noise separation, and using the coupled DFT dictionary can yield a reliable mapping of these estimates to the full-resolution frequency domain.

2.5.4 Computational complexity vs performance

All the evaluated experiments in this work were accelerated using GPUs. The computational complexity of these experiments depends on the length of the temporal context T, the number of exemplars $(J_s + J_n)$ and the dimension of features per frame considered. The average execution time needed for the experiments on the AURORA-2 database, which used 10 000 exemplars each of speech and noise with $T = 30$ frames for various settings are tabulated in Table 2.2. From the evaluations, it is clear that the proposed Mel-DFT setting results in a good ASR and SDR performance without much additional computational cost.

Table 2.2: Average execution time in seconds needed for various settings evaluated on the AURORA-2 database. D is the number of rows in the dictionary used to obtain the NMF-based decomposition. All dictionaries had a total of $20,000$ columns each.

	Mel-Mel	*DFT-DFT*	*Mel-DFT*	*MS-DFT*
Exec.time	5.8s	16.2s	6.0s	14.8s
D	690	3840	690	3450

It is also observed in [8] that increasing the low-pass 3 dB cut-off frequency in the MS exemplar extraction stage can yield an improvement both in terms of SDRs and WERs, in presence of seen noise cases. However, this can have a detrimental effect for signals corrupted with unseen noise and also results in an increased computational complexity as the size of the MS exemplars should also be increased.

Similar to the MS features, the performance of the DFT exemplars depends on the type of noise and the true SNRs in the input noisy signal, and its computational complexity depends solely on the sampling frequency of the input data, given T and the window length t_w^{dft} used to obtain the STFT. On the other hand, the Mel and MS features are more flexible in the sense that their dimensionality can be adjusted by varying choices for M, t_w^{MS} etc., depending on the application and allowable computational complexity.

2.6 Experiments on AURORA-4 database

2.6.1 Results on speech enhancement

ΔSDR, ΔPESQ and ΔSegSNR averaged per test set obtained for the various settings on the AURORA-4 database are presented in Fig. 2.5. As an additional baseline system, a speech enhancement algorithm based on minimum mean-square error log-spectral amplitude estimation [50] with the improved minima controlled recursive averaging (IMCRA) technique for noise variance estimation [40] is included.

It can be seen that the proposed approach using coupled dictionaries results in better SDRs in all cases, consistent with the observations made during the AURORA-2 experiments. It can also be seen that additional evaluations using the PESQ and SegSNR also yielded promising improvements. IMCRA approach yielded better SegSNR for some noise types, but poorer PESQ and

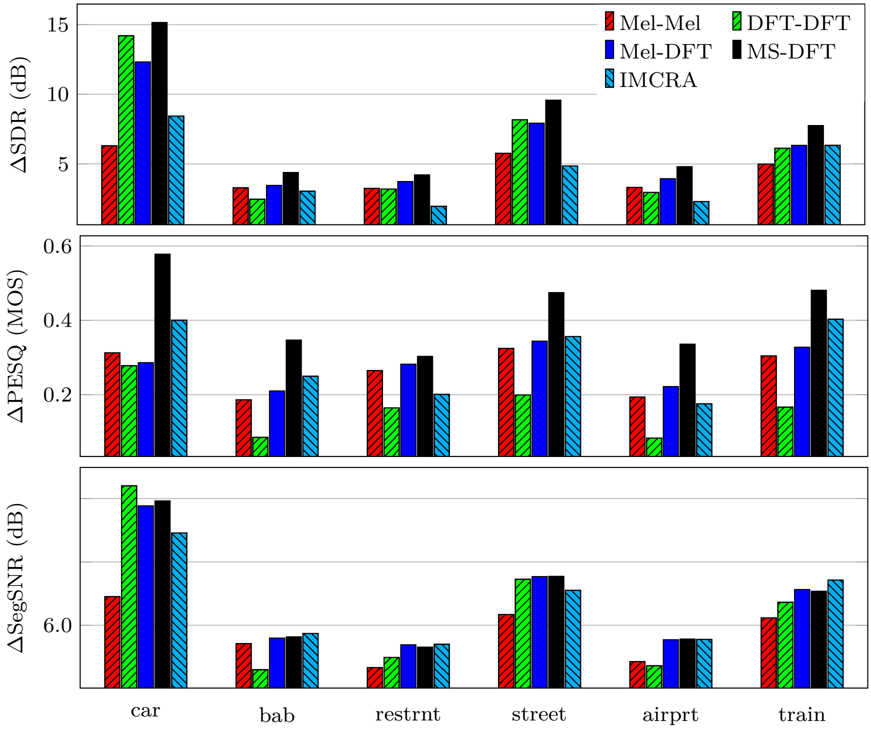


Figure 2.5: Average improvements in speech enhancement performance in terms of ΔSDR, ΔPESQ and ΔSegSNR obtained for each test set on the AURORA-4 database for various settings. From left to right, these noises correspond to test02-07 (car, babble, restaurant, street airport and train noises, respectively). The legends are same for all plots.

SDR improvements were obtained. The MS-DFT setting yielded superior improvements in PESQ MOS evaluation reaffirming the effectiveness of using coupled dictionaries to obtain a reliable reconstruction in the DFT space.

2.6.2 ASR evaluation

The average WERs obtained for the HMM-GMM-based and HMM-DNN-based decoders on various test sets of the NMF-enhanced AURORA-4 data are tabulated in Table 2.3. The results for the retrained scenarios only are presented for both the GMM and DNN based settings.

For acoustic modelling based on retrained GMMs, it can be seen that the various speech enhancement approaches can greatly improve the ASR

Table 2.3: Average WERs obtained in % for various test sets on the AURORA-4 data using the various settings with the HMM-GMM-based and HMM-DNN-based ASR back-ends. Best scores are highlighted in bold font. Shaded rows denote the baseline systems.

(a) Retrained GMM

	Test Sets							
Setting	**A**	**B**						
	01	**02**	**03**	**04**	**05**	**06**	**07**	**Avg.**
No Enh.	5.7	6.2	11.5	22.3	16.7	10.9	15.8	13.9
Mel-Mel	5.1	5.6	8.4	10.6	9.8	8.1	10.1	8.8
DFT-DFT	6.0	5.8	8.9	12.2	10.3	8.7	11.2	9.5
Mel-DFT	4.9	**5.4**	8.0	**10.7**	9.8	7.7	10.2	8.6
MS-DFT	4.9	5.7	**7.3**	11.1	**9.0**	**7.0**	**10.1**	**8.4**
IMCRA	**4.6**	5.6	10.7	15.3	13.8	11.4	14.4	11.9

(b) Retrained DNN

	Test Sets							
Setting	**A**	**B**						
	01	**02**	**03**	**04**	**05**	**06**	**07**	**Avg.**
No Enh.	3.3	4.6	7.3	9.3	8.5	6.6	9.1	7.7
Mel-Mel	**2.9**	**4.1**	6.6	8.8	8.9	6.1	9.2	7.3
DFT-DFT	3.2	**4.1**	6.9	7.8	7.6	6.5	8.0	6.8
Mel-DFT	3.2	4.7	7.5	8.5	8.4	6.9	8.2	7.4
MS-DFT	3.0	4.2	**6.0**	**7.4**	**7.1**	**5.3**	**6.9**	**6.2**
IMCRA	**2.9**	**4.1**	7.2	9.4	9.6	7.6	9.0	7.8

performance over a GMM trained and evaluated on noisy test data. IMCRA yields the best performance on clean speech as it introduces the least distortions on clean speech during speech enhancement. It can also be seen that the MS-DFT setting yields the best performance out of all the evaluated settings with a statistical significance of $p < 0.03$ (over a total of 32 118 words using a binomial independence assumption).

On the other hand, a DNN trained on noisy training data yields around 40% relative improvement over the GMM-based system and is even better than the best performing retrained GMM setting (ref. Table 2.3a), thanks to its multiple hidden layers which can learn and compensate for the noise also. It can be seen that using exemplar-based approaches for speech enhancement and retraining can further improve its performance (ref. Table 2.3b). Also notice that all settings yielded a better WER for clean speech as well, which can be attributed to the ability of sparse representations in moving the test features closer to the training features, thereby minimizing the speaker mismatches in the training and test sets as pointed out in [150].

The MS-DFT setting yielded the best WERs here as well with a statistical significance of $p < 0.001$ over all the other settings yielding an average WER of 6.2% over test B of the AURORA-4 database.

2.7 Contribution to the CHiME-3 challenge

This section investigates the combination of the proposed NMF-based speech enhancement technique using coupled dictionaries with the various ASR backends for evaluation of the CHiME-3 challenge [14]. The CHiME-3 challenge targets the performance of an ASR setting in real world, commercially motivated scenario where the recordings are obtained using a tablet fitted with a six-channel microphone array (more details can be found in Section 1.4.1). In our framework, the six-channel data is converted to a single-channel signal using a beamformer, enhanced using the NMF-based technique and fed to different ASR decoders that use GMM, DNN and CNN-DNN -based acoustic modelling.

This section also presents an extension to the technique using coupled dictionaries by adding adaptive dictionaries that are learned online from the test data. In addition to the fixed speech and noise dictionaries, we add an adaptive dictionary to model unseen noise which is also learned together with the activations from the test data. The coupled atoms in the STFT space for the adaptive dictionary are then learned using the activations (see Section 2.7.1). Such a setting is particularly useful when the noise in the test utterance is not present in the training data, which is often the case in real applications.

2.7.1 Extension using the adaptive dictionary learning

In this section, we extend the approach using coupled dictionaries by adding an adaptive noise dictionary to the existing fixed speech and noise dictionaries. Adaptive noise dictionaries have been effectively used to model unseen noise for the single dictionary cases [87]. In the coupled dictionary framework, the input dictionary will be $\mathbf{A}^{\text{in}} = [\mathbf{S}^{\text{in}}\ \mathbf{N}^{\text{in}}\ \mathbf{N}^{\star\text{in}}]$, where the superscript $\star$ denotes the adaptive part which will also be estimated from the test data as:

$$\mathbf{\Psi}^{\text{in}} \approx \left[\mathbf{S}^{\text{in}} \;\vdots\; \mathbf{N}^{\text{in}} \;\vdots\; \mathbf{N}^{\star\text{in}}\right] \begin{bmatrix} \mathbf{X}^{\text{in}}_{\mathbf{s}} \\ \cdots \\ \mathbf{X}^{\text{in}}_{\mathbf{n}} \\ \cdots \\ \mathbf{X}^{\star\text{in}}_{\mathbf{n}} \end{bmatrix} = \mathbf{A}^{\text{in}}\mathbf{X}^{\text{in}} \tag{2.4}$$

Both $\mathbf{X}^{\text{in}}$ and $\mathbf{N}^{\star\text{in}}$ are estimated by applying alternating multiplicative updates to minimize the KLD between $\mathbf{\Psi}^{\text{in}}$ and $\mathbf{A}^{\text{in}}\mathbf{X}^{\text{in}}$ given in [109], with the adaptive dictionary initialised randomly and activations initialised as $\mathbf{A}^{\text{in}\mathsf{T}}\Psi^{\text{in}}$, where T denotes matrix transpose. Once the activations $\mathbf{X}^{\text{in}}$ are obtained, the coupled adaptive dictionary in the STFT space is obtained such that :

$$\mathbf{\Psi}^{\text{dft}} \approx \left[\mathbf{S}^{\text{dft}} \;\vdots\; \mathbf{N}^{\text{dft}} \;\vdots\; \mathbf{N}^{\star\text{dft}}\right] \mathbf{X}^{\text{in}} \tag{2.5}$$

The multiplicative updates (same as used above from [109]) are applied only on $\mathbf{N}^{\star\text{dft}}$ keeping everything else fixed. The time-varying filter is obtained in the same manner as in Section 2.2 with $\mathbf{A}^{\text{dft}} = [\mathbf{S}^{\text{dft}}\ \mathbf{N}^{\text{dft}}\ \mathbf{N}^{\star\text{dft}}]$. Notice that, this approach is reliable only when the learned atoms model only noise and not speech. Therefore, we use only a limited number of adaptive atoms which is dependent on the length of the utterance and assume that the fixed speech dictionary is sufficient to model speech and the adaptive atoms only model the unseen noise.

2.7.2 Evaluation setup for CHiME-3

The decoders used are based on the Kaldi toolkit [143] scripts provided by the CHiME-3 challenge organisers [14] for GMM, DNN and DNN+sMBR -based evaluations. We used the training data enhanced by the respective front ends to train all the models. In addition, we also include a CNN-based decoder containing 2 convolutive hidden layers and 2 DNN-layers at input followed by 4 fully connected DNN layers as presented in[1] [149].

The GMMs used are trained on MFCC features after applying LDA, MLLT and fMLLR transforms [14]. DNNs are trained on 40 log-Mel features with a temporal context of 5 frames on either side of the central frame. The DNN contains six hidden layers with 2048 sigmoid neurons per layer. For training the CNN system, the input features fed to the 2 CNN layers are 40 log-Mel

[1] The Kaldi CNN-DNN recipe used for the CHiME-3 challenge evaluation is available at `https://github.com/deepakbaby/chime3cnn`

features together with their delta and delta-delta as used in [149] and the 2 DNN layers at the input are fed with the pitch features along with the delta and delta-delta coefficients. The output of the CNN and the DNN are fed to 4 DNN layers.

The speech and noise dictionaries used by the NMF-based speech enhancement setting were created using the clean WSJ0 utterances and background noise recordings respectively, that are provided with the CHiME-3 dataset. The coupled speech and noise dictionaries contained 10 000 and 5 000 exemplars respectively extracted by random sampling. For the adaptive dictionary part, we choose the number of adaptive atoms as $\lceil \alpha \cdot T_f / T \rceil$, where T_f is the number of frames in the test utterance and $0 < \alpha < 1$. We chose $\alpha = 0.2$ as it was found to yield reasonably good enhancement on a few utterances in the development set. The fixed part of all the coupled dictionaries are created only once and are kept fixed for all the evaluations in this paper.

To compare various settings, the evaluations using the output of the beamformer (*enhanced* in the CHiME-3 dataset) is considered as the baseline setting.

2.7.3 Results without the adaptive dictionaries

The average WERs obtained in % for various speech enhancement approaches without adaptive atoms are given in Table 2.4. It can be seen that the CNN-based decoders significantly outperform the DNN-based settings with an absolute WER reduction of 3.3% on real development data (DNN+sMBR vs. CNN-DNN+sMBR) for the baseline system.

The results show that the NMF-based approaches can improve the performance of all the investigated ASR back-ends. Among the evaluated speech enhancement front-ends, Mel-DFT setting performs the best in most of the cases on the development data. Most of the enhancement approaches yield only minor improvements on simulated data, thanks to a larger amount of simulated training data over the real training utterances.

Also notice that the DFT-DFT setting performs better with the test set. It can also be seen that the MS features perform better with the simulated data than with the real data. This is because the modulation envelopes are sensitive to the reverberant or noisy environments and hence it fails to model the highly non-stationary and unseen noise in the real data.

	Baseline		**Mel-Mel**		**Mel-DFT**		**DFT-DFT**		**MS-DFT**	
AM	**REAL**	**SIM**	**REAL**	**SIM**	**REAL**	**SIM**	**REAL**	**SIM**	**REAL**	**SIM**
					Development Set					
GMM	20.9	10.0	**17.9**	**8.8**	**17.9**	**8.8**	18.4	8.9	18.7	9.1
DNN	20.4	**9.3**	20.2	9.4	20.3	9.5	**19.8**	9.8	20.6	**9.3**
+sMBR	17.7	8.4	17.7	**8.3**	**17.5**	**8.3**	17.6	8.6	**17.5**	**8.3**
CNN-DNN	16.1	7.1	15.6	6.7	**15.0**	6.8	15.2	**6.6**	15.6	6.8
+sMBR	14.4	6.2	14.0	6.2	**13.7**	6.2	**13.7**	6.1	14.5	**6.0**
					Test Set					
GMM	37.7	11.1	31.2	**9.6**	**30.8**	9.9	30.9	**9.6**	33.6	**9.6**
DNN	41.9	12.5	40.0	12.6	40.6	12.7	**37.9**	12.8	40.8	**11.7**
+sMBR	34.5	10.6	32.7	11.1	32.9	11.3	**32.4**	11.0	33.2	**10.3**
CNN-DNN	29.7	7.4	27.4	7.7	26.4	**7.6**	**25.8**	**7.6**	28.2	7.7
+sMBR	27.0	6.9	**24.3**	7.0	24.4	6.9	24.9	**6.7**	27.2	6.9

Table 2.4: WERs in % obtained for various ASR back-ends with speech enhancement without the adaptive dictionaries. The best result for each ASR back-end is highlighted in bold font.

	Baseline		**Mel-Mel**		**Mel-DFT**		**DFT-DFT**		**MS-DFT**	
AM	**REAL**	**SIM**	**REAL**	**SIM**	**REAL**	**SIM**	**REAL**	**SIM**	**REAL**	**SIM**
Development Set										
GMM	20.9	10.0	17.3	**8.8**	**16.9**	8.9	17.5	8.9	18.2	8.9
DNN	20.4	9.3	18.8	9.3	18.8	**9.0**	**18.2**	9.1	19.7	9.5
+sMBR	17.7	8.4	16.7	8.4	16.7	**8.1**	**16.5**	8.2	17.2	8.4
CNN-DNN	16.1	7.1	**14.3**	7.0	14.7	7.2	**14.3**	**6.8**	15.2	**6.8**
+sMBR	14.4	6.2	13.2	**6.1**	**12.9**	6.4	13.2	6.2	13.6	6.2
Test Set										
GMM	37.7	11.1	30.2	10.0	**29.6**	9.8	29.7	**9.5**	32.9	9.6
DNN	41.9	12.5	35.4	11.8	36.2	11.3	**34.0**	**10.5**	39.5	11.9
+sMBR	34.5	10.6	30.8	10.5	31.8	10.5	**28.9**	**9.0**	32.8	10.4
CNN-DNN	29.7	7.4	25.7	7.5	25.5	7.8	**24.0**	7.5	28.3	**7.3**
+sMBR	27.0	6.9	24.6	7.1	23.1	6.9	**22.8**	7.0	25.3	**6.6**

Table 2.5: WERs in % obtained for various ASR back-ends with speech enhancement with the adaptive dictionaries. The best result for each ASR back-end is highlighted in bold font.

2.7.4 Results with the adaptive dictionaries

The average WERs obtained in % for various speech enhancement approaches when adaptive atoms are used are tabulated in Table 2.5. It can be seen that the inclusion of adaptive atoms and learning the coupled atoms online can yield significant WER reductions over the fixed dictionary scenarios. Notice that the improvements are mostly obtained on real data and the Mel-DFT setting with adaptive atoms yielded an absolute WER reduction of 0.8% (6% relative) for a state-of-the-art decoder using CNN-DNN+sMBR on real development data, when compared to the same setting without adaptive atoms.

In short, the experiments reaffirm the effectiveness of the NMF-based speech enhancement using coupled dictionaries as front-end for state-of-the-art ASR back-ends and we show that the extension using adaptive atoms can be particularly useful when processing recordings taken from real-world scenarios.

2.7.5 Best results obtained

	Dev. Set		Test Set	
Environment	**REAL**	**SIM**	**REAL**	**SIM**
BUS	15.52	5.88	28.38	5.83
CAF	12.52	8.04	26.45	7.32
PED	11.59	5.32	22.29	6.76
STR	11.84	6.24	15.07	7.84
Avg.	12.87	6.37	23.05	6.94

Table 2.6: Detailed results obtained for the best ASR setting on CHiME-3 challenge data.

In this work, the best result obtained on the real development set is for the setting which uses the Mel-DFT speech enhancement front-end with adaptive atoms with a CNN-DNN+sMBR -based ASR back-end. The detailed results are given in Table 2.6. These results are obtained for a language model weight of 11 and 6 sMBR iterations. We report average WERs of 23.05% and 6.94% on the real and simulated test sets, respectively.

On the real test data, the best setting yields an absolute WER improvement of 4% (15% relative) over the baseline setting which uses the beamformer output. It is also interesting to notice that most of the speech enhancement front-ends yield only slight improvements on the simulated data, thanks to a better acoustic modelling using CNN and a larger amount of simulated training data over the real training data.

2.8 Conclusions

In this work, we proposed using coupled DFT dictionaries, extracted jointly with the input dictionaries used in the exemplar-based speech enhancement systems, for a better mapping from the input space to the DFT space to obtain a better set of filter weights. The approach was found to be effective in overcoming the low-rank approximation where the input dictionary is created using lower-dimensional Mel features and also to obtain a reliable mapping from the MS space to the DFT space. The simulation results revealed that the proposed approach can improve the performance of exemplar-based techniques for both speech enhancement and automatic speech recognition tasks.

The use of modulation spectrogram features, which are inspired from the human auditory processing, was also introduced to the field of exemplar-based techniques in this work, and we showed that using coupled dictionaries can be a reliable way to reconstruct the underlying speech and noise estimates in the DFT domain. The ASR evaluation also revealed that feeding NMF-enhanced data can greatly benefit both the HMM-GMM-based and DNN-HMM-based state-of-the-art ASR systems with and without retraining.

The best performing settings in this work yielded overall average WERs of 5.4% and 14.1% respectively for test sets A and B of the AURORA-2 database, and 7.9% and 5.7% respectively for the GMM-HMM-based and DNN-HMM-based ASR systems on the single microphone sets (test01-test07) in the AURORA-4 database.

In addition, we evaluated the combination of various NMF-based speech enhancement front-ends and ASR back-ends for evaluation of the CHiME-3 challenge. We also introduced an extension to the existing work by adding an adaptive dictionary, the atoms of which are learned online. The evaluations reveal that the speech enhancement approaches together with adaptive atoms can yield significant performance improvements for all the ASR back-ends including the CNN-DNN-based system. The best WER on the real CHiME-3 test data obtained in this work is 23.05%. It is also observed that, in a realistic scenario where the training data available from real conditions is fewer, the NMF-based speech enhancement using coupled dictionaries together with adaptive atoms can be effectively used to mitigate the mismatches between the training and the real test data.

Chapter 3

Hybrid Input Spaces for Exemplar-based Feature Enhancement

This chapter extends the use of coupled dictionaries by introducing hybrid input spaces that are chosen for a more effective separation of speech from background noise. This work investigates the use of two different hybrid input spaces which are formed by incorporating the full-resolution and modulation envelope spectral representations with the Mel features. A coupled output dictionary containing Mel exemplars, which are jointly extracted with the hybrid space exemplars, is used to reconstruct the enhanced Mel features for the ASR back-end. When compared to the system which uses Mel features only as input exemplars, these hybrid input spaces are found to yield improved word error rates on the AURORA-2 database especially with unseen noise cases.

This chapter is adapted from: Deepak Baby and Hugo Van hamme. *Hybrid Input Spaces for Exemplar-based Noise Robust Speech Recognition using Coupled Dictionaries.* 23rd European Signal Processing Conference (EUSIPCO), pp. 1676-1680, September 2015.

3.1 Introduction

Spectral factorization methods based on NMF attempt to decompose the features extracted from a noisy recording as the weighted sum of speech and noise dictionary atoms or exemplars, and are found to be useful for noise-robust ASR [60,191,196]. Most of the conventional exemplar-based ASR systems use exemplars extracted from feature spaces like the Mel [64], Gabor [89], DFT (refers to the magnitude of the discrete-Fourier transform) [186] etc., to obtain the compositional model and enhance the corresponding features. These enhanced features are then used to find the enhanced Mel-frequency cepstral coefficients (MFCCs) to be fed to the ASR back-end.

The efficiency of an exemplar-based NMF approach depends on the ability of the chosen exemplar space in differentiating features originating from speech and noise, and it is found that different exemplar spaces yield different performance depending on the type of added noise and signal-to-noise ratio (SNR) levels [7]. It is also noticed that, apart from increasing the computational complexity, using higher dimensional exemplars derived from feature spaces like the DFT [186], or modulation envelope spectra (MS) [7,8], etc. will result in too detailed modelling of the seen noise cases to generalise well for the unseen noise cases.

In order to address the issues faced by the higher dimensional features and to combine the speech and noise separation properties of different feature spaces, we propose the use of hybrid input spaces to obtain the decomposition. To reconstruct the Mel estimates from this, a variant of the coupled dictionary approach described in [7] is used. In this setup, the exemplars for the coupled hybrid input and the Mel output dictionaries are extracted from the same piece of training data. Then for evaluation, the underlying Mel features are reconstructed using the coupled Mel dictionary, following the decomposition in the hybrid input space.

To obtain a hybrid input space, two feature spaces are chosen first which are called as *primary* and *secondary* feature spaces. A hybrid exemplar is then obtained by concatenating the exemplars belonging to these feature spaces that are extracted from the same piece of training data. In this work, the Mel space is chosen as the primary feature space for its reduced dimensionality and good separation capabilities [7,63] with the DFT or MS representation as the secondary feature space.

To address the "curse" of large dimensionality of the chosen secondary spaces, we propose to use a *trimmed* secondary exemplar space to be concatenated with the full length primary space exemplar. The trimmed exemplar is obtained by reshaping only a subset of the feature frames belonging to the secondary feature space. The decomposition obtained with such a hybrid space will thus rely mainly on the primary feature space with the trimmed secondary space acting as a cue to regularise the separation.

The simulation results obtained on the AURORA-2 database revealed that, even with the secondary space trimmed down to a single frame, both the hybrid input spaces yield improved performances in terms of word error rate (WER) over the baseline system which uses Mel features only. The computational complexity of the proposed approach is also found to be comparable to that of the baseline system as trimmed secondary spaces are used.

3.2 Method

3.2.1 Feature enhancement using NMF

NMF-based compositional models attempt to decompose the features extracted from a noisy recording as a sparse non-negative weighted sum of speech and noise atoms or exemplars stored as columns in a speech and noise dictionary denoted as $\mathbf{S}$ and $\mathbf{N}$, respectively. Exemplars are extracted from training data spanning multiple, say T, frames to capture temporal dynamics, followed by reshaping to form a vector. The representation for the noisy utterance in the exemplar space, $\boldsymbol{\Psi}$, the columns of which are obtained by reshaping sliding windows of length T frames along the length of the utterance [63], is decomposed to get the activations, $\mathbf{X}$, as:

$$\boldsymbol{\Psi} \approx \begin{bmatrix}\mathbf{S} & \mathbf{N}\end{bmatrix} \begin{bmatrix}\mathbf{X}_s \\ \mathbf{X}_n\end{bmatrix} = \mathbf{AX} \quad s.t. \quad \mathbf{X} \geq 0. \tag{3.1}$$

The approximation is done such that it minimises the cost function,

$$\mathcal{C} = D_{KLD}(\boldsymbol{\Psi} \| \mathbf{AX}) + \boldsymbol{\Lambda} \odot \mathbf{X} \tag{3.2}$$

where, D_{KLD} is the element-wise Kullback-Leibler divergence

$$D_{KLD}(x\|y) = x\log(x/y) - x + y \tag{3.3}$$

and $\boldsymbol{\Lambda}$ is the sparsity penalty on the activations $\mathbf{X}$ [64]. $\odot$ denotes element-wise multiplication. The frame-wise speech and noise estimates, $\hat{\mathbf{s}}$ and $\hat{\mathbf{n}}$ are then obtained after removing the windowing effect by adding the frames belonging to the overlapping windows in the windowed estimates $\mathbf{SX}_s$ and $\mathbf{NX}_n$, respectively. A frame-level Wiener-like filter is then obtained after

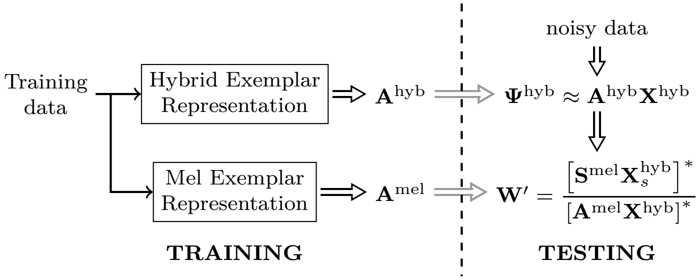

Training data
Hybrid Exemplar Representation
Mel Exemplar Representation
$\mathbf{A}^{\text{hyb}}$
$\mathbf{A}^{\text{mel}}$
noisy data
$\mathbf{\Psi}^{\text{hyb}} \approx \mathbf{A}^{\text{hyb}}\mathbf{X}^{\text{hyb}}$
$\mathbf{W}' = \frac{\left[\mathbf{S}^{\text{mel}}\mathbf{X}_s^{\text{hyb}}\right]^*}{\left[\mathbf{A}^{\text{mel}}\mathbf{X}^{\text{hyb}}\right]^*}$
TRAINING
TESTING

Figure 3.1: Block diagram overview of the proposed system using hybrid input spaces and coupled dictionaries for Mel feature enhancement.

element-wise division as, $\mathbf{G} = \hat{\mathbf{y}} \oslash (\hat{\mathbf{y}} + \hat{\mathbf{w}})$, which when applied to the noisy features yields enhanced features.

3.2.2 Proposed method using hybrid input spaces

In the proposed approach, the activations $\mathbf{X}^{\text{hyb}}$ are obtained using the dictionary $\mathbf{A}^{\text{hyb}} = \left[\mathbf{S}^{\text{hyb}}\ \mathbf{N}^{\text{hyb}}\right]$, which contains exemplars belonging to a hybrid input space, using the NMF approach explained in Section 3.2.1. The windowed Mel speech and noise estimates are then reconstructed using the coupled Mel dictionary, which contains coupled exemplars belonging to the Mel feature space, as $\mathbf{S}^{\text{mel}}\mathbf{X}_s^{\text{hyb}}$ and $\mathbf{N}^{\text{mel}}\mathbf{X}_n^{\text{hyb}}$, respectively. Notice that the corresponding atoms in the coupled dictionaries, $\mathbf{A}^{\text{hyb}}$ and $\mathbf{A}^{\text{mel}}$, are extracted from the same piece of training data which guarantees a reliable reconstruction of the underlying speech and noise estimates in the Mel domain [7,130].

The proposed approach is summarised in Figure 3.1. The notations used to explain the test phase are: $\mathbf{\Psi}^{\text{hyb}}$ for the noisy speech represented in the hybrid exemplar domain and $[\mathbf{Y}]^*$ denotes the matrix obtained after removing the effect of overlapping windows in the windowed observation $\mathbf{Y}$. All matrix divisions should be considered element-wise.

To obtain the hybrid input exemplars, the primary and secondary exemplars are created first from the same piece of training data spanning T frames. Let $\mathcal{T}_S$ be the trimming operator which trims an exemplar spanning T frames down to an exemplar spanning a subset $S \subseteq \{1, 2, \ldots, T\}$ of the T frames. Thus, from an exemplar with frames indexed from 1 through T, the trimming operator $\mathcal{T}_S$ selects only the frames with index contained in S, and reshapes them into a vector.

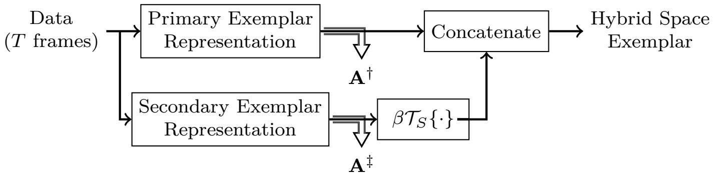

Data
(T frames)
Primary Exemplar Representation
Secondary Exemplar Representation
$\mathbf{A}^{\dagger}$
$\mathbf{A}^{\ddagger}$
$\beta\mathcal{T}_S\{\cdot\}$
Concatenate
Hybrid Space Exemplar

Figure 3.2: Block diagram overview of the processing chain used to obtain the proposed hybrid exemplar representation.

The trimmed secondary exemplars are obtained by applying $\mathcal{T}_S$ on the secondary exemplars, which are also scaled with β to balance its contribution on obtaining the separation. This trimmed and scaled secondary exemplar is then concatenated with the corresponding primary exemplar to get the hybrid representation. Thus, the hybrid exemplar representation for noisy speech $\mathbf{\Psi}^{\text{hyb}}$ and the hybrid dictionary can be expressed as:

$$\mathbf{\Psi}^{\text{hyb}} = \begin{bmatrix} \mathbf{\Psi}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{\Psi}^{\ddagger} \end{bmatrix} \quad \text{and} \quad \mathbf{A}^{\text{hyb}} = \begin{bmatrix} \mathbf{A}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{A}^{\ddagger} \end{bmatrix} \tag{3.4}$$

where, the superscripts † and ‡ denote the primary and secondary exemplar spaces, respectively. The cost function in this setting thus can be expressed as:

$$\begin{aligned} \mathcal{C}' &= D_{KLD}(\mathbf{\Psi}^{\text{hyb}} \| \mathbf{A}^{\text{hyb}}\mathbf{X}^{\text{hyb}}) + \mathbf{\Lambda} \odot \mathbf{X}^{\text{hyb}} \\ &= D_{KLD}\left(\begin{bmatrix} \mathbf{\Psi}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{\Psi}^{\ddagger} \end{bmatrix} \middle\| \begin{bmatrix} \mathbf{A}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{A}^{\ddagger} \end{bmatrix} \mathbf{X}^{\text{hyb}} \right) + \mathbf{\Lambda} \odot \mathbf{X}^{\text{hyb}} \\ &= D_{KLD}(\mathbf{\Psi}^{\dagger} \| \mathbf{A}^{\dagger}\mathbf{X}^{\text{hyb}}) + \beta D_{KLD}(\mathcal{T}_S\mathbf{\Psi}^{\ddagger} \| \mathcal{T}_S\mathbf{A}^{\ddagger}\mathbf{X}^{\text{hyb}}) + \mathbf{\Lambda} \odot \mathbf{X}^{\text{hyb}} \end{aligned}$$

using (3.4) and since the cost function being element-wise. It can thus be seen that the secondary space in effect acts as a regularisation to obtain the activations and β acts as the regularisation weight.

3.3 Description of input spaces

The various input spaces which are chosen to evaluate the proposed approach along with the chosen baseline systems are described in this section.

3.3.1 Mel, DFT and MS only baselines

For a fair evaluation and completeness, three single-input space baseline systems which uses the Mel, DFT and MS representations respectively are evaluated and compared first. All these systems are evaluated using the coupled Mel output dictionary approach depicted in Figure 3.1 with the hybrid exemplars replaced by the Mel, DFT and the MS exemplars, respectively.

Mel baseline: This system uses the *Mel exemplars*, which are created by reshaping the Mel-integrated magnitude spectra of acoustic data spanning T frames. The decomposition of the noisy data expressed in the Mel exemplar domain is obtained using the Mel dictionary, $\mathbf{A}^{\text{mel}} = [\mathbf{S}^{\text{mel}}\ \mathbf{N}^{\text{mel}}]$. The Wiener filter for the noisy Mel enhancement is found using the procedure explained in Section 3.2.1. Also notice that these dictionaries act as the primary exemplar space dictionaries also for the proposed hybrid approach.

DFT baseline: For this setup, the coupled DFT and the Mel dictionaries are obtained first, with the DFT and Mel exemplars extracted from the same piece of training data. To obtain a DFT exemplar, magnitude spectrogram of a training data spanning T frames is reshaped to a vector. For evaluation, the DFT exemplar representation of the noisy data is decomposed using the *DFT dictionary*, $\mathbf{A}^{\text{dft}} = [\mathbf{S}^{\text{dft}}\ \mathbf{N}^{\text{dft}}]$. The activations thus obtained, $\mathbf{X}^{\text{dft}}$ are then applied on to the coupled Mel dictionary to get the speech and noise estimates for noisy Mel enhancement (ref. Section 3.2.2).

MS baseline: The MS representation was proposed as part of a computational model for human hearing which relies on the low frequency amplitude modulations within various frequency bands [141] which are called modulation envelopes. Let M be the number of frequency bands considered. The MS representation for acoustical data is obtained by taking the short-time Fourier transform (STFT) of the modulation envelopes corresponding to each frequency band [71]. For non-negativity, only the magnitude of the STFT is considered.

Because of the low-pass filtering operation, only very few lower bins of the MS will contain significant energy and it is possible to truncate each of the MS to the lowest k bins [16]. All these truncated MS of size $k \times T$ each are then stacked to obtain a matrix of size $(M \cdot k) \times T$ which are referred to as *MS features* [7]. The MS exemplars are then obtained by reshaping the MS features which are stored in the *MS Dictionary*, $\mathbf{A}^{\text{MS}} = [\mathbf{S}^{\text{MS}}\ \mathbf{N}^{\text{MS}}]$. The MS baseline system is then evaluated using the coupled dictionary approach explained in Section 3.2.2 with the decomposition obtained in the MS exemplar space.

3.3.2 Hybrid input spaces: Mel-DFT and Mel-MS spaces

In this work, we investigate the Mel-DFT and Mel-MS hybrid spaces. For this, the Mel, DFT and MS exemplars are created first as explained in Section 3.3.1 from the same piece of data spanning T frames. The trimmed secondary exemplars are then created, by applying $\mathcal{T}_S$ on the DFT and MS exemplars, which are also scaled with β_1 and β_2, respectively. These are then concatenated with the corresponding Mel exemplar (ref. Section 3.2.2) to get the hybrid Mel-DFT and Mel-MS exemplar representations, respectively.

During testing, for every sliding window of length T along the length of the noisy utterance, the Mel and the secondary exemplar representations are obtained. The secondary exemplar representation is then trimmed using $\mathcal{T}_S$ and scaled, followed by concatenating with the Mel exemplar representation to be stored as columns in $\mathbf{\Psi}^{\text{hyb}}$.

3.4 Evaluation experiments

3.4.1 Experimental setup

For evaluation, test sets 'A' and 'B' of the AURORA-2 corpus which contains utterances of digits from '0-9' and 'oh' are used. The training set of the corpus is composed of 8440 clean speech utterances and 6768 noisy utterances which are corrupted by four additive noises (subway, babble, car and exhibition hall). Test Sets A and B are used for evaluating the various techniques discussed in this chapter (refer Section 1.4.1). The WERs obtained after taking the average over the four noise types for clean speech, -5 dB and the combined average of results obtained for SNRs ranging from 20-0 dB are presented.

The noise data required to obtain the noise exemplars are created from the noisy training set using the two step procedure explained in [64]. The clean and the noise samples are pre-processed by removing the dc component and applying pre-emphasis with filter coefficient 0.97. The exemplars for the Mel, and the trimmed DFT and MS spaces are then created using the steps explained in Section 3.3. To extract the coupled exemplars, random pieces of training data spanning 300 ms were used. No supervision was done to avoid the overlap between the chosen random pieces of data or to avoid silence. Then, for each of the chosen random piece of training data:

1. To obtain the Mel exemplars, the DFT of the chosen random piece of training data was first obtained using a window length and hop size of 25 ms and 10 ms respectively with 128 frequency bins within the Nyquist frequency (4 kHz), leading to a DFT representation of size 128 × 30. This is then Mel-integrated with $M = 23$ channels. These frame-level Mel features of size 23 × 30 thus obtained are then reshaped to obtain a Mel exemplar of length 690.

2. To obtain the DFT exemplar, the DFT representation obtained in Step 1 is reshaped to a vector (of length $3,840$).

3. To obtain the MS feature representation, the data is first split across $M = 23$ frequency channels using the equivalent rectangular bandwidth filter banks implemented using Slaney's toolbox [165]. Each of these is then half-wave rectified and low-pass filtered at a 3 dB cut-off frequency of 30 Hz to obtain the modulation envelopes. The modulation spectra for each channel is then found by taking the STFT of each of these envelopes with a window length of 64 ms and hop size 10 ms. With the sampling frequency of 8 kHz and STFT with 128 bins within the Nyquist frequency, each of the spectra was truncated to $k = 5$ bins and are stacked to get the MS features [7] of size 115 × 30. The MS exemplar representation is then obtained after reshaping the MS features to a vector of length $3,450$.

For evaluation, the coupled dictionaries $\mathbf{A}^{\text{mel}}$, $\mathbf{A}^{\text{dft}}$ and $\mathbf{A}^{\text{MS}}$ were created with 10000 speech and noise exemplars each. The hybrid input space dictionaries were then created as explained in Section 3.3.2 for different choices of S, β_1 and β_2. During testing, the corresponding exemplar space representations of the noisy data, $\mathbf{\Psi}$, were obtained as explained in Section 3.3 using the settings given above. The NMF-based decomposition was obtained with 600 multiplicative updates with sparsity constraint. A sparsity penalty of 1.5 for speech and 1 for noise exemplars as in [62] were used for all the evaluated decompositions except for the MS and DFT baselines, both of which used 1.75 and 0.75 respectively as in [7]. GPUs were used to accelerate the NMF iterations using the MATLAB parallel computing toolbox.

For the ASR back-end, a GMM-HMM based decoder using MFCC features was used. Each digit in the HMM topology was described by 16 states together with 3 silence states resulting in a total of 179 states. The GMM models were trained on the MFCCs obtained from the clean training data and enhanced noisy training data using the respective front-ends (referred to as *retraining*), with 13 static features along with their velocity and acceleration coefficients leading to a 39 dimensional feature space. The GMM for each of the HMM state was modelled using 32 Gaussians with diagonal covariance.

Table 3.1: WER in % obtained for various baseline systems as a function of SNR in dB evaluated on a subset of 100 files per test set of the AURORA-2 database. The average execution time per utterance required by the setting is also shown.

		test set A						*test set B*						*Exec.*
	cln	20	15	10	5	0	-5	20	15	10	5	0	-5	*time*
Mel BL	0.2	1.5	1.9	3.7	5.0	**11.6**	**27.6**	1.3	**1.6**	**4.4**	**9.0**	**26.7**	**57.9**	5.8s
DFT BL	0.1	0.9	1.9	2.7	7.2	17.2	33.3	0.6	**1.6**	6.3	14.2	35.1	67.8	12.2s
MS BL	**0.0**	**0.7**	**1.3**	**1.9**	**4.4**	12.5	30.5	**0.5**	1.7	5.1	11.2	34.8	69.0	10.8s

3.4.2 Comparison between the baseline systems

To reduce the experimentation time, we compare the three chosen baseline systems evaluated on a subset of 100 files per test set which is tabulated in Table 3.1. It can be seen that for test set A, the Mel baseline performs better at lower SNRs and as the SNR increases, higher dimensional features yield better separation than the Mel features resulting in improved WERs. The higher dimensionality of these features results in poorer modelling of the unseen cases which explains their inferior performance for test set B. Also notice that the MS and DFT baseline settings are computationally expensive which is almost twice that of the Mel baseline setting.

The different baseline streams were also found to yield complementary results which can also benefit the hybrid input space approach. For the remaining part of this chapter, the Mel exemplar system is chosen as the baseline for its good performance, lower dimensionality and also being the primary input space for the hybrid setup.

3.4.3 Parameters for the hybrid Mel-DFT space

With the baseline system chosen as the Mel exemplars only case, which is the same as the primary exemplar space chosen for the proposed hybrid spaces, the effectiveness of the proposed approach relies on the optimal choices of S and β. These are the two parameters which decide on the contribution of the secondary feature spaces on regularising the speech and noise separation resulting from the Mel baseline system. This section details the analysis of the hybrid Mel-DFT space for different choices of S and β_1 which is summarised in Table 3.2.

Table 3.2: WER in % obtained as a function of SNR in dB on the AURORA-2 database for the hybrid Mel-DFT space approach. The results obtained for various choices of S and β_1 are given.

			test set A		*test set B*	
S	β_1	clean	(20-0)	-5	(20-0)	-5
Mel Baseline		0.4	5.2	28.0	9.2	**59.4**
Hybrid Mel-DFT space						
{1}	0.5	0.4	5.0	27.6	9.3	59.9
{1}	0.2	0.4	4.9	27.1	9.2	60.1
{15}	0.2	0.4	5.1	27.8	9.4	60.4
Switching	0.2	**0.3**	**4.7**	**26.6**	**9.0**	59.7

As the minimum choice, the effect of using the secondary DFT spaces with $|S| = 1$ are investigated. As a pilot experiment, the effect of the first DFT frame i. e., $S = \{1\}$ with $\beta_1 = 0.5$ is investigated, which resulted in marginal performance improvement over the Mel only system. The optimum value of β_1 to get the best separation was then found to be 0.2 after doing a grid search in the range [0.05, 0.5] on a subset of 100 files per noise type . The $S = \{1\}$ system with tuned β_1 is then evaluated over the complete test set which confirmed the effectiveness of the secondary DFT space in significantly improving the recognition results.

With the middle frame more correlated with the other frames in the given temporal context of 30 frames, the choice of $S = \{15\}$ was supposed to be more effective as it can be a better representative of all the DFT frames compared to the first DFT frame. However, on the contrary, the simulation experiments yielded inferior performance when compared to the $S = \{1\}$ case.

An analysis of the $S = \{1\}$ and $S = \{15\}$ cases revealed that such a fall in performance can be attributed to the reshaping operation when considering multiple frames (here, $T = 30$) to obtain the exemplar space representation of the noisy test utterance $\mathbf{\Psi}^{\text{hyb}}$ (ref. Section 3.2). For the utterances in which the speech onset happens before the $15th$ frame, $S = \{15\}$ system was found to fail in detecting the speech onset resulting in a substitution or deletion. To address this and to capture the effectiveness of the middle DFT frame, a switching system which chooses the set S adaptively along the length of the utterance is proposed.

The proposed switching approach is depicted in Fig. 3.3. As explained in Section 3.2, the noisy utterance is first converted to the primary and secondary

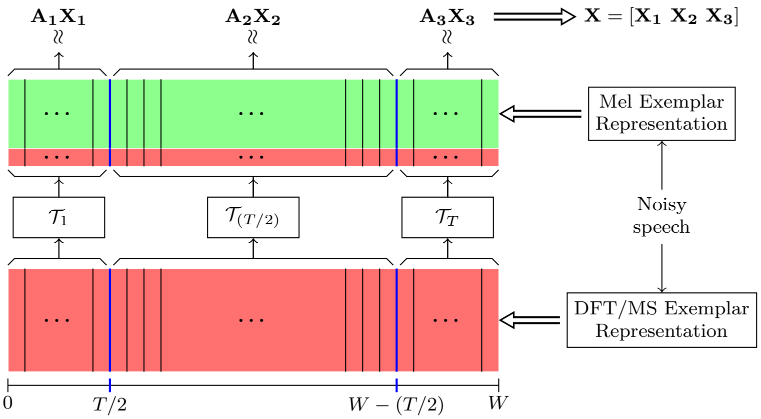


Figure 3.3: Block diagram overview of the proposed switching approach to obtain the activations.

exemplar space representations by means of a sliding window spanning T frames along the length of the utterance. Let W be the total number of resulting sliding windows. In the switching approach, for the first and the last $T/2$ sliding windows of the utterance we use the secondary exemplar with $S = \{1\}$ and $S = \{T\}$ respectively, and $S = \{T/2\}$ for all the remaining windows falling in the middle. Thus we need to use three different dictionaries ($\mathbf{A_1}$, $\mathbf{A_2}$ and $\mathbf{A_3}$) depending on the choice of S in this setup, and the resulting activations are concatenated to obtain the overall activations as $\mathbf{X} = [\mathbf{X_1}\ \mathbf{X_2}\ \mathbf{X_3}]$. It can be seen from Table 3.2 that the assessment of the proposed approach yielded improved WERs over all the other investigated setups.

The performance improvement over the baseline system can be attributed to the inclusion of a secondary feature space which can regularise and improve the speech and noise separation. Also notice that the secondary space is not required to span the entire temporal context considered per exemplar to obtain a significant improvement in separating speech from noise.

3.4.4 Comparison of Mel-DFT and Mel-MS spaces

A comparison between the systems using the proposed hybrid input spaces is presented in this section. To obtain the Mel-MS results, the switching setup is used with a $\beta_2 = 0.1$ which was found after a grid search same as in Section 3.4.3. Table 3.3 summarises the evaluated results.

Table 3.3: WER in % obtained for various approaches as a function of SNR in dB on the AURORA-2 database.

		test set A		*test set B*	
Experiments	clean	(20-0)	-5	(20-0)	-5
GMM trained on clean data					
Mel Baseline	0.4	5.2	28.0	9.2	59.4
Mel-DFT space	**0.3**	**4.7**	**26.6**	9.0	59.7
Mel-MS space	**0.3**	**4.7**	27.2	**8.8**	**59.2**
GMM trained on enhanced noisy data					
Mel Baseline	0.8	2.9	23.0	7.4	55.5
Mel-DFT space	0.6	2.8	22.2	6.5	53.3
Mel-MS space	**0.5**	**2.7**	**21.2**	**6.2**	**52.3**

It can be seen that both the proposed approaches yield statistically significant ($p < 0.01$) improvement in performances when compared to the Mel baseline system for both seen and unseen noise cases. Also notice that a significant 16% relative WER improvement is obtained on test set B SNR(20-0), suggesting that the proposed approach can mitigate the effects of unseen noise cases as well. Inclusion of the MS space as a secondary space was found be more effective when compared to the DFT space. This can be attributed to the better speech and noise separation properties of the MS features when compared to the DFT features as observed in [7,8].

It was also observed in [7] that the MS features can perform well only for the seen noise cases as the MS features lead to more accurate representation of speech and noise, which will not generalise well for the unseen noises. But in the proposed approach, it is found that using trimmed MS exemplars as secondary features can be beneficial for unseen noises also.

The average execution times per utterance are tabulated in Table 3.4. It can be seen that the hybrid exemplar space yields an improved performance at a comparable computational complexity.

3.5 Conclusion and future work

In this work, we presented an exemplar-based feature enhancement method for ASR using hybrid input spaces and coupled dictionaries. The use of hybrid spaces was found to yield improved recognition accuracies over the

Table 3.4: Average execution time needed in seconds per utterance for the various settings with 20000 exemplars in the dictionary. Size (or length) of the exemplars are also shown.

Setting	Mel Baseline	Mel-DFT Space	Mel-MS Space
Exec. time	5.9 s	6.4 s	6.2 s
Size	690	818	805

baseline system. This chapter also presented an effective way of combining multiple input spaces by means of an adaptively trimmed secondary exemplar representation without much increase in the computational complexity. The trimmed representation is also found to be effective in reducing the effects of overtraining to seen noise cases and generalises better to unseen noise cases when compared to full length exemplar representations.

Further, possibly adaptive, feature dimensionality reduction and its effect on reducing overfitting are to be investigated. Another future work is to study the effect of the number of noise exemplars and sparsity penalties in modelling unseen noise cases.

Chapter 4

Investigating Modulation Spectrogram Features for DNN-based ASR

Deep neural network (DNN) based acoustic modelling has been shown to yield significant improvements over Gaussian Mixture Models (GMM) for a variety of automatic speech recognition (ASR) tasks. In addition, it is also becoming popular to use rich speech representations, such as full-resolution spectrograms and perceptually motivated features, as input to the DNNs as they are less sensitive to the increase in the input dimensionality. In this chapter, we evaluate the performance of a DNN trained on the perceptually motivated modulation envelope spectrogram features that model the temporal amplitude modulations within sub-band speech signals. The proposed approach is shown to outperform DNNs trained on a variety of conventional features such as Mel, PLP and STFT features on both TIMIT phone recognition and the AURORA-4 word recognition tasks. It is also shown that the approach outperforms a sophisticated auditory model based on Gabor filter bank features on TIMIT and the channel matched conditions of the AURORA-4 database.

This chapter is adapted from: Deepak Baby and Hugo Van hamme. *Investigating Modulation Spectrogram Features for Deep Neural Network-based Automatic Speech Recognition.* Proc. INTERSPEECH, ISCA, pp. 2479–2483, September 2015.

4.1 Introduction

Gaussian Mixture Model (GMM) -based hidden Markov models (HMMs) have traditionally been the state-of-the-art in the field of automatic speech recognition (ASR) technology. Recent advances in deep neural network based approaches have shown significant performance improvements over the GMM based approaches on a variety of ASR tasks [42,46,81], thanks to its multiple hidden layers learning rich multiple projections. It is also shown to be robust to various kinds of distortions when compared to the GMMs [45,162], sometimes improving the performance by large margins.

However, the DNN performance is still far from that of humans especially in noisy environments. Therefore, there is still a growing interest in feature-related research that focuses on applying our knowledge about human auditory processing into this framework. Traditionally, GMMs needed uncorrelated observations due to its diagonal covariance design and this forced most of these attempts to make use of a feature decorrelation step in the end. On the other hand, it is shown that DNNs are less sensitive to the increase in input dimensionality and correlation between features. In particular, Mel-filter bank outputs are shown to yield better performance than the conventional lower dimensional features such as MFCC or PLP coefficients [46,128]. This allows us to use richer, physiologically motivated features to train the DNNs and aim a better cross-fertilization between the human speech recognition (HSR) and the ASR communities.

There exist some studies that evaluate the performance of DNNs trained on physiologically motivated features. Most of these analyses take into account the poorer frequency resolution of the basilar membrane and the role of spectral and temporal modulations in human hearing. In [126], a comparison of various features such as Gammatone filter coefficients, damped oscillator coefficients etc. extracted from the time domain signal without explicitly going to the frequency domain are presented. Another approach is to extract the various spectro-temporal modulation patterns from the log-compressed Mel spectrogram. An investigation based on the Gabor filter analysis and amplitude modulation filter-banks are presented in [121]. Most of these features were found to yield better performance over the Mel-filter bank features.

In this work, we investigate the performance of an auditory model which relies on the amplitude modulations within frequency bands [141]. These are computationally modelled as modulation envelopes that capture the amplitude envelope of the half-wave rectified sub-band speech signals [71]. These features have been successfully used for noise robust speech recognition [100] and phone classification [39]. Since the human speech contains very low modulation frequencies of the order of 20-30 Hz [158], a low-pass filter with a cut-off frequency of around 30 Hz is employed to capture the speech information.

The low-pass filtering used to obtain the modulation envelopes has two benefits; One, it helps in getting rid of the added noise containing higher modulation frequencies. Two, it yields a compact representation of speech in the spectral domain. Therefore, the spectrograms of these envelopes are taken and are truncated to the lowest few significant bins that fall below the 3 dB cut-off frequency of the low-pass filter used. This representation of modulation envelopes in the spectral domain are referred to as modulation spectrogram (MS) features. In our previous works, these features have been successfully used for exemplar-based speech enhancement as a front-end for DNN-based ASR [5].

In this work, the MS features are used to train and evaluate a DNN-based recogniser and the results are compared with the traditional Mel, STFT and PLP features on TIMIT and AURORA-4 databases. We also include a comparison with the Gabor filter bank features investigated in [155]. The rest of the chapter is as follows: Section 2 details the MS feature extraction and other baseline features together with the DNN architecture used for evaluation. Section 3 details the evaluation setup followed by the results and discussion in Section 4. Section 5 concludes the chapter along with some suggestions for future work.

4.2 Methods

4.2.1 MS features

The MS representation was proposed as part of a computational model for human hearing which relies on the low frequency amplitude modulations within various frequency bands [141]. The processing chain used to obtain the MS features is depicted in Figure 4.1.

To obtain the MS features, the input speech signal is first filtered using a filter-bank having M channels to model the poor frequency resolution of the basilar membrane. This is implemented using an equivalent rectangular bandwidth (ERB) filter bank whose center frequencies are equally spaced along the log-frequency axis that also model the non-linear frequency resolution property of cochlea as defined in [139]. In this work we used ERB filter bank implemented using Gammatone filters [164]. The frequency response of these filters are shown in Figure 4.2. The resulting M band-limited signals are half-wave rectified to model non-negative nerve firings. The modulation envelopes are

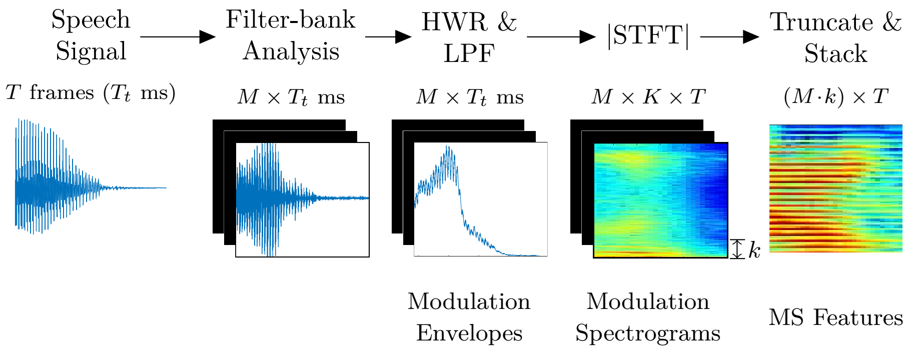


Figure 4.1: Block diagram overview of the processing steps to obtain the proposed MS features. The corresponding sizes of each of the representation are also shown.

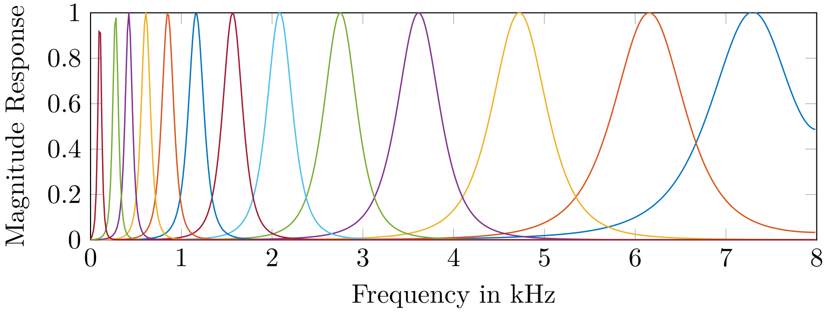


Figure 4.2: Frequency response of the equivalent rectangular bandwidth filters used to model the basilar membrane.

obtained by low-pass filtering these rectified sub-band signals at a 3 dB cut-off frequency of around 30 Hz, since human speech contains very small amplitude modulations.

From these envelopes which contain only low frequency signals, the modulation spectrograms are obtained by taking the magnitude STFT, resulting in M modulation spectrograms [71] of size $K \times T$ each, where K is the number of modulation frequency bins used to obtain the STFT and T is the number of frames in the signal. As there is a low-pass filtering operation, it is possible to truncate each of these modulation spectrograms to their lowest few, say k, bins [8,16], i.e, each modulation spectrogram now has size $k \times T$. Only the positive half of the magnitude modulation spectrogram is considered. To obtain a compact two-dimensional representation, we stack these modulation spectrograms originating from M channels to a matrix of size $(M \cdot k) \times T$. These are then log compressed to model the non-linear intensity to loudness variation of the ear. These are referred to as the MS features. Notice that k denotes the number of amplitude modulation frequencies within each frequency sub-band that are used in the model.

The dimensionality of the MS features depends on the value of K which is approximately equal to the window-length used during the STFT step, the sampling frequency f_s and the 3 dB cut-off frequency of the low-pass filter f_{3dB} used to obtain the modulation envelope. The value of k thus will be roughly $\geq f_{3dB} \cdot K/f_s$. i.e., a higher value of K and/or k can be used to capture more temporal amplitude modulation frequencies.

4.2.2 Baseline features

In this work, we compare the proposed set of features with the conventional Mel, short-time Fourier transform (STFT) and the perceptual linear prediction (PLP) features. We also include the comparison with another physiologically inspired features using Gabor filters, dubbed GBFB features [155]. GBFB features are computed by processing the log-Mel spectrogram with 31 frequency channels by a number of 2D modulation filters. In this setup, the 2D Gabor filters are defined as the product of a complex sinusoidal function and a Hann envelope function, such that they cover a wide range of spectro-temporal modulation patterns [155]. With the setting described in [121], 59 spectro-temporal filters are used per Mel channel which resulted in a total of 1829 components. These are then reduced to 657 features per frame by removing redundant features. For further details, we refer the reader to [121,155].

4.2.3 DNN decoder

The evaluations are done using the “recipe” DNN-HMM-based recogniser in the Kaldi toolkit [143]. A DNN is simply a multi-layer perceptron with multiple hidden layers between its inputs and outputs. Performing back-propagation training on such a network can result in a poor local optimum with a randomly initialised network weights. To circumvent this, a pre-training is done first by considering each pair of adjacent layers as restricted Boltzmann machines (RBM) [80] and then a back propagation training is done over the entire network such that it provides posterior probability estimates for the HMM states [180]. All DNNs used are comprised of 6 hidden layers with 2 048 sigmoid neurons per layer. The input layer used a temporal context of 11 frames.

To perform ASR using a DNN-HMM-hybrid setting, the state emission likelihoods generated by the GMMs are replaced by the pseudo-likelihoods or scaled-likelihoods generated by the DNN.

4.3 Evaluation setup

4.3.1 TIMIT database

TIMIT is a benchmark database for evaluating and comparing the phone recognition accuracy of various ASR systems in clean conditions. The description of the database can be seen in Section 1.4.1. The core test set of the database is used for evaluation, which contains 192 utterances with 8 sentences each from 24 speakers. The phone error rates (PER) in % are reported for evaluations on the TIMIT database.

4.3.2 AURORA-4 database

AURORA-4 database is a large vocabulary continuous speech recognition database based on the WSJ0 corpus of read English speech. The details of the database can be found in Section 1.4.1. The multicondition training data containing 7 138 utterances with channel variations and added noise is used for training the DNNs. The ASR settings are evaluated on test01-test14 sets of the database that contains both single microphone and multiple microphone test conditions. The development set of the database was used for parameter tuning and cross-validation. Word error rates (WER) in % is used to compare the various systems evaluated on this database.

4.3.3 Feature extraction

All the testing and training data are first pre-processed using a DC removal filter and a pre-emphasis filter of coefficient 0.97 before extracting the features. The STFT features are obtained by taking the STFT of the signal with a window length of 25 ms and a window shift of 10 ms with 512 bins. The absolute values of the positive half of the STFT is taken to obtain STFT features of size $B = 256$ per frame. To obtain the Mel features, the STFT features are Mel integrated with $M = 40$ channels resulting in 40 Mel features per frame. The log compressed Mel and STFT features are used to train the DNNs as they were found to yield better results than the raw format. The PLP features are

Table 4.1: Summary of the MS settings evaluated along with the modulation frequencies considered and the number of features per frame.

Setting	K	k	**Mod. freqs. taken (Hz)**	**Size**
$MS_{1024;5}$	1024	5	0, 15, 30, 45, 60	200
$MS_{1024;3}$	1024	3	0, 15, 30	120
$MS_{2048;5}$	2048	5	0, 7.5, 15, 22.5, 30	200

extracted using the Kaldi feature extraction script with 40 Mel channels and 13 PLP coefficients per frame are computed. The GBFB features are extracted using the code provided in [121] which yielded 657 Gabor features per frame.

To obtain the MS features, equivalent rectangular bandwidth filter bank containing $M = 40$ channels, implemented using Slaney's toolbox [165], is used to obtain the sub-band signals. These are then half-wave rectified and low-pass filtered at a 3 dB cut-off frequency of 30 Hz to obtain the modulation envelopes within each sub-band. Then an analysis is made for various choices of K, which decides the resolution of the modulation frequencies used, and k which decides the set of amplitude modulation frequencies considered. Two choices of K are used; a window length of 64 ms with $K = 1024$, and a window length of 128 ms with $K = 2048$. The resolution of the modulation spectra will be roughly 15 Hz and 7.5 Hz with $K = 1024$ and 2048, respectively. The evaluations are then made for various choices of k. The settings evaluated are summarised in Table 4.1.

Since the alignments used for the DNN training are taken from a GMM-based back-end which used a shorter window length (25 ms) than the ones used by the MS, there will be a state-frame misalignment when MS features are used. It is found that the MS features with window length 64 ms and 128 ms lead the GMM features by 2 and 4 frames respectively and the alignments are corrected by delaying the MS features by the respective number of frames. Also notice that the MS features take into account a temporal context of 165 ms when 11 consecutive frames are used for DNN training, whereas all the baseline features span only 115 ms context. For a fair comparison, we also include another baseline system based on Mel features which uses a temporal context of 15 frames (splice = 7) which adds to 165 ms context (denoted as Mel_{splice7}).

Table 4.2: Average PER in % obtained for the TIMIT speaker-independent phone recognition task with DNNs trained on various input features.

Features	**PER in %**
Mel	21.5
Mel_{splice7}	21.8
$STFT$	22.1
PLP	21.6
$GBFB$	21.0
$MS_{1024;5}$	**19.6**
$MS_{1024;3}$	19.8
$MS_{2048;5}$	20.6

4.4 Results and discussion

4.4.1 Results on TIMIT database

The PER results obtained for various settings on the TIMIT database are presented in Table 4.2. Notice that no speaker adaptation is done on any of these features. It is observed that using a splice of 7 frames with Mel features is found to be performing worse than the splice equal to 5 setting. It can be seen that both the perceptually motivated models (GBFB and MS) outperform the traditional features and the MS features yield the best result with a phone recognition accuracy of more than 80 % on the TIMIT database. Given the splice 5 vs. splice 7 comparison with the Mel features, this improvement cannot be attributed to a longer temporal context used by the MS features.

It can also be seen that including more modulation frequencies ($MS_{1024;5}$ vs. $MS_{1024;3}$) indeed can benefit the PER performance. It is also seen that increasing the modulation spectral resolution by increasing K could be detrimental (ref. $MS_{2048;5}$) mainly because of its too long temporal context (11 frames correspond to 218 ms) which could cover multiple phones at a time and may result in a poorer classifier.

When compared to the GBFB features, MS features gave an absolute PER improvement of 1.4 % (7% relative). This is in fact one of the best results reported on the TIMIT database with the given DNN architecture without any speaker adaptive training.

Table 4.3: Summary of results on the AURORA-4 database with DNNs trained on various input features.

Features	**test A**	**test B**	**test C**	**test D**	**Avg.**
Mel	3.5	7.4	10.3	21.5	13.3
Mel_{splice7}	3.3	7.5	9.8	21.0	13.2
STFT	3.3	7.6	10.8	22.0	13.7
PLP	3.9	8.6	10.4	22.8	14.5
GBFB	3.5	7.3	**8.0**	**19.0**	**12.1**
$MS_{1024;5}$	**2.6**	7.1	8.7	19.5	12.2
$MS_{1024;3}$	2.8	**6.9**	10.0	20.5	12.7

4.4.2 Results on AURORA-4 database

Next, the noise robustness of the features are evaluated on the AURORA-4 database. The WERs obtained on each of the test sets in the AURORA-4 for various input features are detailed in Table 4.4. It can be seen that both GBFB and MS features yield a better robustness to both channel variation and noisy conditions over the Mel, STFT and PLP features. The $MS_{2048;5}$ is not evaluated as it gave poorer performance on the TIMIT database. MS features yielded the best performance on the single microphone cases. In particular, a significant WER improvement even on clean speech is obtained which is even better than the results obtained for the same DNN setting trained on Mel features extracted from the clean training data of the database (2.9 % reported in [5]).

The summary of WERs obtained on various test sets are presented in Table 4.3. It can also be seen that including more modulation frequencies improves the performance in channel mismatched conditions (ref. test C and test D results of $MS_{1024;5}$ vs. $MS_{1024;3}$). For multiple microphone cases GBFB features performed better because of its sophisticated design in which the features are chosen such that they exhibit robustness to channel variations and noisy conditions. However, no such adaptation is done for the MS feature extraction. Also notice that the MS features use fewer features per frame when compared to the GBFB features. These results reaffirm the effectiveness of combining perceptually motivated rich features as inputs to the DNNs.

Additional experiments were also conducted by concatenating the Mel features with the MS features. However, the evaluations using these concatenated features (not shown) yielded more or less similar results as the MS features. It implies that the information provided to the DNN by the MS and Mel features are not complementary in general and no additional information is introduced by the Mel features.

4.5 Conclusions

In this chapter, we evaluated the performance of the perceptually motivated modulation spectrogram features as input features to DNNs. The approach yielded a PER of 19.6 % on the TIMIT database which is among the best results published on the database without any speaker adaptive training. Further, the noise robustness of these features are evaluated and compared on the AURORA-4 database and it is shown that MS features yield robust performance in all cases when compared to the Mel, STFT and PLP features. When compared to the GBFB features, MS features gave a better performance on single microphone cases. These results reaffirm that DNNs can be effectively combined with perceptually motivated features to bridge the gap between the ASR and HSR performances.

Further evaluations of MS features with other choices of low-pass cut of frequencies f_{3dB} and other values of K, to vary the number amplitude modulation frequencies considered, are to be done. Other future work is to incorporate channel adaptation and speaker adaptation into the MS feature extraction framework.

Table 4.4: Average WERs in % obtained on each test set of the AURORA-4 database for DNNs trained on various input features.

	Mic 1							**Mic 2**							**Avg.**
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	
Mel	3.5	4.3	6.7	9.7	8.2	6.6	8.7	10.3	14.5	20.9	24.6	24.4	20.1	24.4	13.3
Mel_{splice7}	3.3	4.5	7.1	9.8	8.3	6.7	8.6	9.8	14.1	20.5	24.6	23.0	19.6	24.4	13.2
STFT	3.3	4.6	6.9	9.6	8.9	6.7	8.9	10.8	14.9	21.3	25.6	24.5	20.5	25.2	13.7
PLP	3.9	5.3	7.8	10.7	9.8	7.8	10.0	10.4	14.9	22.6	25.9	26.0	21.4	25.8	14.5
GBFB	3.5	4.6	6.7	9.1	8.0	6.8	8.2	**8.0**	12.9	**18.8**	**23.2**	**20.7**	18.1	**20.4**	**12.1**
$MS_{1024;5}$	**2.6**	4.0	6.6	**8.8**	8.4	**6.4**	8.7	8.7	**11.8**	20.0	23.7	21.5	**18.0**	22.0	12.2
$MS_{1024;3}$	2.8	**3.9**	**6.5**	**8.8**	**7.7**	6.7	**8.0**	10.0	14.6	20.7	23.6	22.6	18.7	23.1	12.7

Chapter 5

Joint Denoising and Dereverberation with Decaying Norm Constraint

This chapter extends the exemplar-based technique for single-channel speech enhancement in noisy reverberant environments using a novel approximation of the noisy reverberant speech in the frequency domain and non-negative matrix deconvolution (NMD). In the proposed model, the room impulse response (RIR) in the magnitude STFT domain is defined such that its decaying structure can also be estimated from the test data itself, whereas the existing models used a sub-optimal bin-wise clamping procedure to impose such a decaying structure which does not hold in a typical RIR. This chapter presents multiplicative updates for estimating the RIR, its decay and the underlying anechoic speech and noise. The proposed model is evaluated on a synthetically created dataset created by convolving TIMIT recordings with RIRs measured from different rooms and varying speaker-and-microphone locations, and adding background noises taken from the CHiME corpus. Simulation results show that the proposed model results in a better RIR estimate over the existing model and improves various instrumental speech quality measures.

This chapter is adapted from: Deepak Baby and Hugo Van hamme. *Joint Denoising and Dereverberation using Exemplar-based Sparse Representations and Decaying Norm Criterion.* Submitted to IEEE/ACM Trans. on Audio, Speech and Language Processing, 2016.

5.1 Introduction

Speech signals recorded using a distant microphone in a noisy enclosed space are comprised of the original speech, its reflections from various surfaces in the room and the additive background noise. In such noisy reverberant environments, these distortions highly degrade the speech intelligibility for hearing impaired listeners [18,104], recognition accuracy in automatic speech recognition [101,102] and speaker identification systems [148,201]. It is therefore desirable to have a mechanism for noise suppression and dereverberation in these applications. Most of the traditional systems make use of a two step procedure which is comprised of a source separation or denoising stage followed by a dereverberation step. This work focusses on a setting that can simultaneously achieve denoising and dereverberation.

There exist a few unsupervised techniques for joint denoising and dereverberation. Huang *et al.* [84] used prior-knowledge about single-talk periods for channel identification followed by signal estimation. Another technique presented in [198] also requires a-priori knowledge about the noise statistics and speech absence periods. A joint denoising and dererverberation approach using the TRINICON technique that employs the higher order statistics of speech is presented in [22,23]. The conditional separation and dereverberation technique presented in [199] aims at achieving a similar task in a tandem manner where a blind source separation and blind dereverberation approaches are alternated to optimise the task.

This chapter concentrates on a joint denoising and dereverberation approach that operates on the magnitude spectrogram of the noisy reverberant speech. In this work, the magnitude short-time Fourier transform (STFT) of the reverberant speech at every frequency bin is approximated as a convolution of the magnitude STFT of clean (anechoic) speech signal with that of the room impulse response (RIR) in the corresponding frequency bin. Notice that this model also neglects the cross-band leakage due to windowing and such an approximation based on the non-negative transfer function has been successfully used for system identification and speech dereverberation under noise-free conditions [97,106,116,131,163,176]. Even though such a model holds only approximately, it does not require to explicitly model the phase of the RIR, which is difficult especially in noisy conditions [97]. This work extends this model to noisy scenarios where the magnitude STFT of the noisy spectrogram is decomposed as the sum of the magnitude STFTs of the reverberant speech and of the additive noise.

Notice that the proposed decomposition is comprised of three parameters that are to be estimated from the single-channel noisy reverberant speech recording;

magnitude STFTs of clean speech, additive noise and the RIR. In order to get meaningful estimates from such an approximation, it is important to make use of some constraints that capture the specific properties of speech, noise and the RIR. For modelling speech and noise, an exemplar-based technique based on non-negative matrix deconvolution (NMD) [167] is used where the speech and noise estimates are approximated as a convolution of speech and noise exemplars stored as atoms in a dictionary with the corresponding weights or activations which also capture the temporal continuity of speech and noise. Exemplars are directly sampled versions of the training speech and noise data spanning multiple consecutive frames expressed in the required feature space (magnitude STFT in this case) [62,89]. Thus the dictionary atoms model the spectral structure of speech and noise and the activations model the discovery of these underlying patterns in the recording. Such an NMD-based model using exemplars has been successfully used for speech denoising and noise robust automatic speech recognition (ASR) [86,89]. This work also extends such a system to model the reverberation under noisy scenarios.

In addition, it is also important to impose some constraints on the RIR in order to get a meaningful and reliable decomposition. One of the main apriorily known constraints applied on the RIR is to force it to have a decaying structure [113]. The techniques presented in [131] and [96] also use a similar approximation where non-negative matrix factorisation (NMF) is used to obtain the activations in noise-free and noisy scenarios, respectively. This work differs from these existing approaches in three key aspects. First, the proposed technique makes use of NMD to obtain the approximation which is shown to outperform NMF for the denoising task when there are fewer dictionary atoms [89]. Second, this work deals with dereveberation in noisy conditions which makes the estimation problem even more challenging. Notice that [96] also deals with noisy scenarios, but the RIR is estimated based on both speech and noise estimates. However, such an estimate is not reliable when there are multiple and/or moving noise sources which results in varying RIRs over time. In this chapter, the RIR estimate is obtained based only on the speech estimate assuming the speaker is stationary when the recording is made. Finally, the decaying norm constraint on the RIR is not implicitly modelled in the previous works and the constraint is imposed separately which may result in sub-optimal estimates. This chapter proposes a novel approach where a decaying norm criterion is imposed on the RIR which equips the model to uncover a better RIR estimate together with its decay shape from the noisy speech.

The proposed model thus has an additional parameter that estimates how the RIR decays over time (or frame in this work since the approximation is in the magnitude STFT space). The estimators for the activations, RIR and the decays are derived for the proposed model that minimise the Kullback-

Leibler divergence between the original noisy reverberant speech and its approximation. Experimental results show that the proposed model improves various instrumental speech quality measures when compared to the previously proposed joint denoising and dereverberation technique where the decaying structure of the RIR is not implicitly modelled [6].

The rest of the chapter is organised as follows. Section 5.2 details an existing model detailed in [6,131] on which this work is based on, estimators for obtaining the model parameters and its limitations. The proposed technique together with the estimators for the activations, RIR and the decays are described in Section 5.3. The evaluation setup is explained in Section 5.4 followed by results and discussion in Section 5.5. Section 5.6 concludes the chapter with some suggestions for future work.

5.2 Existing model

5.2.1 Non-negative representation of noisy reverberant speech

Let $y[n]$ and $w[n]$ be the clean and noise signals, respectively. The room impulse response (RIR) is assumed to be a finite-impulse response $h[n]$ of length L_t. The noisy reverberant speech recorded by the microphone can be written as $z[n] = h[n] * y[n] + w[n]$. Notice that, we are only interested in the RIR that is being convolved with the clean speech signal and everything else is modelled as noise. Such an assumption will be more realistic in scenarios where there are multiple and/or moving noise sources. In the complex STFT domain $z[n]$ can be approximated as [66,133,176]:

$$\mathcal{Z}(f,t) \approx \sum_{p=1}^{L} \mathcal{H}(f,p)\mathcal{Y}(f,t-p+1) + \mathcal{W}(f,t) \tag{5.1}$$

where $\mathcal{Z}$, $\mathcal{H}$, $\mathcal{Y}$ and $\mathcal{W}$ denote the complex-valued STFT of $z[n]$, $h[n]$, $y[n]$ and $w[n]$, respectively. f and t denote the frequency-bin and frame indices, respectively. L denotes the length of the RIR in the STFT space. Let the STFT be obtained for $2B$ frequency bins and $\mathcal{Z}$ contains F frames.

For a non-negative representation, the noisy reverberant speech is approximated in the magnitude STFT domain as $\mathbf{Z}(f,t) \approx \sum_{p=1}^{L} \mathbf{H}(f,p)\mathbf{Y}(f,t-p+1) + \mathbf{W}(f,t)$, where $\mathbf{Z} = |\mathcal{Z}|$, $\mathbf{H} = |\mathcal{H}|$, $\mathbf{Y} = |\mathcal{Y}|$ and $\mathbf{W} = |\mathcal{W}|$. Such an approximation has been successfully used for system identification in [96,125,131]. For ease of notations, this approximation can be written in matrix form as:

$$\mathbf{Z} \approx \sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \overset{(p-1)\to}{\mathbf{Y}} + \mathbf{W} \tag{5.2}$$

where $[\mathbf{H}]_p$ is the p-th column of the matrix $\mathbf{H}$ and $\overset{p\to}{\mathbf{Y}}$ denotes the right-shifting operation by adding p columns of zeros to the left and removing the last p columns of $\mathbf{Y}$. The operation $\mathbf{h} \circledcirc \mathbf{Y}$ stands for the element-wise multiplication of a vector $\mathbf{h}$ with the all the columns of $\mathbf{Y}$. Since these magnitude STFTs are symmetric, only the positive half of the magnitude STFT is considered for the rest of the chapter, i.e., $\mathbf{Z}, \mathbf{Y}, \mathbf{W} \in \mathbb{R}_{+}^{B\times F}$ and $\mathbf{H} \in \mathbb{R}_{+}^{B\times L}$.

The goal of the dereverberation task is thus to obtain reliable estimates for $\mathbf{H}$, $\mathbf{Y}$ and $\mathbf{W}$ from the magnitude STFT of the noisy reverberant speech $\mathbf{Z}$. However, obtaining the estimates using such an under-complete system is prone to local optima and may not capture the typical characteristics of the speech, such as its low-rank nature and spectral structure along harmonics, and results in a very poor enhancement performance. Therefore, it is advised to make use of some a-priori knowledge of speech and noise in order to achieve a better and more reliable decomposition.

5.2.2 Modelling speech and noise using exemplar-based sparse representations

This work makes use of exemplar-based techniques where the speech and noise are approximated as a sparse linear combination of atoms (exemplars) in a dictionary. Such models have been successfully used for speech enhancement [7] and speech recognition [64,200]. Exemplars are randomly chosen magnitude STFT patches obtained from a training set containing speech and noise only recordings and are stored as columns in the speech dictionary $\mathbf{S}$ and the noise dictionary $\mathbf{N}$, respectively. The exemplars may span multiple frames, of length say T frames, in order to capture the temporal continuity of speech and noise. These patches are reshaped to a vector of length BT to form an exemplar. Let the dictionaries are comprised of J_s speech and J_n speech and noise exemplars, respectively.

In this work, we use the non-negative deconvolution (NMD) based technique [167] to obtain the approximation of speech and noise spectra as a linear combination of the atoms in the dictionary.

$$\mathbf{Y} \approx \tilde{\mathbf{Y}} = \sum_{t=1}^{T} \mathbf{S_t} \overset{(t-1)\to}{\mathbf{X_s}} \quad \text{and} \quad \mathbf{W} \approx \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\to}{\mathbf{X_n}} . \tag{5.3}$$

The matrix $\mathbf{S_t}$ denotes the t-th block matrix obtained by partitioning $\mathbf{S}$ into T block rows each of size $K \times J_s$ [166]. $\mathbf{N_t}$ is also defined in the same manner from $\mathbf{N}$. The approximation is obtained such that mixing weights or activations $\mathbf{X_s}$ and $\mathbf{X_n}$ are also non-negative. Thus, the model for noisy reverberant speech becomes,

$$\tilde{\mathbf{Z}} = \sum_{p=1}^{L} \sum_{t=1}^{T} [\mathbf{H}]_p \circledcirc \left(\mathbf{S_t} \overset{\tau\rightarrow}{\mathbf{X_s}} \right) + \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\rightarrow}{\mathbf{X_n}} \tag{5.4}$$

using (5.2) and (5.3), where $\tau = p + t - 2$. The problem thus boils down to estimating $\mathbf{H}$ and the activations $\mathbf{X_s}$ and $\mathbf{X_n}$.

5.2.3 Computing the estimates

In the existing model [6], the estimates are obtained such that they minimise the Kullback-Leibler divergence (KLD) between the magnitude STFT of the noisy reverberant speech $\mathbf{Z}$ and its approximation $\tilde{\mathbf{Z}}$. The KLD between z and $\tilde{z}$ is given as:

$$D_{KLD} \left(z \| \tilde{z} \right) = z \, \log \frac{z}{\tilde{z}} + \tilde{z} - z. \tag{5.5}$$

In addition, we apply sparsity constraints such that the resulting activations have a sparse structure. Such a sparse solution can be achieved by adding $\lambda_s \|\mathbf{X_s}\|_1$ and $\lambda_n \|\mathbf{X_n}\|_1$ to the cost function, where $\| \cdot \|_1$ denotes the sum of all the elements in a matrix. Thus a larger value of λ forces a sparser solution for the activations. In principle, the λ value can be specifically tuned to get a better approximation for every frame if extra information such as the speech activity and SNR in the frame are available. In this work, no such knowledge is available a-priori and same sparsity penalties are used for all frames. The optimal values for λ_s and λ_n are typically obtained after parameter tuning on a development set.

The estimates for the RIR and the activations are obtained such that they minimise the cost function,

$$\mathcal{C} = \sum_{f,t} D_{KLD}(\mathbf{Z} \| \tilde{\mathbf{Z}}) + \lambda_s \|\mathbf{X_s}\|_1 + \lambda_n \|\mathbf{X_n}\|_1. \tag{5.6}$$

The set of estimates that minimise this cost can be obtained by alternately applying the following multiplicative updates until convergence. The derivations are provided in Appendix A.1.

$$[\mathbf{H}]_p \leftarrow [\mathbf{H}]_p \odot \frac{\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1} \odot \ [\mathbf{R}]_l}{\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1}}$$

$$\mathbf{X_s} \leftarrow \mathbf{X_s} \odot \frac{\sum_{t=1}^{T} \sum_{p=1}^{L} \mathbf{S_t}^{\intercal} \left([\mathbf{H}]_p \circledcirc \overset{\leftarrow \tau}{\mathbf{R}} \right)}{\sum_{t=1}^{T} \sum_{p=1}^{L} \mathbf{S_t}^{\intercal} \left([\mathbf{H}]_p \circledcirc \overset{\leftarrow \tau}{\mathbf{1}} \right) + \lambda_s}$$

$$\mathbf{X_n} \leftarrow \mathbf{X_n} \odot \frac{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow (t-1)}{\mathbf{R}}}{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow (t-1)}{\mathbf{1}} + \lambda_n}$$

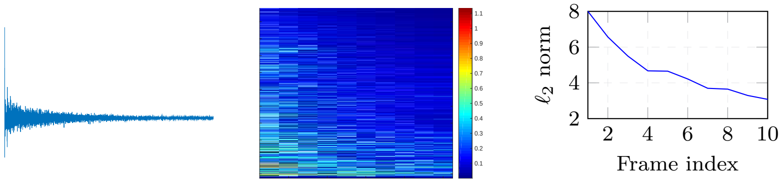
The element-wise ratio $\mathbf{Z} \oslash \tilde{\mathbf{Z}}$ is denoted as $\mathbf{R}$, $\mathbf{1}$ is a matrix of ones of the same size as $\mathbf{Z}$ and $\odot$ denotes element-wise multiplication. The operation $\overset{\leftarrow \tau}{\mathbf{R}}$ shifts the matrix to the left by removing the first τ columns and adding τ columns of zeros to the right.

Notice that obtaining the estimates without any additional constraint may result in a scaling ambiguity since there are two free parameters: activations and the RIR. Then for any scaled value of the RIR $\gamma\mathbf{H}$, the algorithm can give the same minimum cost with the appropriately scaled activations $\mathbf{X_s}/\gamma$. This results in an infinite number of possible solutions that can yield the same minimum cost. Apart from such a non-uniqueness problem, not constraining the RIR may also result in an unrealistic solution that may not capture the typical characteristics of an RIR. It is observed that the RIR has a decaying structure and therefore in the STFT domain, it is safe to assume that the columns of $\mathbf{H}$ have a decreasing ℓ_2 norm. In order to impose such a characteristic, the following operations are done on the RIR estimate after every multiplicative update.

1. In order to force a decaying structure, every element in the RIR matrix $\mathbf{H}$ is clamped such that $\mathbf{H}(f,t) \leq \mathbf{H}(f,t-1)$.

2. Every row of $\mathbf{H}$ is scaled to have a unit ℓ_1 norm in order to bound on the energy introduced by the RIR per frequency-bin.

Once the optimal estimates are obtained, they are used to enhance the noisy STFT by element-wise multiplying with the time-varying filter $\mathcal{Z} \odot \tilde{\mathbf{Y}} \oslash \tilde{\mathbf{Z}}$. The time-domain signal is then obtained using the overlap-add method. Such a setting has been successfully used for dereverberation in [6,133]. This setting is referred to as the *NMD+R* setting.



(a) RIR in time domain

(b) Magnitude STFT of the RIR

(c) ℓ_2 norm of the frame vs. frame index

Figure 5.1: The time and frequency domain representations of an arbitrary room impulse response in the Aachen impulse response database. The ℓ_2 norm vs. frame index is also shown. It can be seen that the ℓ_2 norm of the RIR in magnitude spectral domain decays with frame and no bin-wise decaying dependency is observed.

5.2.4 Limitations of the existing model

Even though the NMD+R setting is shown to yield a decent dereverberation performance [6], the setting is still sub-optimal since neither the cost function nor the multiplicative updates takes the decaying structure of $\mathbf{H}$ into account. Rather, such a decaying structure is imposed by modifying $\mathbf{H}$ after every multiplicative update. Such a setting has the following disadvantages.

1. Since the multiplicative update for $\mathbf{H}$ is derived such that it does not preserve any decaying structure on the columns of $\mathbf{H}$, such a setting limits the discovery of the actual underlying pattern of $\mathbf{H}$ from the noisy reverberant speech. Therefore it is advised to incorporate the decaying nature of $\mathbf{H}$ in the cost function and derive multiplicative updates such that the decaying structure is preserved.

2. Notice that the clamping is done for every frequency-bin with respect to the same bin in the previous frame. However, such a bin-wise decaying structure is not observed in a typical RIR. This step may alter a correctly identified $\mathbf{H}$ and result in a wrong estimate. Figure 5.1 depicts an RIR in the time and magnitude spectrogram domain together with the ℓ_2 norm of every frame in the magnitude spectrogram. It is clear that the RIR does not contain any bin-wise dependency over frames and the ℓ_2 norms decay with frame index.

3. In the existing model, the clamping constraint applied on the RIR is non-linear and there exists no straight-forward way to integrate such an operation to the cost function or to derive the multiplicative updates.

In this work, we propose a novel dereverberation scheme that can address these disadvantages and obtain a more realistic and better RIR estimate.

5.3 Proposed model

This section describes the proposed model for noisy reverberant speech such that an RIR with an inherent decaying structure can be estimated. This model enables the shape of the RIR decay over frames to be estimated from the noisy reverberant speech itself and we present the multiplicative updates for estimating it.

5.3.1 Incorporating decaying norm constraint

In the proposed setting, the RIR matrix $\mathbf{H}$ is replaced by $\alpha_p[\mathbf{H}]_p/\|[\mathbf{H}]_p\|$, where $\|\cdot\|$ denotes the ℓ_2 norm of a vector and α_p is a scalar. The approximations for the clean speech and noise spectrograms are kept the same. The term $[\mathbf{H}]_p/\|[\mathbf{H}]_p\|$ ensures columns having unit norms and α_p controls the norm of every column in the RIR. Thus, in order to have a decaying structure, only the scalar values α_p are to be designed so that $\alpha_p \leq \alpha_{p-1}, \;\; \forall p = 2, \ldots, L$. Notice that these α values are also to be estimated from the noisy reverberant speech.

In order to estimate the α values using multiplicative updates and to force the decaying structure, we define $\alpha_p = (1 + c_p)\alpha_{p+1}$ where c_p is a non-negative number that captures the scaling difference between α_p and α_{p+1} and ensures the decaying structure. Let α be a vector of entries α_p. Thus an arbitrary l-th element of α can be written as $\alpha_l = \prod_{i=l}^{L}(1 + c_i)$ with $c_L = 0$ and $\alpha_L = 1$. Further, the α vector is also constrained to have a unit ℓ_2 norm in order to avoid the ambiguity due to scaling as mentioned in Section 5.2.3, i.e., $\bar{\alpha}_p \triangleq \alpha_p/\|\alpha\|$. Thus the approximation for the noisy reverberant speech in the proposed model can be written as,

$$\begin{aligned}\tilde{\mathbf{Z}} &= \sum_{p=1}^{L} \bar{\alpha}_p \left[\bar{\mathbf{H}}\right]_p \circledcirc \overset{(p-1)\rightarrow}{\tilde{\mathbf{Y}}} + \tilde{\mathbf{W}} \\ &= \sum_{p=1}^{L} \frac{\prod_{i=p}^{L}(1 + c_i)}{\|\alpha\|} \left[\bar{\mathbf{H}}\right]_p \circledcirc \overset{(p-1)\rightarrow}{\tilde{\mathbf{Y}}} + \tilde{\mathbf{W}}. \end{aligned} \tag{5.7}$$

where, $[\bar{\mathbf{H}}]_p = [\mathbf{H}]_p/\|[\mathbf{H}]_p\|$. Thus the problem boils down to estimating the RIR $\mathbf{H}$, decay coefficients c_i and the activations $\mathbf{X_s}$ and $\mathbf{X_n}$.

Notice that there is no bin-wise constraint on the RIR matrix and for every frame the bin-values are given full degree of freedom. In addition, since the decaying norm structure is imposed using non-negative coefficients c_i, they can also be estimated using the multiplicative updates. Such a model with decaying norm constraint can reliably estimate the underlying RIR structure and yield a more realistic solution. This setting is denoted as the *NMD+R+DN* setting.

5.3.2 Multiplicative updates for the estimates

The multiplicative updates are derived such that they minimise the cost function in (5.6) with the proposed approximation for $\tilde{\mathbf{Z}}$ given in (5.7). The derivation of these updates are provided in Appendix A.2. The multiplicative updates for all the required parameters are given below. In addition, after every multiplicative update, the rows of $\mathbf{H}$ is scaled to have a unit ℓ_1 norm to bound the energy contributed by the RIR matrix per frequency-bin.

$$\mathbf{X_s} \longleftarrow \mathbf{X_s} \odot \frac{\sum_{t=1}^{T} \sum_{p=1}^{L} \bar{\alpha}_p \cdot \mathbf{S_t}^{\intercal} \left([\bar{\mathbf{H}}]_p \circledcirc \overset{\leftarrow\tau}{\mathbf{R}} \right)}{\sum_{t=1}^{T} \sum_{p=1}^{L} \bar{\alpha}_p \cdot \mathbf{S_t}^{\intercal} \left([\bar{\mathbf{H}}]_p \circledcirc \overset{\leftarrow\tau}{\mathbf{1}} \right) + \lambda_s} \tag{5.8}$$

$$\mathbf{X_n} \longleftarrow \mathbf{X_n} \odot \frac{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow(t-1)}{\mathbf{R}}}{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow(t-1)}{\mathbf{1}} + \lambda_n} \tag{5.9}$$

$$\left[\bar{\mathbf{H}}\right]_p \longleftarrow \left[\bar{\mathbf{H}}\right]_p \odot \frac{\sum_{l=1}^{F} \left([\mathbf{R}]_l \odot [\mathbf{Y}]_{l-p+1} \right) + [\bar{\mathbf{H}}]_p [\bar{\mathbf{H}}]_p^{\intercal} \sum_{l=1}^{F} [\mathbf{Y}]_{l-p+1}}{[\bar{\mathbf{H}}]_p [\bar{\mathbf{H}}]_p^{\intercal} \sum_{l=1}^{F} \left(\mathbf{R}]_l \odot [\mathbf{Y}]_{l-p+1} \right) \; + \; \sum_{l=1}^{F} [\mathbf{Y}]_{l-p+1}} \tag{5.10}$$

$$c_i \longleftarrow c_i \cdot \frac{\sum_{l=1}^{i} \bar{\alpha}_l [\bar{\mathbf{H}}]_l^{\intercal} \sum_{j=1}^{F} [\mathbf{R}]_j \odot [\mathbf{Y}]_{j-l+1} + \sum_{j=1}^{i} \bar{\alpha}_j^2 \sum_{l=1}^{L} \bar{\alpha}_l [\bar{\mathbf{H}}]_l^{\intercal} \sum_{j=1}^{F} [\mathbf{Y}]_{j-l+1}}{\sum_{l=1}^{i} \bar{\alpha}_l [\bar{\mathbf{H}}]_l^{\intercal} \sum_{j=1}^{F} [\mathbf{Y}]_{j-l+1} + \sum_{j=1}^{i} \bar{\alpha}_j^2 \sum_{l=1}^{L} \bar{\alpha}_l [\bar{\mathbf{H}}]_l^{\intercal} \sum_{j=1}^{F} [\mathbf{R}]_j \odot [\mathbf{Y}]_{j-l+1}} \tag{5.11}$$

5.4 Experimental setup

5.4.1 Dataset used

To evaluate and compare the performance of the various dereverberation methods, recordings from the TIMIT database are used to generate noisy reverberated data. In order to consider a wide variety of conditions, RIRs from the Aachen impulse response database [94] are used. The reverberation times of the rooms are between 0.23 s and 1.20 s. The recordings from the TIMIT database are convolved with a randomly chosen RIR from the Aachen RIR database to obtain the noise-free reverberated signal. These are then added with a randomly chosen noise segment provided with the CHiME challenge [35] at SNRs 5, 10, 15 and 20 dB to create the noisy reverberant data. The core test set of the TIMIT dataset containing 192 recordings of read speech is used to synthetically create the test data. The sampling frequency is 16 kHz.

The RIR database contains impulse responses measured from various rooms such as lecture room, stairway, meeting room and a cathedral with different microphone and source locations. The background noise used also contains a significant amount of reverberation with multiple noise sources that are recorded from a domestic environment.

5.4.2 Dictionary creation

The speech and noise dictionaries used in this work are composed of STFT exemplars extracted from TIMIT training set and the CHiME background noises, respectively. In order to create an STFT exemplar, a random segment of training speech or background noise spanning T frames is chosen and is converted to the magnitude STFT domain with a window-length of 25 ms and a window-shift of 10 ms. The number of FFT bins used is 512 and zero-padding is used whenever necessary. Only the positive half of the magnitude STFT is considered which yields an STFT of size $256 \times T$ which is reshaped to a vector of length $256 \cdot T$ to create an exemplar.

In this work, $J_s = 5000$ speech exemplars and $J_n = 2500$ noise exemplars are used to create the speech and noise dictionaries respectively. A temporal context of $T = 10$ is used resulting in speech and noise dictionaries of size 2560×5000 and 2560×2500, respectively. The sparsity penalties used are $\lambda_s = 1.6$ and $\lambda_n = 0.8$ as they were found to yield the best decomposition in our previous work [6].

For all the evaluated NMD-based settings, the multiplicative updates for activations and RIR are applied 100 times. A RIR length of $L = 10$ frames is used in all the experiments. Both the activations and the RIR are initialised to a matrix of all ones of appropriate size before applying the multiplicative updates. For the NMD+R+DN setting, the decay coefficients are initialised to 0.1 and the multiplicative updates are applied 100 times.

5.4.3 Evaluated methods

In addition to the NMD+R and the proposed NMD+R+DN settings, the speech enhancement using NMD without any RIR model is evaluated first as a baseline setting. This is the same as the NMD+R setting with $L = 1$ and the resulting RIR matrix will be an all one vector. Notice that such a setting is capable of doing only the denoising and is denoted as the *NMD* setting. As an additional baseline system, a speech enhancement algorithm based on minimum mean-square error log-spectral amplitude estimation [50] with the improved minima controlled recursive averaging (IMCRA) technique for noise variance estimation [40] is included.

The speech enhancement performance is evaluated using the following measures: perceptual evaluation of speech quality (PESQ) [147] in terms of mean opinion score (MOS), signal-to-distortion ratio (SDR) in dB, frequency-weighted segmental SNR (fwsegSNR) in dB, cepstral distance (CD) in dB, log-likelihood ratio (LLR) and speech-to-reverberation modulation energy ratio (SRMR) [51] in dB. The SDR was obtained using the BSS evaluation toolkit [181] and the CD, LLR, fwsegSNR and SRMR measures were obtained using the implementations provided with the REVERB challenge [102]. Higher values of PESQ, SDR, fwsegSNR and SRMR, and lower values of CD and LLR indicate a better performance. For better readability, the improvements in these measures (shown as ΔPESQ, ΔSDR, ΔfwsegSNR, ΔCD, ΔLLR and ΔSRMR) are used for comparing the results. The Δs for PESQ, SDR, fwsegSNR and SRMR are obtained by subtracting the metric obtained on the noisy data from that of the enhanced data, whereas the ΔCD and ΔLLR measures are obtained by subtracting the metric obtained on the enhanced data from that of the noisy data. In short, a higher Δ value implies a better performance for all the measures.

5.5 Results and discussion

5.5.1 Results on speech enhancement

Figure 5.2 depicts the improvement in speech enhancement quality measures for the various evaluated settings. For the PESQ measure, the NMD+R and the NMD+R+DN settings yield a similar performance. But for all the other measures, the NMD+R+DN technique significantly outperforms all the other evaluated settings. In particular, the improvement on the fwsegSNR and CD measures using the proposed method is around 1 dB and 0.2 respectively for all SNR conditions when compared to the other speech enhancement settings.

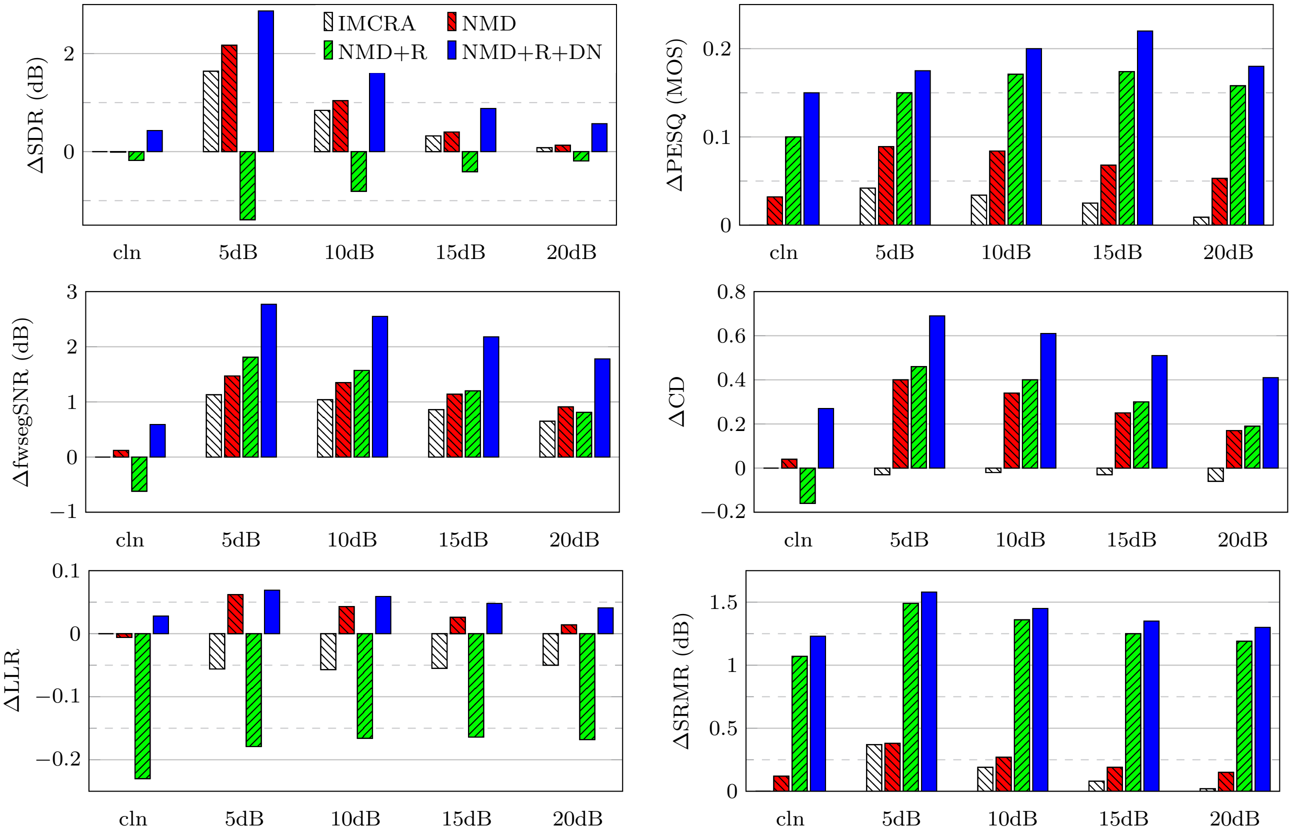


Figure 5.2: Improvements in SDR, PESQ, frequency-weighted segmental SNR (fwsegSNR), CD, LLR and SRMR obtained for various dereverberation techniques. Same legend is used for all plots.

Also notice that the IMCRA and the NMD+R settings resulted in poorer LLR measures whereas the NMD+R+DN setting yielded a slightly improved LLR over the unprocessed data. These results confirms that the proposed technique consistently outperforms the other speech enhancement settings by means of a better RIR estimate.

5.5.2 Analysis on enhancement performance

To compare the joint denoising and dereverberation performance of the various settings, the corresponding enhanced spectrograms from an arbitrarily chosen noisy reverberant recording with noise added at 20 dB are shown in Figure 5.3. The PESQ and the SDR values are also shown below each of the spectrograms. The NMD technique that does only noise suppression was able to only slightly improve the PESQ measure whereas the SDR is improved by 0.25 dB.

In the NMD+R model which captures the reverberation, both the measures are improved where the SDR is further improved by 1 dB on both the the noise-free and noisy reverberant cases. The non-negative approximations used in this work model reverberation in the magnitude STFT domain as a spectral leakage from one frame to a few upcoming frames. It can be seen that the NMD+R model reduces such spectral leakages (comparing the high frequency regions in the spectrograms for noise-free reverberant speech and the NMD+R enhanced speech) and results in a better dereverberation and speech quality. The NMD+R+DN setting is further able to reduce these spectral leakages and results in almost 1 dB SDR improvement over the NMD+R setting, thanks to a better RIR estimate.

5.5.3 Convergence of the cost function and computational complexity

In all the proposed models, the multiplicative updates are computed such that they minimise the defined cost function. This section discusses on the convergence of the cost function upon alternately applying the multiplicative updates. Figure 5.4 depicts the average cost per frame vs. the iteration count. The cost is computed for 10 randomly chosen utterances in the dataset and the total cost is averaged over the total number of frames. The costs for all the approaches converge after 60-80 iterations suggesting that the multiplicative updates indeed result in a decaying cost. Notice that the cost is not zero upon convergence since the cost function includes the sparsity penalties as well. It is also observed that the final cost of the proposed setting is higher than that

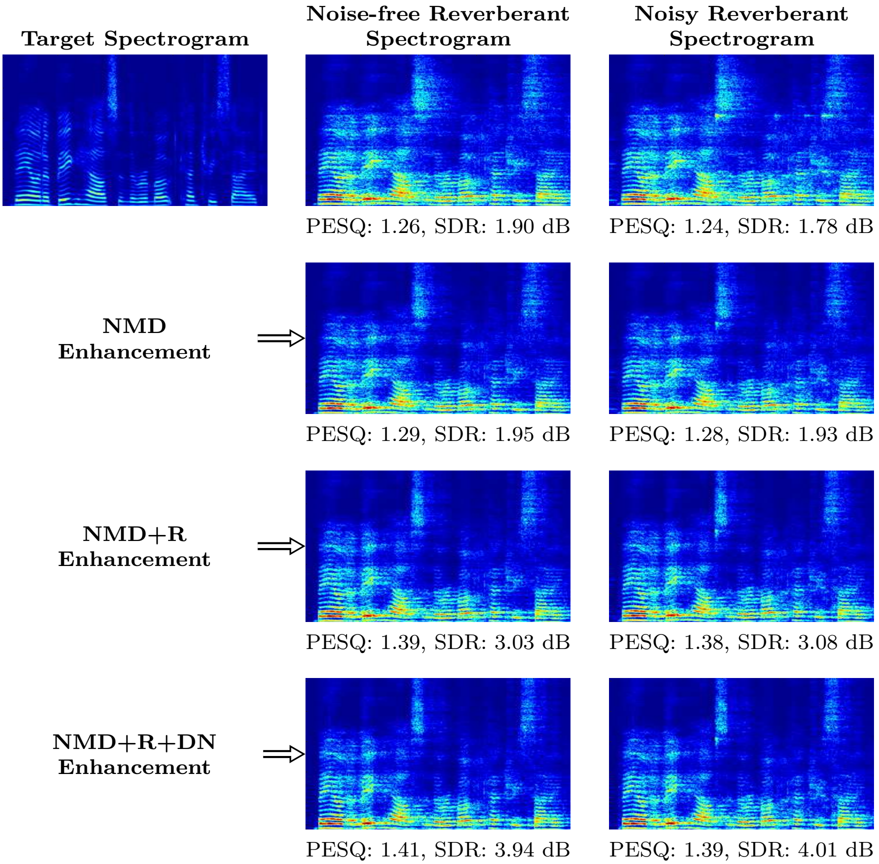


Figure 5.3: Comparison of enhanced speech spectrograms obtained for various settings for an arbitrary noisy and noise-free reverberated signal from the database corrupted with additive noise at an SNR of 20 dB. Log spectrograms are shown for a better visualisation. All figures used the same colormap. The resulting SDRs and PESQ values are also shown below each of the spectrograms.

of the NMD solution, which is due to the increased sparsity cost arising from the row normalisation of the RIR matrix.

Another point to be noted is that in the proposed model, the decay coefficients c_i are defined per RIR frame and multiplicative updates are also applied separately. In principle, after every update for c_i, the updates for activations and the RIR should be done before updating c_{i+1}. But in this work, to save execution time, the updates for c_i, $\forall i$ are done together. It can be seen from the costs curve that such a setting indeed converges and the evaluation results

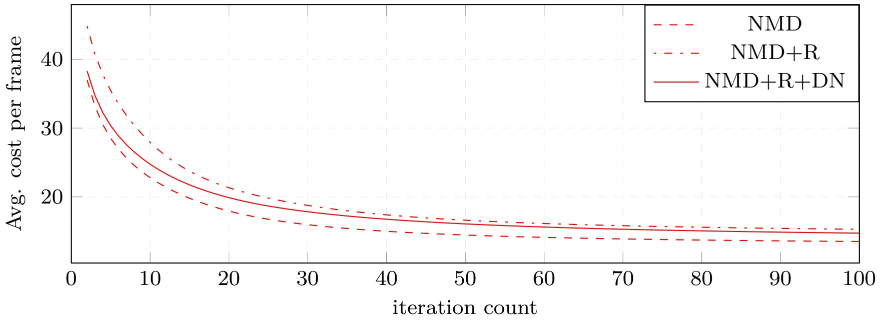


Figure 5.4: Average cost per frame vs iteration count for various enhancement techniques based on NMD.

using various speech quality measures suggest that the setting yields reliable estimates.

In order to compare the computational complexity, the total execution times for the 10 randomly chosen utterances are measured and divided by the total duration in seconds which yields the execution time required to process 1 second of data. The execution times are computed for the multiplicative updates implemented in MATLAB together with GPUs for acceleration. The average execution time to process one second of test data using the NMD, NMD+R and NMD+R+DN techniques are 1.08, 6.2 and 9.8 seconds, respectively. Such an increased computational complexity is expected since the NMD+R+DN setting has to estimate more parameters.

5.6 Conclusions and future work

This chapter proposed a novel technique to estimate the RIR with decaying norms from a noisy reverberant single-channel recording. The proposed RIR model is such that the decaying structure is inherently forced on the RIR estimate and the estimators for obtaining the decay are also presented. The estimators for obtaining the parameters in all the methods are obtained such that they minimise the Kullback-Leibler divergence between the magnitude STFT of the noisy reverberant speech and its approximation. In order to impose a decaying structure on the RIR, an existing model clamped the energies in every frequency-bin with respect to the same bin in the previous frame whereas no such bin-wise dependency is observed in a typical RIR. To overcome this, the proposed model used a novel approximation where the RIR bins are given full degree of freedom and the decaying structure of the RIR is estimated from the data itself. The model is also defined in such a manner that the decay coefficients can be estimated by applying multiplicative updates.

For evaluating and comparing the performance of the proposed technique with that of the existing techniques, a noisy reverberant dataset is artificially created using the TIMIT corpus, RIRs from the Aachen impulse response database and the challenging background noise conditions from the CHiME corpus. Evaluation results confirm the effectiveness of the proposed approach by improving the various instrumental speech quality measures including PESQ, SDR, SRMR, fwsegSNR, CD and LLR. In particular, the proposed model yielded SDR, fwsegSNR, CD and SRMR improvements of around 0.5 dB, 1.0 dB, 0.2 dB and 0.15 dB, respectively when compared to the existing model where the RIR decay is not incorporated in the approximation (NMD+R).

Investigating such a speech enhancement setting as a front-end for the ASR systems is a suggested future work. As detailed before, the proposed technique also estimates the decaying structure of the RIR from the test data itself and such a setting may be useful for estimating the T_{60} of the room using the recordings. Such an investigation might require some other changes such as modifications to how the decays are defined and using different spectral representations which is also a suggested future work.

Chapter 6

Application to Neuroscience Research

In this chapter, the proposed speech enhancement schemes are applied to the field of clinical neuroscience for the pre-operative planning on patients with brain tumor. During the pre-operative planning, a neurosurgeon has to decide if the affected brain region is essential for the major functions such as motor movement and language related processes. To identify the functional relevance of a brain region for language related processes, picture naming task together with magnetic stimulation of the relevant brain region (called transcranial magnetic stimulation or TMS) [17,74] has been effectively used. The methodology followed now is to record the responses and to manually check the accuracy and the reaction time by listening to it. However, such a process is prone to substantial intra- and inter-observer variabilities [105,171]. A novel automatic and objective evaluation routine for the picture naming task using ASR and the proposed speech enhancement schemes is developed.

This chapter is adapted from: Deepak Baby, Laura Seynaeve, Patrick Dupont, Wim Van Paesschen and Hugo Van hamme. *An automatic evaluation routine for picture naming task with transcranial magnetic stimulation using machine speech recognition.* Submitted to the Journal of Neuroscience Methods, 2016.

6.1 Introduction

Transcranial magnetic stimulation (TMS) was introduced as a non-invasive technique for magnetic stimulation of the human cortex [13]. A TMS device generates a perpendicular time-varying magnetic field that can penetrate the scalp without any attenuation. This magnetic field is produced by passing a strong current through a stimulation coil and can interact with the cortical processing non-invasively by stimulating the neuronal axons. Such a stimulation occurs since the magnetic field induce a small and short-lived current at the site of the stimulation which can either excite or inhibit the stimulated area [17,74].

TMS has been successfully used for understanding the functional relevance of the stimulated region, for example a motor evoked potential is used as a measure to study the effect of TMS when applied over the primary motor cortex [17]. However, for studying cognitive functions such as language processing, no such direct measurement is available and are typically quantified either using behavioural measurements (reaction times and accuracy of a specific task with and without TMS) or changes in neural activation [76]. This paper concentrates on the object naming task in presence of navigated repetitive TMS (rTMS) for studying the functional relevance of the stimulated region for language related tasks.

In many studies in healthy volunteers, reaction times (RTs) for the object naming task are used as a sensitive marker to pick up an effect of TMS on cognitive functioning [136,142,178,190]. TMS has also been used in patients with brain tumors to determine what areas around the tumor are involved in speaking and/or language [177]. This may aid in an objective risk-benefit assessment of a planned surgery and precisely targeted smaller craniotomies. Such preoperative planning would be a safer alternative for patients that cannot undergo awake craniotomy [140,171]. In addition, the rTMS method has been accepted by the Food and Drug Administration (FDA) of US for presurgical speech mapping [48] and it probably will have wider applications in the near future.

One of the main challenges in the object naming task is that it requires manual review [159] to evaluate the accuracy of the responses and to find the RT. Such a procedure has three main disadvantages: 1) it involves a high manual effort, 2) it is nearly impossible to measure the exact RT in the presence of rTMS noise and 3) even in no rTMS conditions, these measurements are susceptible to a high intra- and inter-observer variability [105,171]. Such variabilities reduce the repeatability, scalability and reliability of object naming tasks in general and objective evaluation routines are to be developed. Also, the methodology varies between different surgical groups applying it and no standardised procedure is defined yet.

To overcome the observer variabilities, previous studies made use of objective measurement schemes such as the responses of throat muscles, non-verbal reaction times like button-presses (typically on healthy volunteers), or by neglecting the reaction times completely to arrive at a conclusion. These approaches are still sub-optimal and are not capturing the real variable of interest. There exist some techniques to automate the evaluation task. The technique presented in [117] makes use of a video classification algorithm to evaluate the accuracy of the responses. An accelerometer-based approach to detect speech onsets is presented in [187], which still requires manual review to score the accuracy of the responses. Both these techniques required video recordings and in this work, we concentrate on a simpler naming task setup where only the audio recordings that are started simultaneously with the stimuli onset are required. Such a setting is cheaper, requires far lesser storage and is also advantageous in scenarios where the subject is not comfortable with recording the video.

In order to evaluate the object naming task using the audio recordings, an automatic speech recogniser setting is employed which can recognise the word being said in the recording together with its temporal information. Since it is noticed that the recogniser output may go wrong in presence of rTMS noise, a noise suppression system is also used. In this paper, we propose a novel automatic evaluation scheme for the object naming task that can generate text files indicating the accuracy and the RT of every response. One of the main challenges in automating the object naming task is when the subject gives a synonym, hypernym or hyponym as the response. The proposed setting also includes a functionality to add acceptable synonyms for every response, thereby increasing the flexibility and objectivity of the setting.

6.2 Materials and methods

6.2.1 Experimental design

To evaluate the proposed setting, an object naming task with 140 objects was conducted on 8 healthy Dutch speaking subjects (3 males and 5 females). The subjects were asked to name black-and-white line drawings based on the Snodgrass and Vanderwart picture set [169]. The stimuli were presented and the corresponding responses were recorded using Presentation version 14.8 (Neurobehavioralsystems, USA). Every stimulus lasted for 3 seconds and the

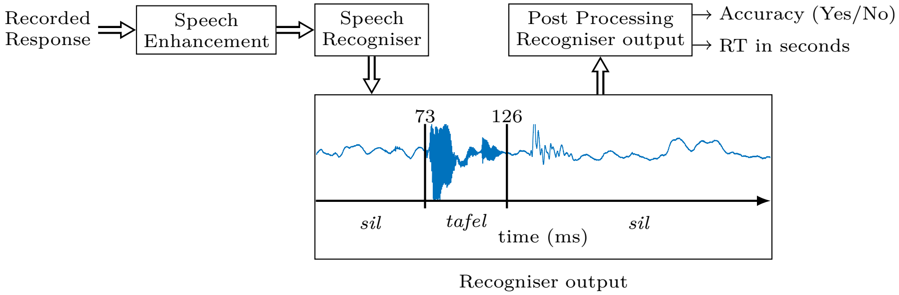


Figure 6.1: Block diagram overview of the proposed automatic evaluation routine. An example output of the recogniser is shown for the stimuli *tafel* (table in English). *sil* denotes the silence regions as given by the speech recogniser.

corresponding recording started simultaneously with the projection of each picture. The recordings taken from the 8 subjects are denoted as *S1* to *S8* and combination of all the recordings are denoted as *S1-S8*.

The dataset thus generated contains 1120 recordings. The RTs for these recordings were manually annotated by a speech technology expert using Praat [19] software with an accuracy of upto 10 ms, since the resolution of the proposed automatic routine is also 10 ms. To obtain the RTs as correct as possible, additional information such as spectrograms, pitch contours and voice activity detection are used. This dataset is denoted as *Noise-free*. Notice that the term noise-free denotes that the recordings do not contain the rTMS noise and they still might contain other recording room noises like fan-noise, breathing and patient movements.

The recordings in presence of rTMS were artificially created from the already collected noise-free recordings to reduce the manual effort in annotating the noisy data that are also error-prone since it is difficult to manually find the speech onset in presence of the rTMS noise. Moreover, a reliable ground truth is required to test the RT measurement accuracy of the proposed routine. The noisy recordings were therefore synthetically created using a two step procedure. First, the noise-free data was delayed for a random time period by padding zeros at its beginning to model the delay in RT in the presence of rTMS. Notice that the RT measurement is more challenging in scenarios where the speech onset overlaps with the rTMS noise. To simulate such conditions, we chose a random delay between 50 and 300 ms and these delays were added with the original RT for comparison with the estimated RT. In principle we can choose a larger delay than 300 ms, but that may make the speech onset not overlapping with the rTMS noise and makes the problem less challenging. These delayed responses were then added with a previously recorded rTMS noise at the beginning of the recording. The rTMS noise was recorded during a TMS stimulation using a figure-8 coil (Magstim, United Kingdom) which contains 5 consecutive pulses of 5 Hz. To simulate a variety of rTMS conditions, the rTMS noise was recorded by placing the coil on different locations of the head and a random noise recording is chosen at random during the artificial generation of the database in the presence of rTMS. The database thus simulates various rTMS conditions where the coil is placed on different locations of the head. This procedure generated 1120 noisy delayed recordings that contain the rTMS noise at random positions and the dataset is denoted as *Noise + Delay* set.

In addition to the above two sets, a test set where no delay is present was also created artificially. This set was created by adding rTMS noise to the beginning of the noise-free set to model the scenario where a sham coil is used and this set is denoted as *Noisy* set.

6.2.2 Proposed evaluation routine

The proposed automatic routine is derived from an automatic speech recognition system presented in [47]. Figure 6.1 depicts the outline of the proposed automatic evaluation routine. The speech recogniser takes in the recorded response and generates the recognised word together with its timing information. This output is then post-processed to check if the response is correct and if yes, it also outputs the RT. A response is marked as a correct response if the recogniser output matches the expected picture name or one of its synonyms. The internals of each block are described below.

The statistical speech model required for the speech recogniser is trained using the Flemish recordings contained in the CGN corpus [160]. A detailed description of the speech recogniser can be found in [47]. Only the specific details that are relevant for the object naming task are summarised here. Based on the input recording of the response, the speech recogniser can handle the following conditions.

1. *Model the expected response* : The speech recognition stage takes in one recording at a time and it makes use of the expected response to model the word in the recording. For this, the internal parameters of the speech recogniser are updated for every response so that a higher weight is given to the expected stimuli for it to be recognised. Notice that a much higher weight will produce errors since it increases the chance of some background noise or a wrong response being recognised as the expected response and therefore a compromise must be pursued. In this task, the setting is adjusted so

that all the wrong or no response events are detected correctly that can be applicable to rTMS tasks where detecting such events are more important. Notice that such a setting may reduce the recognition accuracy for correct responses. The automatic routine is flexible so that the observer can vary the parameters if a higher detection accuracy on correct responses is required.

A sample output for the input recording for the stimulus *tafel* (table in English) can also be seen in Figure 6.1. The recogniser output can contain the following: the expected stimulus, *sil*, *garbage* and/or *stut*. The *sil* output refers to the regions recognised as silence. The *stut* output (not shown in Figure 6.1) models stuttering which is detailed below. The *garbage* output is comprised of phone sequences that are either random corresponding to background noises or corresponding to a wrong response given by the subject. The recogniser will yield the *garbage* output under the following scenarios: wrong response from the subject, no-response event in presence of some noise or if the recogniser fails to identify a correct response.

2. *Modelling stuttering* : During the picture naming task, it is observed that sometimes the subject stutters or hesitates before saying the actual response. Stuttering corresponds to repeating the parts of the stimuli name before uttering the complete name. The recogniser can identify such events by explicitly modelling the stuttering (the output in this case is *stut*) and by modelling other hesitation sounds as *garbage*. The *stut* is modelled by including phone sequences that skip parts of the stimuli name in the language model graph [47].

3. *Synonyms* : Another challenge in automating the naming task is when the subject says a synonym of an expected response. As described before, the speech recogniser only outputs the expected response and any other sound or response is modelled as *garbage*. The speech recogniser is modified so that the synonym is also considered as an additional expected response. Notice that not all the synonyms will be included in the basic evaluation package. The observer has to listen to those recordings that are marked as wrong responses by the automatic routine and if the subject used a synonym, a separate routine is used to add synonyms to the existing setting. Such a routine adds flexibility to the system as the set of synonyms to be added is decided by the observer and the synonyms list has to be updated until all the acceptable synonyms are added. Notice that after adding a synonym, the automatic routine has to be executed again in order to update the RT and accuracy results on the corresponding responses. After evaluating on a few patients, the synonyms list is expected to converge and no more updating is required.

4. *Timing information* : The recogniser can also yield the timing information of the various outputs (expected response, *sil*, *stut* and *garbage*) along the

length of the utterance upto an accuracy of 10 ms. The details of how these temporal alignments are obtained can be found in [47].

In order to reduce the manual effort in going through all the recogniser outputs corresponding to every response, a post processing stage is added which generates a text file that summarises the accuracy and RTs of all the responses recorded in a session. The functions of the post-processing module are:

1. *Accuracy* : The post-processing routine yields 'Yes' or 'No' output depending on whether the expected response is present in the recogniser output by means of a string comparison script.
2. *RT estimate* : If the recogniser output contains the expected response, the post-processing stage next extracts the timing information corresponding to the expected response. It is observed that the recogniser alignments will contain some bias depending on the beginning sound (phoneme) of the response. These biases are also not expected to be subject dependent since the recognition engine used is designed for speaker independent speech recognition. Therefore the biases for every class of phonemes are computed experimentally (based on the subset S1 and S2) and are compensated by the post-processing stage to yield RTs as close as possible to the manually obtained RTs (more details can be found in Section 6.3). Notice that, if the user is interested only in the difference in RT in two scenarios, this bias correction is not necessary since these systematic biases will be cancelled out when the difference in RT is computed.

6.2.3 Speech enhancement front-end

Since the speech models used by the recogniser is based on the CGN corpus which contains noise-free recordings, such a recogniser is sensitive to distortions introduced to the recording due to the rTMS noise and other background room noises. Therefore, a speech enhancement front-end that suppresses the noises and reduce distortions may improve the recogniser performance. Since the responses are recorded in a controlled environment where the patient's language (Dutch) and the types of background noise (rTMS noise, room noises, etc.) are known a-priori, a speech enhancement setting that makes use of this knowledge can be used to achieve a better noise suppression. In this paper, the speech enhancement technique detailed in [9] is used. Such a speech enhancement setting is previously shown to significantly improve the performance of a state-of-the-art speech recogniser [5]. The enhanced recordings are fed to the speech recogniser system for evaluation.

6.2.4 Splitting the data into subsets

It is observed that the speech recogniser yields temporal alignments with differing offsets (or biases) depending on the beginning phoneme of the response. In order to correct these biases, the test set is divided into 7 subsets based on the starting phoneme. The subsets are named after the class of the starting phoneme that are; vowels (VOW), voiced stops (VSTP), unvoiced stops (USTP), nasals (NAS), sibilant fricatives (SIB), non-sibilant fricatives (NSIB) and liquids (LIQ). The stimuli present in each subset can be found in B. The analysis and the RT bias correction are done for every subset separately to study and correct the systematic effects of the recogniser.

6.2.5 Evaluation metric

In the picture naming task, the responses are grouped into two categories: events where the response is correct and where there is a wrong or no response. The performance of the automatic routine in predicting the accuracy of the response is computed by looking at the number of cases where these events are detected correctly. In the picture naming task, the wrong or no response events are expected to be detected at an accuracy of close to 100 %.

For the RT prediction performance, the error in RT with respect to the manually obtained RT is computed using $e_{RT} = RT_{est} - RT_{true}$, where RT_{est} and RT_{true} are the estimated and the actual RTs, respectively. Then an evaluation metric E_t is defined, which is the percentage of cases where the prediction error is between $\pm t$ ms, i.e., percentage of estimated RTs satisfying $-t\ ms \leq e_{RT} \leq t\ ms$. Notice that the automatic routine does not yield RTs corresponding to the recordings that are detected as wrong responses and these are omitted while calculating and analysing the E_t and e_{RT} measures.

6.3 Results

6.3.1 Evaluating the accuracy of the responses

The response detection accuracy of the automatic routine evaluated on data collected from various subjects are tabulated in Table 6.1. The overall accuracy

Table 6.1: Accuracy of the automatic routine in identifying the type of responses from the subjects under different test conditions with and without the speech enhancement (SE) front-end.

Subject	Sex	Condition	Correct Responses			Wrong/No Responses			Overall	
	(F/M)		Occuring	Detected		Occuring	Detected		Accuracy (%)	
				No SE	SE		No SE	SE	No SE	SE
S1	M	Noise-Free	135	125	126	5	5	5	92.9	93.6
		Noisy		132	132		5	5	97.0	97.9
		Noise + Delay		131	130		5	5	97.1	96.4
S2	F	Noise-Free	124	111	113	16	16	16	90.7	92.1
		Noisy		110	116		16	16	90.0	94.3
		Noise + Delay		111	114		16	16	90.7	92.9
S3	M	Noise-Free	134	109	112	6	4	4	80.7	82.9
		Noisy		114	114		4	4	84.3	84.3
		Noise + Delay		110	115		3	5	80.7	85.7
S4	F	Noise-Free	139	118	119	1	1	1	85.0	85.7
		Noisy		117	119		1	1	84.3	85.7
		Noise + Delay		121	123		1	1	87.1	88.6
S5	F	Noise-Free	133	125	126	7	7	7	94.3	95.0
		Noisy		123	123		7	7	92.9	92.9
		Noise + Delay		120	125		7	7	90.7	94.3
S6	M	Noise-Free	131	124	124	9	9	9	95.0	95.0
		Noisy		119	121		9	9	91.4	92.9
		Noise + Delay		115	120		9	9	88.6	92.1
S7	F	Noise-Free	136	99	106	4	4	4	73.6	78.6
		Noisy		105	105		4	4	77.9	77.9
		Noise + Delay		112	112		4	4	82.9	82.9
S8	F	Noise-Free	137	122	123	3	2	2	88.6	89.3
		Noisy		120	124		2	2	87.1	90.0
		Noise + Delay		122	122		2	2	88.6	88.6
Overall		Noise-Free	1069	933	949	51	48	48	87.7	89.1
		Noisy		940	954		48	48	88.3	89.6
		Noise + Delay		942	961		47	49	88.4	90.4

in % in the last column is computed as:

$$\frac{\begin{array}{c}\text{number of correct responses detected +}\\ \text{number of wrong/no responses detected}\end{array}}{140}\times 100. \tag{6.1}$$

Out of the 1069 correct responses in the noise-free scenario, the automatic routine is able to detect 933 and 949 of the cases without and with speech enhancement, respectively. The 51 wrong/no response events are comprised of 28 wrong responses and 23 no-responses, out of which the evaluation routine correctly identified more than 25 wrong responses and all the 23 no-response events.

In general, the speech enhancement front-end is able to improve the response detection accuracy especially under noisy conditions (except for S1). As noted before, the speech recogniser is trained using the Dutch speech component of the CGN corpus and that model is used to recognise a response that is recorded with a different microphone in different room conditions. Some of the detection errors are caused by this mismatch. The speech enhancement front-end was able to improve the accuracy even in noise-free conditions, thanks to its ability to reduce the mismatches between the test and training data as pointed out in [5]. The automatic routine is also robust to false-detection of responses in noisy conditions yielding an overall detection accuracy of 92.4% on wrong-response events and 100% for no-response events with speech enhancement under the noise+delay condition.

6.3.2 Analysis on RT measurement

In order to analyse the accuracy in predicting the RT, the E_t measures are plotted for varying t from 0 ms to 430 ms (which was the highest error) with and without speech enhancement in Figure 6.2. For ease of visualisation, the E_t measures are obtained on the S1-S8 dataset. The speech enhancement front-end was able to significantly improve the RT prediction in noisy conditions. The automatic routine together with speech enhancement yielded an absolute e_{RT} of less than 40 ms for more than 88% of the cases under all conditions.

It is also observed that some responses are more prone to yield higher e_{RT} (e.g. oog, vis, vos, fiets, etc.) and it is possible to further improve the RT prediction performance by omitting these from the stimuli set.

In order to analyse the RT prediction on various subsets on the test data, the E_t measure at $t = 40$ ms is given in Table 6.2. The value $t = 40$ ms is chosen since it is the knee-point in the E_t curve shown in Figure 6.2 where the curve reaches a plateau. As described before, the subsets were created based on the type of the first phoneme in the expected response. The automatic routine was found to be robust to rTMS noise when the responses are starting with a vowel or nasal sound. The RT prediction on responses starting with a voiced stop sound (/b/, /d/) or a sibilant fricative sound (/s/,/z/) were found to be the most error-prone in presence of rTMS and a speech enhancement front-end

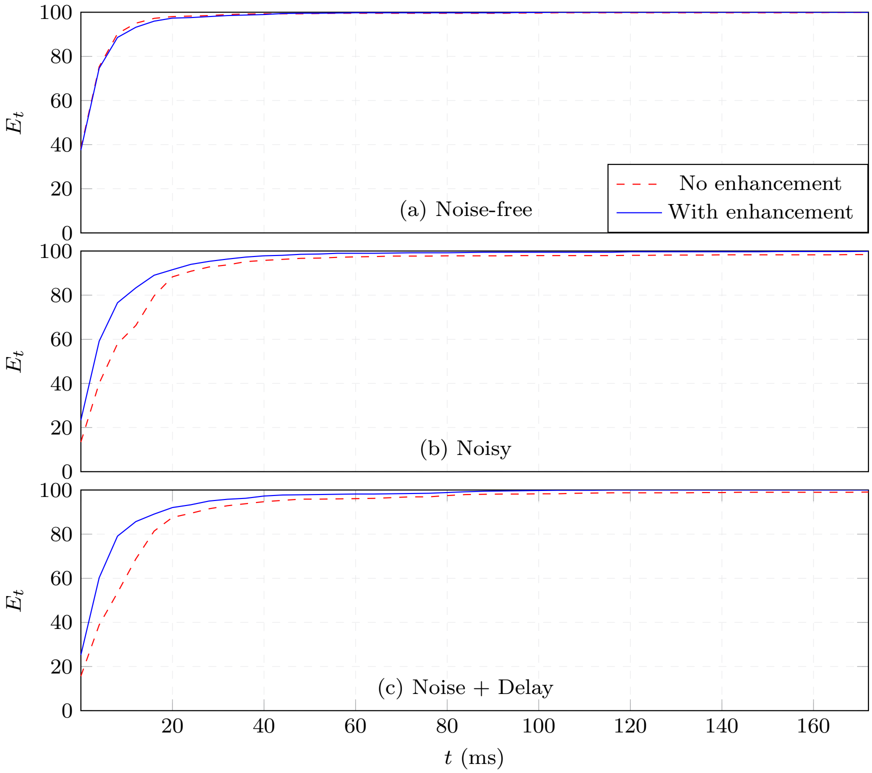


Figure 6.2: E_t (percentage of cases where the prediction error is between $\pm t$ ms) plot obtained on the complete test set (S1-S8). All figures uses the same legend and x axis values.

was able to reduce many of these errors. However, for responses starting with an unvoiced stop (/k/,/p/,/t/), the speech enhancement front-end sometimes resulted in a poorer RT prediction. This is mainly because the unvoiced stop sound is confused with the impulsive rTMS noise and the speech enhancement scheme suppresses a part of the actual unvoiced sound as well resulting in a wrong RT. It is also observed that the speech enhancement front-end was not able to improve the RT prediction on the LIQ subset (/y/,/l/) because the speech onset of such sounds are rather vague since the beginning of these can be elongated. For such cases, even manual annotation is really difficult and an objective prediction setting would be more desirable.

As an additional evaluation metric, the mean and standard deviation of the absolute e_{RT} are also shown in Table 6.3. It can be seen that the proposed

Table 6.2: E_{40} (percentage of cases where the prediction error is between ±40 ms) values obtained in % for various test cases.

		Without Speech Enhancement								With Speech Enhancement							
Subject	Condition	VOW	USTP	VSTP	SIB	NSIB	LIQ	NAS	ALL	VOW	USTP	VSTP	SIB	NSIB	LIQ	NAS	ALL
S1	Noise-free	100	100	92.3	100	96.6	100	100	98.4	100	97.1	85.7	95.5	96.2	100	100	96.0
	Noisy	100	85.3	81.2	86.4	88.9	100	100	89.3	100	97.1	80.0	100	96.0	90.0	100	95.3
	Noise+Delay	100	91.4	68.8	86.4	92.3	100	100	90.0	100	97.1	66.7	100	92.6	100	100	93.8
S2	Noise-free	100	93.3	100	100	95.7	100	100	97.3	93.3	100	100	100	100	100	100	99.1
	Noisy	100	78.6	75.0	90.0	79.2	87.5	100	84.5	100	77.4	92.3	90.0	80.8	87.5	100	86.1
	Noise+Delay	100	90.0	92.3	66.7	78.3	87.5	100	85.6	100	83.3	100	80.0	80.0	87.5	100	86.8
S3	Noise-free	93.8	95.8	92.3	95.2	96.0	100	100	95.4	100	96.0	92.3	95.2	96.3	100	100	96.4
	Noisy	81.2	74.1	38.5	90.0	78.6	100	100	77.2	93.8	96.3	66.7	90.0	78.6	83.3	100	86.7
	Noise+Delay	100	79.2	66.7	90.9	77.8	100	75.0	83.6	100	96.2	75.0	90.9	82.1	100	75.0	89.6
S4	Noise-free	100	96.6	100	81.0	87.5	100	100	93.2	100	100	92.3	77.3	78.3	100	100	90.6
	Noisy	100	86.7	53.3	60.0	77.3	88.9	100	78.6	100	90.0	61.5	70.0	75.0	100	100	83.1
	Noise+Delay	100	87.1	71.4	85.0	63.0	87.5	100	81.7	100	93.8	64.3	80.0	64.0	100	100	83.2
S5	Noise-free	100	96.8	100	100	96.3	100	100	98.4	100	96.8	100	100	96.3	100	100	98.4
	Noisy	100	86.2	58.3	100	75.0	100	100	87.0	100	88.9	66.7	100	78.6	100	100	89.3
	Noise+Delay	93.8	86.2	66.7	87.0	71.4	100	100	83.3	94.1	93.5	75.0	95.7	67.9	100	100	87.0
S6	Noise-free	88.9	100	93.8	100	96.4	100	100	96.8	88.9	96.6	93.8	95.0	96.3	100	100	95.2
	Noisy	100	92.9	53.3	83.3	82.1	100	75.0	84.9	86.7	96.4	66.7	94.7	82.1	100	50.0	86.6
	Noise+Delay	94.1	85.2	46.2	88.2	82.1	100	75.0	82.6	93.8	96.4	71.4	88.9	86.2	100	75.0	89.0
S7	Noise-free	100	91.3	90.9	84.2	90.9	100	100	91.9	100	86.4	91.7	93.8	76.2	83.3	100	88.4
	Noisy	100	84.6	80.0	84.2	76.2	100	100	86.0	100	88.9	68.8	80.0	69.2	100	100	82.6
	Noise+Delay	92.3	88.9	60.0	76.2	79.2	100	100	82.1	92.3	85.7	73.3	94.4	88.5	100	100	88.4
S8	Noise-free	100	100	100	100	100	88.9	100	99.2	100	100	92.9	91.3	92.6	77.8	100	94.3
	Noisy	100	90.3	66.7	87.0	72.0	87.5	100	84.9	100	96.3	72.7	78.3	87.5	87.5	100	88.4
	Noise+Delay	94.1	83.3	76.9	82.6	61.5	88.9	75.0	79.5	100	96.7	92.9	83.3	70.8	66.7	75.0	86.1
ALL	Noise-free	97.7	96.9	96.3	95.2	95.1	98.5	100	96.5	97.7	96.9	93.4	93.4	92.0	95.6	100	95.0
	Noisy	97.7	85.0	63.6	85.5	78.8	95.8	96.3	84.1	97.7	91.3	72.0	88.0	80.9	94.3	93.1	87.4
	Noise+Delay	96.9	86.7	68.5	83.1	75.6	95.5	90.3	83.6	97.6	92.9	76.9	89.2	79.2	94.0	90.3	88.0

automatic routine is able to estimate the RT with a small error mean and standard deviation. The speech enhancement front-end is able to reduce the standard deviation in all conditions in addition to making the mean absolute error closer to zero. Notice in particular that the noisy test case without any speech enhancement had a high standard deviation of 101.4 ms on the overall

Table 6.3: The mean (μ) and standard deviation (σ) of the absolute e_{RT} in ms obtained on various test sets under varying testing conditions with and without the speech enhancement (SE) front-end.

Subject	Condition	No SE		With SE	
		μ	σ	μ	σ
S1	Noise-free	11.0	12.7	10.9	11.8
	Noisy	27.3	88.2	13.9	20.9
	Noise+Delay	30.1	92.2	15.2	24.3
S2	Noise-free	12.5	40.3	9.4	10.8
	Noisy	36.8	110.5	20.8	30.2
	Noise+Delay	26.5	65.7	22.2	44.4
S3	Noise-free	20.7	78.3	13.4	26.4
	Noisy	57.2	156.7	21.3	23.8
	Noise+Delay	44.8	124.2	20.9	32.3
S4	Noise-free	13.5	21.3	16.6	22.7
	Noisy	31.8	66.5	25.1	36.4
	Noise+Delay	30.8	53.4	27.1	44.8
S5	Noise-free	11.0	12.9	10.3	12.7
	Noisy	27.2	62.6	22.7	39.1
	Noise+Delay	27.8	44.1	21.2	30.2
S6	Noise-free	12.0	23.0	13.5	25.1
	Noisy	28.5	54.3	25.1	40.1
	Noise+Delay	32.6	55.7	24.7	42.9
S7	Noise-free	17.3	33.4	21.1	34.7
	Noisy	47.1	167.4	42.3	147.9
	Noise+Delay	37.8	91.2	21.3	32.5
S8	Noise-free	7.9	10.8	13.1	26.1
	Noisy	23.7	32.6	20.1	29.4
	Noise+Delay	34.4	61.4	22.4	38.7
Overall	Noise-free	13.0	34.9	13.4	23.2
	Noisy	34.5	101.4	23.8	60.1
	Noise+Delay	33.0	77.1	21.8	36.8

S1-S8 set which is reduced to 60.1 ms by using the speech enhancement front-end.

6.4 Discussion

The proposed routine can automatically generate the RT and test the accuracy of an object naming task from the recorded responses. A speech enhancement front-end is also employed to improve the accuracy especially in presence of the rTMS noise. The proposed routine was evaluated on responses obtained from eight subjects with 140 stimuli each. To simulate the responses under rTMS, two test sets were artificially created, one for sham rTMS pulses where no delay in RT is expected a and second one for the actual rTMS condition with both delay and rTMS noise. To analyse the prediction errors, the RTs given by the automatic routine were compared with the manually obtained RTs. The absolute error was found to be less than 40 ms in 95.0% and 88.0% of the responses obtained without and with rTMS, respectively. The automatic routine was able to yield absolute errors with small mean and standard deviation.

Our setup can be a good alternative for the manual annotation of the recorded response to obtain the RT especially in presence of the rTMS noise. The routine yields an objective estimate of the RT thereby reducing the subjectivity and increases the repeatability and reliability of the object naming experiments. The proposed framework is flexible so that the observer can add synonyms to the setting. The observer can decide whether a response is a synonym or not depending on the level of abstraction needed for the experiment. This improves the flexibility of the routine to suit the specific requirements of the task.

For the RT prediction routine, the speech enhancement front-end is found to improve the prediction accuracy in general, especially in reducing the number of outliers having high RT prediction errors. An analysis based on the type of starting phone for the automatic RT measurement is also made and the paper also gives advice on the type of stimuli to be chosen in order to get a reliable performance using the proposed routine. To summarise, stimuli starting with a vowel and nasal sound are more preferred for a better RT prediction. For stimuli starting with a stop consonant (/b/, /d/, /k/, /p/, /t/), the RT measurement in presence of the rTMS is difficult in general since the rTMS noise has very similar characteristics as stop consonants. The speech enhancement front-end was able to reduce RT errors in unvoiced stop consonants. For objects starting with a liquid sound, an objective RT measurement system may be more desirable.

Notice that the post-processing stage also corrects the RT prediction offset depending on the type of the starting phone. These offsets are observed to be pretty consistent for a given starting phone. Therefore, such an offset correction will not be needed when the objective of the task is to measure the difference in RTs with and without rTMS as these offsets will be cancelled out.

Notice that the recordings were obtained using a head-mounted microphone which has higher chances of recording the breathing and microphone tapping sounds. The automatic routine may be further improved by using a microphone mounted on a stand with pop filter, thereby reducing distortions and mismatches between the test and training recordings. The speech enhancement setting is also flexible enough to incorporate knowledge of other noise sources so that such noises can also be suppressed by the setting.

6.5 Conclusion

In this work, we proposed an automatic routine for testing the accuracy and obtaining the RT for an object naming task in presence of rTMS. The approach made use of an automatic speech recognition system to obtain the RTs and the response. A speech enhancement front-end is also employed to improve the system accuracy in presence of the rTMS noise. The automatic routine is found to yield small prediction errors with small standard deviation which can be reliably used for automating the object naming task. The method can produce tables indicating the accuracy and the RT of the response thereby adding reliability, objectivity and repeatability to the rTMS object naming analysis. The proposed setting is also flexible as it has the functionality to add synonyms depending on the task definition.

Chapter 7

Conclusions and Future Work

This chapter concludes the thesis with a concise review of the original contributions of this work with some suggestions for future research.

7.1 Original contributions

This thesis focused on speech enhancement using the spectrogram factorisation techniques NMF and NMD. Several novel approaches have been proposed that are shown to be beneficial for improving speech intelligibility and as a front-end for ASR systems.

- **Coupled dictionaries for exemplar-based speech enhancement**
  The idea of using coupled dictionaries to obtain a better mapping from one feature space to the STFT space is one of the main contributions of this thesis. The proposed setting can simultaneously benefit from a better speech enhancement performance of one exemplar space and can directly map the resulting estimates to the full resolution frequency domain for a better set of filter weights. The setting was shown to be effective across different databases such as AURORA-2, AURORA-4 and CHiME-3. Chapter 2 details the proposed setting together with an extension to learn the coupled noise atoms online from the test data.

- **MS features for exemplar-based speech enhancement**
  The perceptually motivated MS features were introduced to the exemplar-based speech enhancement framework. One of the main drawbacks

of the MS features was that there is no direct mapping to the STFT domain, despite having a good speech and noise separation capability. Using the previously proposed coupled dictionary-based speech enhancement setting, a reliable mapping from the MS domain to the STFT domain was achieved. The MS features were also shown to yield superior improvements in speech enhancement quality and a better ASR performance in AURORA-2 and AURORA-4 databases (in Chapter 2). It is also observed that the MS exemplar space, mostly due to its higher dimensionality, does not generalise well to unseen noise scenarios.

- **Exemplar-based speech enhancement as a front-end to neural network-based ASR settings**
  This work showed that using an exemplar-based speech enhancement setting can improve the accuracy of neural network-based ASR settings, while most of the traditional speech enhancement settings such as spectral subtraction were found to be ineffective. The evaluations were performed using a DNN-based ASR setting on the AURORA-4 database, and both DNN- and CNN-DNN-based decoders on CHiME-3 (in Chapter 2).

- **Hybrid exemplar spaces for NMF-based decomposition**
  Since it was observed that different exemplar spaces behave differently in presence of different noise scenarios, a method to jointly obtain the decomposition across different exemplar spaces was proposed in Chapter 3. The method was shown to result in a better speech and noise separation especially in unseen noise conditions. A novel switching mechanism to obtain an improved performance was also proposed.

- **DNN training using MS features**
  DNN-based acoustic modelling has been shown to benefit from richer feature representations and to be less sensitive to the dimensionality of the input features. A DNN trained on MS features was investigated for ASR tasks on the benchmark databases TIMIT and AURORA-4 in Chapter 4. A comparison between different features, such as PLP, FBANK, GBFB and STFT for DNN training was presented. The perceptually motivated MS and GBFB features were shown to yield a better performance over the conventional features.

- **Joint denoising and dereverberation**
  Another important contribution of this thesis is to incorporate a reverberation model in the NMD-based speech enhancement framework. The proposed approach explicitly modelled a decaying RIR and provided multiplicative updates to jointly estimate the RIR, its decay together with the speech and noise activations. This setting was shown to yield a better

speech enhancement quality based on various instrumental objective quality measures in Chapter 5.

- **Application to clinical neuroscience**
  The final contribution of this thesis was to develop an automatic evaluation routine for a picture naming task on brain-tumor patients. The software provides an objective evaluation measure of the reaction times in addition to checking the accuracy of the response. The evaluations were performed on the recordings obtained from 8 subjects and were compared against the manual annotations. The coupled dictionary-based speech enhancement setting was also effectively used to suppress the noise from the magnetic stimulation device and other room noises for a better reaction time estimation (in Chapter 6). This is a pioneering work where the concepts of ASR and speech enhancement are applied to automate the evaluation of a picture naming task in clinical neuroscience that aids pre-operative planning and studying the functional relevance of the stimulated brain region for language related processes.

7.2 Suggestions for future work

This section suggests a few possible extensions to the proposed denoising and dereverberation algorithms for a better speech and noise separation, better noise modelling when the noise characteristics are not known a-priori and a better ASR performance.

- **Offline learned speech and noise dictionaries**
  While this thesis concentrates mostly on exemplar-based techniques which use randomly sampled spectro-temporal patches as exemplars stored in overcomplete dictionaries, another research area is to learn compact representations of speech and noise using NMF with an undercomplete dictionary. It is recently shown in [109] that properly learned dictionaries with tuned sparsity can outperform overcomplete dictionaries for speech enhancement under certain noise scenarios. Using undercomplete dictionaries also reduces the computational complexity. Since the proposed coupled dictionary-based approach also can reduce the computational complexity by obtaining the decomposition in lower dimensional exemplar spaces and yield a better mapping to the STFT domain, it is worth combining these two paradigms that learn coupled dictionaries. Thus during the learning phase, the spectro-temporal patches from one feature space and the corresponding STFT patch are concatenated and a compact representation using NMF (or NMD)

is learned. Then during the enhancement phase, the factorisation is obtained using the learned dictionary corresponding to the lower dimensional feature space and the mapping to the STFT space is obtained using the jointly learned STFT dictionary. This approach can also use some insights from Chapter 3 where hybrid exemplar spaces are investigated such as relative weighting of feature spaces during the learning phase.

- **Discriminative online learning of noise and/or speech atoms**
  Most of the work in this thesis considers the scenario where the noise examples are known a-priori. But such a scenario may not be always realistic since the noise repository may not capture all the noise characteristics and hence the adaptive learning of a few atoms to model these unseen noise characteristics are found to be effective. This thesis describes such an online noise learning in Chapter 3. However, such a model is based on the assumption that the speech dictionary sufficiently captures all the speech characteristics. Therefore, the number of atoms to be learned is crucial in the performance of such a setting since if there are too many adaptive atoms that are learned online, they might start modelling speech as well which can degrade the performance. Therefore, one possible future research direction is to add some constraint so that the learned noise atoms have different characteristics from speech by means of some discriminative constraint. Some examples of prior work on NMF where such approaches are applied to source separation can be found in [189,192]. Investigating better discriminative criteria, extending it to the coupled dictionary framework and investigating such models on MS features are also possible research directions.

  Another possibility comes under the blind source separation scenario where both the speech and noise atoms are learned online. There is scope for extensive investigations using coupled dictionaries in this case as well. These can also be derived from the aforementioned online learning of noise atoms.

- **Perceptually motivated features for neural network training**
  In Chapter 4, it is observed that the perceptually motivated MS and GBFB features result in a better acoustic modelling using a state-of-the-art DNN setting when compared to the traditional GMM-based approaches. However, these features are sensitive to reverberation and to the type of background noise. Therefore some feature enhancement schemes, say the proposed exemplar-based techniques, may benefit such systems. Such an evaluation using enhanced features is not covered in this thesis. Application of neural network based denoising settings such as stacked autoencoders [182] is also a suggested future work.

- **Better RIR models for dereverberation**
  This work proposed a novel algorithm to jointly estimate the RIR, its decay, and the activations with decaying constraint on the RIR. While the proposed constraint is shown to yield an improved performance, further constraints can be introduced for a better RIR estimate based on the observations made from the STFT of a typical RIR shown in Chapter 5. For instance, the RIR spectrum decays faster at higher frequencies when compared the lower frequencies. Incorporating such constraints into the system without affecting the convergence properties of the cost function can result in further improvements.

- **Improvements to the picture naming task evaluation routine**
  The automatic evaluation routine presented in Chapter 6 makes use of a general purpose GMM-based acoustic model trained using the CGN corpus. Since it is already shown that DNN-based acoustic modelling outperforms GMM-based models, developing a DNN-based acoustic model for the automatic routine is a suggested future work. Another possible extension is to apply speaker adaptation on such models since it is possible to record a few sentences from the patient before the picture naming task. Thus the model can adapt to the speaker variations and may result in better recognition accuracies.

Appendix A

Derivation of multiplicative updates

This appendix details the derivation of multiplicative updates for obtaining the estimates discussed in chapter 5. The updates are derived such that they minimise the Kullback-Leibler divergence (KLD) between the noisy reverberant speech $\mathbf{Z} \in \mathbb{R}_{+}^{B \times F}$ and its approximation $\tilde{\mathbf{Z}} \in \mathbb{R}_{+}^{B \times F}$ in the magnitude STFT domain. The KLD between $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ is defined as:

$$\begin{aligned}\mathcal{D} &\triangleq D_{KLD}\left(\mathbf{Z} \| \tilde{\mathbf{Z}}\right)\\ &= \sum_{f=1}^{F} \sum_{b=1}^{B} \left(\mathbf{Z}(b,f) \log \frac{\mathbf{Z}(b,f)}{\tilde{\mathbf{Z}}(b,f)} - \mathbf{Z}(b,f) + \tilde{\mathbf{Z}}(b,f) \right)\end{aligned}$$

In general, the cost function used for solving the dereverberation problem in this work is:

$$\mathcal{C} = \mathcal{D} + f(\Theta) \tag{A.1}$$

where, Θ denotes the variable set used in the approximation $\tilde{\mathbf{Z}}$ and $f(\Theta)$ denotes some additional constraint imposed on these variables so that a meaningful approximation is obtained. The optimal estimate for a variable θ in the parameter set Θ is obtained by iteratively applying the following multiplicative updates,

$$\theta \longleftarrow \theta \, \frac{\nabla_{\theta}^{-} \mathcal{C}}{\nabla_{\theta}^{+} \mathcal{C}} \tag{A.2}$$

where, $\nabla_{\theta}^{-} \mathcal{C}$ and $\nabla_{\theta}^{+} \mathcal{C}$ are respectively the negative and the positive terms in the derivative $\nabla_{\theta} \mathcal{C} = \partial \mathcal{C} / \partial \theta$ as used in [109,185]. The derivative of $\mathcal{C}$ with respect to the parameter θ is :

$$\nabla_\theta \mathcal{C} = \nabla_\theta \mathcal{D} + \nabla_\theta f(\Theta) \tag{A.3}$$

$$\nabla_\theta \mathcal{C} = \sum_{f=1}^{F} \sum_{b=1}^{B} \left(-\frac{\mathbf{Z}(b,f)}{\tilde{\mathbf{Z}}(b,f)} \frac{\partial \tilde{\mathbf{Z}}(b,f)}{\partial \theta} + \frac{\partial \tilde{\mathbf{Z}}(b,f)}{\partial \theta} \right) + \frac{\partial f(\Theta)}{\partial \theta}. \tag{A.4}$$

The element-wise ratio $\mathbf{Z} \oslash \tilde{\mathbf{Z}}$ is denoted as $\mathbf{R}$. In the proposed models, the approximation used for clean speech and noise in the STFT domain are $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{W}}$, respectively.

$$\tilde{\mathbf{Y}} = \sum_{t=1}^{T} \mathbf{S_t} \overset{(t-1)\rightarrow}{\mathbf{X_s}} \quad \text{and} \quad \tilde{\mathbf{W}} = \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\rightarrow}{\mathbf{X_n}} . \tag{A.5}$$

An arbitrary l-th column in $\tilde{\mathbf{Z}}$ can be written as $[\tilde{\mathbf{Z}}]_l = \sum_{t=1}^{T} \mathbf{S_t}[\mathbf{X_s}]_{l-t+1} + \sum_{t=1}^{T} \mathbf{N_t}[\mathbf{X_n}]_{l-t+1}$. Thus every k-th column in $\mathbf{X_s}$ and $\mathbf{X_n}$ appears in columns in the range $[k, k+T-1]$ of $\tilde{\mathbf{Z}}$. The derivative of $\mathcal{D}$ with respect to a column $[\mathbf{X_s}]_k$ is:

$$\nabla_{[\mathbf{X_s}]_k} \mathcal{D} = -\sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} [\mathbf{R}]_{k+t-1} + \sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \mathbf{1}_{B\times 1} \tag{A.6}$$

which in matrix form can be written as:

$$\nabla_{\mathbf{X_s}} \mathcal{D} = -\sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \overset{\leftarrow(t-1)}{\mathbf{R}} + \sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \overset{\leftarrow(t-1)}{\mathbf{1}} \tag{A.7}$$

where, $\mathbf{1}$ is a matrix of ones of size $B \times F$. The same derivation can be used to obtain $\nabla_{\mathbf{X_n}} \mathcal{D}$.

In this work, the constraint $f(\Theta)$ used is to force the activations to have sparse solutions which is given as $f(\Theta) = \lambda_s \|\mathbf{X_s}\|_1 + \lambda_n \|\mathbf{X_n}\|_1$, where $\| \cdot \|_1$ denotes the sum of all the elements in a matrix. Thus a larger λ value ensures a sparser solution. The derivative of $f(\Theta)$ with respect to any element in $\mathbf{X_s}$ and $\mathbf{X_n}$ is thus λ_s and λ_n, respectively.

A.1 Derivations for the NMD+R setting

In this setting, the approximation for the noisy reverberant speech used is

$$\tilde{\mathbf{Z}} = \sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \overset{(p-1)\rightarrow}{\tilde{\mathbf{Y}}} + \tilde{\mathbf{W}} \tag{A.8}$$

$$\Rightarrow \quad \tilde{\mathbf{Z}} = \sum_{p=1}^{L} \sum_{t=1}^{T} [\mathbf{H}]_p \circledcirc \left(\mathbf{S_t} \overset{\tau\rightarrow}{\mathbf{X_s}} \right) + \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\rightarrow}{\mathbf{X_n}} \tag{A.9}$$

where $\tau = p + t - 2$ and $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{W}}$ are the STFT estimates of clean speech and noise, respectively. Using the notations defined before, an arbitrary l-th column of $\tilde{\mathbf{Z}}$ can be written as

$$\begin{aligned} [\tilde{\mathbf{Z}}]_l &= \sum_{p=1}^{L} [\mathbf{H}]_p \odot [\tilde{\mathbf{Y}}]_{l-p+1} \; + \; [\tilde{\mathbf{W}}]_l \\ &= \sum_{p=1}^{L} [\mathbf{H}]_p \odot \left(\sum_{t=1}^{T} \mathbf{S_t} [\mathbf{X_s}]_{l-\tau} \right) + \sum_{t=1}^{T} \mathbf{N_t} [\mathbf{X_n}]_{l-t+1}. \end{aligned} \tag{A.10}$$

A.1.1 Multiplicative updates for activations

It is clear from the formulation that every k-th column in $\mathbf{X_s}$ appears in columns in the range $[k, k+T+L-2]$ of $\tilde{\mathbf{Z}}$. The derivatives $\nabla_{\mathbf{X_s}}\mathcal{C}$ and $\nabla_{\mathbf{X_n}}\mathcal{C}$ can be obtained as described in Section A.

$$\begin{aligned} \nabla_{\mathbf{X_s}}\mathcal{C} = -\sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \left(\sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \overset{\leftarrow\tau}{\mathbf{R}} \right) + \\ \sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \left(\sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \overset{\leftarrow\tau}{\mathbf{1}} \right) + \lambda_s \end{aligned}$$

$$\nabla_{\mathbf{X_n}}\mathcal{C} = -\sum_{t=1}^{T} \mathbf{N_t}^{\mathsf{T}} \overset{\leftarrow(t-1)}{\mathbf{R}} + \sum_{t=1}^{T} \mathbf{N_t}^{\mathsf{T}} \overset{\leftarrow(t-1)}{\mathbf{1}} + \lambda_n.$$

The multiplicative updates for the activations are (using (A.2)),

$$\mathbf{X_s} \leftarrow \mathbf{X_s} \odot \frac{\sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \left(\sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \overset{\leftarrow\tau}{\mathbf{R}} \right)}{\sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \left(\sum_{t=1}^{T} \mathbf{S_t}^{\mathsf{T}} \overset{\leftarrow\tau}{\mathbf{1}} \right) + \lambda_s}$$

$$\mathbf{X_n} \leftarrow \mathbf{X_n} \odot \frac{\sum_{t=1}^{T} \mathbf{N_t}^{\mathsf{T}} \overset{\leftarrow(t-1)}{\mathbf{R}}}{\sum_{t=1}^{T} \mathbf{N_t}^{\mathsf{T}} \overset{\leftarrow(t-1)}{\mathbf{1}} + \lambda_n}$$

A.1.2 Multiplicative updates for RIR

The derivative of $\mathcal{C}$ with respect to an arbitrary p-th column of $\mathbf{H}$ can be obtained using (A.4) and (A.10) as follows.

$$\nabla_{[\mathbf{H}]_p} \mathcal{C} = -\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1} \odot \ [\mathbf{R}]_l + \sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1}$$

The multiplicative update for the RIR estimate is,

$$[\mathbf{H}]_p \leftarrow [\mathbf{H}]_p \odot \frac{\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1} \odot \ [\mathbf{R}]_l}{\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1}}. \tag{A.11}$$

A.2 Derivations for the NMD+R+DN setting

In the proposed setting, the approximation for the noisy reverberant signal used is

$$\tilde{\mathbf{Z}} = \sum_{p=1}^{L} \bar{\alpha}_p \left[\bar{\mathbf{H}}\right]_p \circledcirc \overset{(p-1)\rightarrow}{\tilde{\mathbf{Y}}} + \tilde{\mathbf{W}} \tag{A.12}$$

where, $[\bar{\mathbf{H}}]_p \triangleq [\mathbf{H}]_p / \|[\mathbf{H}]_p\|_2$. The $\bar{\alpha}$ vector is used to impose a decaying structure on the ℓ_2 norm of the RIR matrix by setting the l-th element in α as $\alpha_l = (1 + c_l)\alpha_{l+1} = \prod_{n=l}^{L}(1 + c_l)$, where $c_l \geq 0$, $\forall l = 1, \ldots, L$, where $\alpha_L = 1$ and $c_L = 0$. The decays are also constrained to have an ℓ_2 norm of 1, denoted as $\bar{\alpha}$, to avoid indeterminacy due to scaling (the algorithm may converge to any scaled value of α and this scaling difference can be captured by the activations. This results in an infinite number of solutions with the same minimum cost). The ℓ_2 norm of the α vector is $\|\alpha\| \triangleq \sqrt{\sum_{l=1}^{L} \prod_{n=l}^{L} (1 + c_n)^2}$.

Thus the elements of $\bar{\alpha}$ are:

$$\bar{\alpha}_l = \frac{\prod_{n=l}^{L}(1+c_n)}{\|\alpha\|} \tag{A.13}$$

Thus the approximation problem boils down to estimating the activations $\mathbf{X}$, RIR $\mathbf{H}$ and the decay coefficients c_i.

A.2.1 Multiplicative updates for activations

Obtaining the multiplicative updates for activations is straight-forward using the steps described in Section A.1.1 by replacing $[\mathbf{H}]_p$ with $\bar{\alpha}_p[\bar{\mathbf{H}}]_p$.

$$\mathbf{X_s} \leftarrow \mathbf{X_s} \odot \frac{\sum_{p=1}^{L} \bar{\alpha}_p [\bar{\mathbf{H}}]_p \circledcirc \left(\sum_{t=1}^{T} \mathbf{S_t}^{\intercal} \overset{\leftarrow\tau}{\mathbf{R}}\right)}{\sum_{p=1}^{L} \bar{\alpha}_p [\bar{\mathbf{H}}]_p \circledcirc \left(\sum_{t=1}^{T} \mathbf{S_t}^{\intercal} \overset{\leftarrow\tau}{\mathbf{1}}\right) + \lambda_s}$$

$$\mathbf{X_n} \leftarrow \mathbf{X_n} \odot \frac{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow(t-1)}{\mathbf{R}}}{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow(t-1)}{\mathbf{1}} + \lambda_n}$$

A.2.2 Multiplicative updates for RIR

In this section, we describe the multiplicative updates for the RIR $\mathbf{H}$ such that its ℓ_2 norm is always preserved to be equal to 1. To obtain the derivative w.r.t. an arbitrary p-th column of $\mathbf{H}$, we define the following. $\mathbf{h} = [\mathbf{H}]_p$, $\bar{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|_2$ and h_i be the i-th element of $\mathbf{h}$. We also make use of the known derivative $\dfrac{\partial\|\mathbf{h}\|}{\partial\mathbf{h}} = \dfrac{\mathbf{h}}{\|\mathbf{h}\|}$. Then,

$$\frac{\partial}{\partial h_i}\mathcal{C} = \sum_{j=1}^{B} \frac{\partial \bar{h}_j}{\partial h_i}\frac{\partial \mathcal{C}}{\partial \bar{h}_j}$$

$$\frac{\partial \bar{h}_j}{\partial h_i} = \frac{\partial}{\partial h_i}\frac{h_j}{\|\mathbf{h}\|} = \begin{cases} \dfrac{1}{\|\mathbf{h}\|} - \dfrac{h_i^2}{\|\mathbf{h}\|^3} & i = j \\ -\dfrac{h_i h_j}{\|\mathbf{h}\|^3} & i \neq j \end{cases}$$

$$\Rightarrow \quad \frac{\partial}{\partial h_i}\mathcal{C} = \frac{1}{\|\mathbf{h}\|}\left(\frac{\partial \mathcal{C}}{\partial \bar{h}_i} - \sum_{j=1}^{B} \frac{h_i h_j}{\|\mathbf{h}\|^2}\frac{\partial \mathcal{C}}{\partial \bar{h}_j}\right)$$

Using $\bar{h}_i = h_i / \|\mathbf{h}\|$ we can write the above expression as:

$$\nabla_{\mathbf{h}}\mathcal{C} = \frac{1}{\|\mathbf{h}\|}\left(\mathbf{I}_B - \bar{\mathbf{h}}\bar{\mathbf{h}}^\intercal\right)\nabla_{\bar{\mathbf{h}}}\mathcal{C} \tag{A.14}$$

The derivative $\nabla_{\bar{\mathbf{h}}}\mathcal{C}$ can be directly obtained using the derivation given in Section A.1.2.

$$\nabla_{\bar{\mathbf{h}}}\mathcal{C} = -\underbrace{\bar{\alpha}_p \sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1} \odot [\mathbf{R}]_l}_{\nabla^-_{\bar{\mathbf{h}}}\mathcal{C}} + \underbrace{\bar{\alpha}_p \sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1}}_{\nabla^+_{\bar{\mathbf{h}}}\mathcal{C}} \tag{A.15}$$

Using (A.14) and (A.15), the required derivative and the multiplicative updates can be obtained as follows.

$$\begin{aligned}
\nabla_{\mathbf{h}}\mathcal{C} &= -\frac{1}{\|\mathbf{h}\|}\left(\mathbf{I}_B - \bar{\mathbf{h}}\bar{\mathbf{h}}^\intercal\right)\left(-\nabla^-_{\bar{\mathbf{h}}}\mathcal{C} + \nabla^+_{\bar{\mathbf{h}}}\mathcal{C}\right)\\
&= -\frac{1}{\|\mathbf{h}\|}\left(\nabla^-_{\bar{\mathbf{h}}}\mathcal{C} \;+\; \bar{\mathbf{h}}\bar{\mathbf{h}}^\intercal\nabla^+_{\bar{\mathbf{h}}}\mathcal{C}\right) +\\
&\quad \frac{1}{\|\mathbf{h}\|}\left(\bar{\mathbf{h}}\bar{\mathbf{h}}^\intercal\nabla^-_{\bar{\mathbf{h}}}\mathcal{C} \;+\; \nabla^+_{\bar{\mathbf{h}}}\mathcal{C}\right)\\
\Rightarrow\;\; \mathbf{h} &\longleftarrow \mathbf{h} \odot \frac{\nabla^-_{\bar{\mathbf{h}}}\mathcal{C} \;+\; \bar{\mathbf{h}}\bar{\mathbf{h}}^\intercal\nabla^+_{\bar{\mathbf{h}}}\mathcal{C}}{\bar{\mathbf{h}}\bar{\mathbf{h}}^\intercal\nabla^-_{\bar{\mathbf{h}}}\mathcal{C} \;+\; \nabla^+_{\bar{\mathbf{h}}}\mathcal{C}}
\end{aligned}$$

The updates are such that the ℓ_2 norm of every column of $\mathbf{H}$ is preserved to unity.

A.2.3 Multiplicative updates for the decay coefficients

The multiplicative updates for the decay coefficients c_i can be found using

$$\nabla_{c_i}\mathcal{C} = \sum_{l=1}^{L} \frac{\partial\bar{\alpha}_l}{\partial c_i} \cdot \nabla_{\bar{\alpha}_l}\mathcal{C}$$

$$\nabla_{\bar{\alpha}_l}\mathcal{C} = -\underbrace{[\bar{\mathbf{H}}]_l^\intercal \sum_{j=1}^{F} [\mathbf{R}]_j \odot [\mathbf{Y}]_{j-l+1}}_{\nabla^-_{\bar{\alpha}_l}\mathcal{C}} + \underbrace{[\bar{\mathbf{H}}]_l^\intercal \sum_{j=1}^{F} [\mathbf{Y}]_{j-l+1}}_{\nabla^+_{\bar{\alpha}_l}\mathcal{C}}$$

$$\frac{\partial \bar{\alpha}_l}{\partial c_i} = \begin{cases} \prod_{n=l}^{L}(1+c_n)\frac{\partial}{\partial c_i}\frac{1}{\|\alpha\|} & i < l \\ \frac{\prod_{\substack{n=l \\ n\neq i}}^{L}(1+c_n)}{\|\alpha\|} + \prod_{n=l}^{L}(1+c_n)\frac{\partial}{\partial c_i}\frac{1}{\|\alpha\|} & i \geq l \end{cases}$$

$$= \begin{cases} -\frac{\prod_{n=l}^{L}(1+c_n)}{\|\alpha\|^2}\frac{\partial}{\partial c_i}\|\alpha\| & i < l \\ \frac{\prod_{n=l}^{L}(1+c_n)}{\|\alpha\|(1+c_i)} - \frac{\prod_{n=l}^{L}(1+c_n)}{\|\alpha\|^2}\frac{\partial}{\partial c_i}\|\alpha\| & i \geq l \end{cases}$$

$$= \begin{cases} -\frac{\bar{\alpha}_l}{\|\alpha\|}\frac{\partial}{\partial c_i}\|\alpha\| & i < l \\ \frac{\bar{\alpha}_l}{(1+c_i)} - \frac{\bar{\alpha}_l}{\|\alpha\|}\frac{\partial}{\partial c_i}\|\alpha\| & i \geq l \end{cases}$$

$$\Rightarrow \quad \nabla_{c_i}\mathcal{C} = \sum_{l=1}^{i}\frac{\bar{\alpha}_l}{(1+c_i)}\nabla_{\bar{\alpha}_l}\mathcal{C} - \sum_{l=1}^{L}\frac{\bar{\alpha}_l}{\|\alpha\|}\frac{\partial}{\partial c_i}\|\alpha\|\nabla_{\bar{\alpha}_l}\mathcal{C}$$

The derivative of $\|\alpha\|$ w.r.t. c_i can be found as:

$$\frac{\partial\|\alpha\|}{\partial c_i} = (1+c_i)\frac{\sum_{j=1}^{i}\prod_{\substack{n=j \\ n\neq i}}^{L}(1+c_n)^2}{\|\alpha\|}$$

$$= \frac{\sum_{j=1}^{i}\prod_{n=j}^{L}(1+c_n)^2}{\|\alpha\|(1+c_i)}$$

$$= \|\alpha\|\frac{\sum_{j=1}^{i}\bar{\alpha}_j^2}{(1+c_i)}$$

Thus the derivative w.r.t. c_i becomes,

$$\nabla_{c_i}\mathcal{C} = \sum_{l=1}^{i}\frac{\bar{\alpha}_l}{(1+c_i)}\nabla_{\bar{\alpha}_l}\mathcal{C} -$$

$$\sum_{l=1}^{L}\frac{\bar{\alpha}_l}{\|\alpha\|}\|\alpha\|\frac{\sum_{j=1}^{i}\bar{\alpha}_j^2}{(1+c_i)}\nabla_{\bar{\alpha}_l}\mathcal{C}$$

$$= \frac{1}{(1+c_i)}\left(\sum_{l=1}^{i}\bar{\alpha}_l\nabla_{\bar{\alpha}_l}\mathcal{C} - \sum_{j=1}^{i}\bar{\alpha}_j^2\sum_{l=1}^{L}\bar{\alpha}_l\nabla_{\bar{\alpha}_l}\mathcal{C}\right)$$

The multiplicative updates can now be obtained as follows.

$$c_i \longleftarrow c_i \cdot \frac{\sum_{l=1}^{i} \bar{\alpha}_l \nabla^{-}_{\bar{\alpha}_l} \mathcal{C} + \sum_{j=1}^{i} \bar{\alpha}_j^2 \sum_{l=1}^{L} \bar{\alpha}_l \nabla^{+}_{\bar{\alpha}_l} \mathcal{C}}{\sum_{l=1}^{i} \bar{\alpha}_l \nabla^{+}_{\bar{\alpha}_l} \mathcal{C} + \sum_{j=1}^{i} \bar{\alpha}_j^2 \sum_{l=1}^{L} \bar{\alpha}_l \nabla^{-}_{\bar{\alpha}_l} \mathcal{C}} \tag{A.16}$$

Appendix B

Stimuli Set used for the picture naming task

The list of stimuli used in the picture naming task in Chapter 6 divided into subsets are given below. The bias corrections used for every subset is also given in brackets.

1. **Vowels (VOW)** (30 ms) : aap, aardbei, accordeon, ananas, anker, appel, arm, artisjok, asbak, asperge, auto, eekhoorn, eend, eikel, ezel, olifant, oog, oor, uil
2. **Voiced stops (VSTP)** (30 ms) : bal, ballon, banaan, bed, beer, beitel, berg, bij, bloem, boek, boom, bril, broek, bus, deegrol, deur, duim
3. **Unvoiced stops (USTP)** (40 ms) : clown, kaars, kam, kanon, kerk, kers, kever, kikker, kip, koe, koffer, konijn, kreeft, krokodil, kroon, paard, paddenstoel, paprika, paraplu, pauw, peer, pijp, pinguin, pompoen, pop, potlood, tafel, tandenborstel, tang, telefoon, tijger, tomaat, ton, trompet, trui
4. **Nasals (NAS)** (30 ms) : maan, mand, mes, neushoorn
5. **Sibilant Fricatives (SIB)** (20 ms) : citroen, schaap, schaar, schildpad, schoen, schommel, schommelstoel, schroevendraaier, sigaar, sigaret, slak, sleutel, spin, spinnenwiel, sprinkhaan, ster, stoel, strijkplank, struisvogel, zaag, zebra, zeepaardje, zon, zwaan
6. **Non-sibilant Fricatives (NSIB)** (10 ms) : fiets, gieter, glas, giraf, gitaar, haan, hand, harp, hart, helicopter, hert, hoed, hond, huis, vaas, varken,

verkeerslicht, vingerhoed, viool, vis, vlag, vlieg, vlieger, vliegtuig, vlinder, vork, vos, wasknijper, wiel, wolk, wortel

7. **Liquids and approximants (LIQ)** (10 ms) : jojo, laars, ladder, leeuw, lepel, ring, rits, rok, rolschaats, rups

Bibliography

[1] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., and Penn, G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2012), pp. 4277–4280. pages 16

[2] Amodei, D. e. a. Deepspeech 2: End-to-end speech recognition in enlish and mandarin. In *International Conference in Machine Learning (ICML)* (New York, USA, 2016). pages 18

[3] Aradilla, G., Vepa, J., and Bourlard, H. Using posterior-based features in template matching for speech recognition. In *INTERSPEECH* (Pittsburgh, PA, USA, September 2006), ISCA. pages 13

[4] Atal, B., and Schroeder, M. Predictive coding of speech signals and subjective error criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing 27*, 3 (June 1979), 247–254. pages 14

[5] Baby, D., Gemmeke, J. F., Virtanen, T., and Van hamme, H. Exemplar-based speech enhancement for deep neural network based automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2015), pp. 4485–4489. pages 72, 78, 106, 109

[6] Baby, D., and Van hamme, H. Supervised speech dereverberation in noisy environments using exemplar-based sparse representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2016), pp. 156–160. pages 85, 87, 88, 89, 92

[7] Baby, D., Virtanen, T., Barker, T., and Van hamme, H. Coupled dictionary training for exemplar-based speech enhancement. In *IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy, May 2014), pp. 2883–2887. pages 27, 32, 33, 36, 37, 44, 57, 59, 61, 63, 67, 86

[8] Baby, D., Virtanen, T., Gemmeke, J. F., Barker, T., and Van hamme, H. Exemplar-based noise robust speech recognition using modulation spectrogram features. In *IEEE Spoken Language Technology Workshop* (South Lake Tahoe, USA, December 2014), pp. 519–524. pages 19, 33, 34, 36, 45, 57, 67, 73

[9] Baby, D., Virtanen, T., Gemmeke, J. F., and Van hamme, H. Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 23*, 11 (November 2015), 1788 –1799. pages 106

[10] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473* (2014). pages 18

[11] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai, China, March 2016), pp. 4945–4949. pages 13, 18

[12] Ban, S. M., and Kim, H. S. Weight-space viterbi decoding based spectral subtraction for reverberant speech recognition. *IEEE Signal Processing Letters 22*, 9 (September 2015), 1424–1428. pages 5

[13] Barker, A. T., Jalinous, R., and Freeston, I. L. Non-invasive magnetic stimulation of human motor cortex. *Lancet 1* (1985), 1106–1107. pages 101

[14] Barker, J., Marxer, R., Vincent, E., and Watanbe, S. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2015). pages 21, 22, 25, 48, 49

[15] Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. D. The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language 27*, 3 (2013), 621 – 633. Special Issue on Speech Separation and Recognition in Multisource Environments. pages 22

[16] Barker, T., and Virtanen, T. Non-negative tensor factorization of modulation spectrograms for sonaural sound source separation. In *INTERSPEECH* (2013), ISCA, pp. 827–831. pages 33, 35, 61, 73

[17] Bestmann, S. The physiological basis of transcranial magnetic stimulation. *Trends in cognitive sciences 12*, 3 (March 2008), 81–83. pages 19, 99, 101

[18] Beutelmann, R., and Brand, T. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America 120* (2006), 331–342. pages 83

[19] Boersma, P., and Weenink, D. Praat, a system for doing phonetics by computer. *Glot International 5*, 9/10 (2001), 341–345. pages 103

[20] Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing 27*, 2 (April 1979), 113–120. pages 5, 11, 27

[21] Bregman, A. S. *Auditory scene analysis: The perceptual organization of sound.* The MIT Press, 1990. pages 33

[22] Buchner, H., Aichner, R., and Kellermann, W. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *Speech and Audio Processing, IEEE Transactions on 13*, 1 (January 2005), 120–134. pages 83

[23] Buchner, H., Aichner, R., and Kellermann, W. TRINICON-based blind system identification with application to multiple-source localization and separation. In *Blind Speech Separation*, S. Makino, H. Sawada, and T. Lee, Eds., Signals and Communication Technology. Springer Netherlands, 2007, pp. 101–147. pages 83

[24] Cappe, O. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Speech and Audio Processing 2*, 2 (April 1994), 345–349. pages 5

[25] Cerisara, C., Demange, S., and Haton, J.-P. On noise masking for automatic missing data speech recognition: A survey and discussion. *Computer Speech & Language 21*, 3 (2007), 443 – 457. pages 4

[26] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. Listen, attend and spell. *arXiv:1508.01211* (2015). pages 18

[27] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai, China, March 2016), pp. 4960–4964. pages 13

[28] Chen, R., Chan, C. F., and So, H. C. Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 4 (May 2012), 1324–1336. pages 4

[29] Chen, S. F., and Goodman, J. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 1996), Association for Computational Linguistics, pp. 310–318. pages 17

[30] Cho, K., Van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, October 2014), Association for Computational Linguistics, pp. 1724–1734. pages 18

[31] Cho, Y. D., Al-Naimi, K., and Kondoz, A. Mixed decision-based noise adaptation for speech enhancement. *Electronics Letters 37*, 8 (April 2001), 540–542. pages 4

[32] Cho, Y. D., and Kondoz, A. Analysis and improvement of a statistical model-based voice activity detector. *IEEE Signal Processing Letters 8*, 10 (October 2001), 276–278. pages 4

[33] Chorowski, J., Bahdanau, D., Cho, K., and Bengio, Y. End-to-end continuous speech recognition using attention-based recurrent NN: first results. *arXiv:1412.1602* (2014). pages 18

[34] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. Attention-based models for speech recognition. In *Annual Conference on Neural Information Processing Systems (NIPS)* (Montreal, Quebec, Canada, December 2015), pp. 577–585. pages 13, 18

[35] Christensen, H., Barker, J., Ma, N., and Green, P. D. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *INTERSPEECH* (2010), ISCA, pp. 1918–1921. pages 22, 92

[36] Cichocki, A., Cruces, S., and Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy 13*, 1 (2011), 134–170. pages 8

[37] Cichocki, A., Zdunek, A., Phan, A. H., and Amari, S. *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* Wiley, 2009. pages 8

[38] Cichocki, A., Zdunek, R., and Amari, S. Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. In *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA)* (Charleston, SC, USA, 2006), pp. 32–39. pages 8

[39] Clark, P., Sell, G., and Atlas, L. A novel approach using modulation features for multiphone-based speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2011), pp. 5264–5267. pages 71

[40] Cohen, I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing 11*, 5 (September 2003), 466–475. pages 4, 45, 93

[41] Cohen, I., and Berdugo, B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters 9*, 1 (January 2002), 12–15. pages 4

[42] Dahl, G., Yu, D., Deng, L., and Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 1 (January 2012), 30–42. pages 16, 71

[43] De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., and Van Compernolle, D. Template-based continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 4 (May 2007), 1377–1390. pages 13

[44] Demuynck, K., Roelens, J., Van Compernolle, D., and Wambacq, P. SPRAAK: an open source "speech recognition and automatic annotation kit". In *INTERSPEECH* (Brisbane, Australia, September 2008), p. 495. pages 24

[45] Deng, L., Hinton, G., and Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2013), pp. 8599–8603. pages 71

[46] Deng, L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., and Acero, A. Recent advances in deep learning for speech research at Microsoft. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2013), pp. 8604–8608. pages 71

[47] Duchateau, J., Kong, Y., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquiére, P., Verhelst, W., and Van hamme, H. Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. *Speech Communication 51*, 10 (2009), 985 – 994. pages 104, 105, 106

[48] Eldaief, M., Press, D., and Pascual-Leone, A. Transcranial magnetic stimulation in neurology: A review of established and prospective applications. *Neurology: Clinical Practice 3*, 6 (2013), 519–526. pages 101

[49] Ephraim, Y. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Transactions on Signal Processing 40*, 4 (1992), 725–735. pages 27

[50] Ephraim, Y., and Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing 33*, 2 (April 1985), 443–445. pages 5, 45, 93

[51] Falk, T. H., Zheng, C., and Chan, W. Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing 18*, 7 (September 2010), 1766–1774. pages 93

[52] Févotte, C., Bertin, N., and Durrieu, J. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation 21*, 3 (2009), 793–830. pages 8

[53] Févotte, C., and Idier, J. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation 23*, 9 (2011), 2421–2456. pages 8

[54] Fiscus, J. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding* (1997), pp. 347–354. pages 38

[55] Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 2 (1936), 179–188. pages 14

[56] Gannot, S. Speech processing utilizing the kalman filter. *IEEE Instrumentation Measurement Magazine 15*, 3 (June 2012), 10–14. pages 4

[57] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. Timit acoustic-phonetic continuous speech corpus ldc93s1, 1993. pages 21

[58] Geiger, J. T., Gemmeke, J. F., Schuller, B., and Rigoll, G. Investigating NMF speech enhancement for neural network based acoustic models. In *INTERSPEECH* (2014), ISCA, pp. 2405–2409. pages 6, 27, 32, 35, 36, 37

[59] Geiger, J. T., Vipperla, R., Bozonnet, S., Evans, N. W. D., Schuller, B. W., and Rigoll, G. Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization. In *INTERSPEECH* (September 2012), ISCA, pp. 2154–2157. pages 9

[60] Geiger, J. T., Weninger, F., Hurmalainen, A., Gemmeke, J. F., Wöllmer, M., Schuller, B., Rigoll, G., and Virtanen, T. The TUM+TUT+KUL Approach to the CHiME Challenge 2013: Multi-Stream ASR Exploiting BLSTM Networks and Sparse NMF. In *Proceedings of the 2nd CHiME workshop* (June 2013), pp. 25–30. pages 57

[61] Gemmeke, J. F., Cranen, B., and Remes, U. Sparse imputation for large vocabulary noise robust ASR . *Computer Speech & Language 25*, 2 (2011), 462 – 479. pages 4

[62] Gemmeke, J. F., and Van hamme, H. Advances in noise robust digit recognition using hybrid exemplar based systems. In *INTERSPEECH* (2012), ISCA, pp. 2134–2137. pages 6, 27, 35, 38, 42, 63, 84

[63] Gemmeke, J. F., and Virtanen, T. Noise robust exemplar-based connected digit recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2010), pp. 4546 –4549. pages 7, 57, 58

[64] Gemmeke, J. F., Virtanen, T., and Hurmalainen, A. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 7 (September 2011), 2067 –2080. pages 10, 18, 35, 37, 42, 57, 58, 62, 86

[65] Ghoshal, A., Povey, D., Agarwal, M., Akyazi, P., Burget, L., Feng, K., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R. C., Schwarz, P., and Thomas, S. A novel estimation of feature-space MLLR for full-covariance models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2010), pp. 4310–4313. pages 14

[66] Gilloire, A., and Vetterli, M. Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation. *IEEE Transactions on Signal Processing 40*, 8 (August 1992), 1862–1875. pages 85

[67] Gönen, M. Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning. *Pattern Recognition Letters 38* (2014), 132–141. pages 28

[68] Grancharov, V., Samuelsson, J., and Kleijn, B. On causal algorithms for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing 14*, 3 (2006), 764–773. pages 27

[69] Graves, A., Fernández, S., Gomez, F. J., and Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Twenty-Third International Conference on Machine Learning (ICML)* (Pittsburgh, Pennsylvania, USA, June 2006), pp. 369–376. pages 13, 18

[70] Graves, A., and Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning (ICML)* (June 2014), pp. 1764–1772. pages 18

[71] Greenberg, S., and Kingsbury, B. The modulation spectrogram: in pursuit of an invariant representation of speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1997), vol. 3, pp. 1647–1650. pages 19, 28, 33, 61, 71, 73

[72] Griffin, D., and Lim, J. Signal estimation from modified short-time Fourier transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1983), vol. 8, pp. 804–807. pages 34

[73] Haeb-Umbach, R., and Ney, H. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 1992), vol. 1, pp. 13–16 vol.1. pages 38

[74] Hallett, M. Transcranial magnetic stimulation and the human brain. *Nature 406* (July 2000), 147–150. pages 19, 99, 101

[75] Hannun, A. Y., Maas, A. L., Jurafsky, D., and Ng, A. Y. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. In *arXiv:1408.2873* (2014). pages 18

[76] Hartwigsen, G. The neurophysiology of language: Insights from non-invasive brain stimulation in the healthy human brain. *Brain and Language 148* (2015), 81 – 94. pages 101

[77] Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America 87*, 4 (1990), 1738–1752. pages 14

[78] Hermansky, H., and Morgan, N. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing 2*, 4 (Oct 1994), 578–589. pages 14

[79] Hermus, K., Wambacq, P., and Van hamme, H. A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Advances in Signal Processing 2007* (September 2007). pages 4

[80] Hinton, G. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade (2nd ed.)*. 2012, pp. 599–619. pages 39, 74

[81] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29*, 6 (Nov 2012), 82–97. pages 14, 15, 38, 71

[82] Hirsch, H. G., and Pearce, D. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Automatic Speech Recognition: Challenges for the new Millenium, ISCA Tutorial and Research Workshop (ITRW)* (Paris, France, 2000), pp. 29–32. pages 20, 34, 38

[83] Huang, X., Acero, A., and Hon, H. W. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, 1st ed. ed. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001. pages 17

[84] Huang, Y., Benesty, J., and Chen, J. A blind channel identification-based two-stage approach to separation and dereverberation of speech

signals in a reverberant environment. *Speech and Audio Processing, IEEE Transactions on 13*, 5 (September 2005), 882–895. pages 83

[85] Hurmalainen, A., Gemmeke, J. F., and Virtanen, T. Non-negative matrix deconvolution in noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Prague, Czech Republic, 2011), pp. 4588–4591. pages 9

[86] Hurmalainen, A., Gemmeke, J. F., and Virtanen, T. Non-negative matrix deconvolution in noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2011), pp. 4588 –4591. pages 84

[87] Hurmalainen, A., Gemmeke, J. F., and Virtanen, T. Modelling non-stationary noise with spectral factorisation in automatic speech recognition. *Computer Speech & Language 27*, 3 (May 2013), 763–779. pages 9, 48

[88] Hurmalainen, A., Saeidi, R., and Virtanen, T. Group sparsity for speaker identity discrimination in factorisation-based speech recognition. In *INTERSPEECH* (September 2012), ISCA, pp. 2138–2141. pages 9

[89] Hurmalainen, A., and Virtanen, T. Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 4113–4116. pages 6, 27, 57, 84

[90] Inoue, T., Saruwatari, H., Takahashi, Y., Shikano, K., and Kondo, K. Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 6 (August 2011), 1770–1779. pages 5

[91] Jabloun, F., and Champagne, B. Signal subspace techniques for speech enhancement. In *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer-Verlag, Berlin, Germany, 2005, ch. 7, pp. 135–159. pages 4

[92] Jaitly, N., Nguyen, P., Senior, A. W., and Vanhoucke, V. Application of pretrained deep neural networks to large vocabulary speech recognition. In *INTERSPEECH* (Portland, Oregon, USA, September 2012), pp. 2578–2581. pages 16

[93] Jensen, J., Benesty, J., Christensen, M., and Jensen, S. Enhancement of single-channel periodic signals in the time-domain. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 7 (2012), 1948–1963. pages 27

[94] Jeub, M., Schäfer, M., and Vary, P. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of the 16th International Conference on Digital Signal Processing* (2009), pp. 550–554. pages 92

[95] Jolliffe, I. T. *Principal component analysis.* Springer, New York, 2002. pages 14

[96] Kallasjoki, H., Gemmeke, J. F., Pallomaki, K. J., and Beeston, A. V. Recognition of reverberant speech by missing data imputation and NMF feature enhancement. In *Proc. REVERB Workshop* (Florence, Italy, May 2014). pages 84, 85

[97] Kameoka, H., Nakatani, T., and Yoshioka, T. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2009), pp. 45–48. pages 83

[98] Katz, S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing 35*, 3 (March 1987), 400–401. pages 17

[99] Kavalekalam, M. S., Christensen, M. G., Gran, F., and Boldt, J. B. Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2016), pp. 191–195. pages 4

[100] Kingsbury, B., Morgan, N., and Greenberg, S. Robust speech recognition using the modulation spectrogram. *Speech Communication 25*, 1-3 (1998), 117–132. pages 33, 71

[101] Kingsbury, B. E. D., and Morgan, N. Recognizing reverberant speech with RASTA-PLP. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (April 1997), vol. 2, pp. 1259–1262. pages 83

[102] Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., and Maas, R. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (Oct 2013), pp. 1–4. pages 83, 93

[103] Kneser, R., and Ney, H. Improved backing-off for m-gram language modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 1995), vol. 1, pp. 181–184 vol.1. pages 17

[104] Kokkinakis, K., Hazrati, O., and Loizou, P. C. A channel-selection criterion for suppressing reverberation in cochlear implants. *The Journal of the Acoustical Society of America 129*, 5 (2011), 3221–3232. pages 83

[105] Krieg, S. M., Sollmann, N., Hauck, T., Ille, S., Meyer, B., and Ringel, F. Repeated mapping of cortical language sites by preoperative navigated transcranial magnetic stimulation compared to repeated intraoperative dcs mapping in awake craniotomy. *BMC Neuroscience 15*, 1 (2014), 1–10. pages 20, 99, 101

[106] Kumar, K., Singh, R., Raj, B., and Stern, R. Gammatone subband magnitude-domain dereverberation for ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2011), pp. 4604–4607. pages 83

[107] Le Bouquin-Jeannés, R., and Faucon, G. Proposal of a voice activity detector for noise reduction. *Electronics Letters 30*, 12 (June 1994), 930–932. pages 4

[108] Le Bouquin-Jeannés, R., and Faucon, G. Study of a voice activity detector and its influence on a noise reduction system. *Speech Communication 16*, 3 (1995), 245 – 254. pages 4

[109] Le Roux, J., Hershey, J. R., and Weninger, F. Sparse NMF – half-baked or well done? In *Techical Report, Mitsubishi Electric Research Labs (MERL)* (Cambridge, USA, 2015), no. TR2015-023. pages 6, 10, 49, 117, 121

[110] Lee, D. D., and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature 401* (October 1999), 788–791. pages 6

[111] Lee, D. D., and Seung, H. S. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems* (2001), MIT Press, pp. 556–562. pages 6, 10

[112] Lefèvre, A., Bach, F. R., and Févotte, C. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, NY, USA, 2011), pp. 313–316. pages 8

[113] Lehmann, E. A., and Johansson, A. M. Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America 124*, 1 (2008), 269–277. pages 84

[114] Leonard, R. A database for speaker-independent digit recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (March 1984), vol. 9, pp. 328–331. pages 20

[115] Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 4 (April 2014), 745–777. pages 11

[116] Liang, D., Hoffman, M. D., and Mysore, G. J. Speech dereverberation using a learned speech model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2015), pp. 1871–1875. pages 83

[117] Lioumis, P., Zhdanov, A., Mäkelä, N., Lehtinen, H., Wilenius, J., Neuvonen, T., Hannula, H., Deletis, V., Picht, T., and Mäkelä, J. P. A novel approach for documenting naming errors induced by navigated transcranial magnetic stimulation. *Journal of Neuroscience Methods 204*, 2 (2012), 349 – 354. pages 102

[118] Loizou, P. C. *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1 ed. CRC Press, 2007. pages 37

[119] Lu, Y., and Loizou, P. C. A geometric approach to spectral subtraction. *Speech Communication 50*, 6 (2008), 453 – 466. pages 5

[120] Martin, R. An efficient algorithm to estimate the instantaneous SNR of speech signals. In *EUROSPEECH* (Berlin, Germany, September 1993). pages 4

[121] Martinez, A., Moritz, N., and Meyer, B. T. Should deep neural nets have ears? The role of auditory features in deep learning approaches. In *INTERSPEECH* (2014), ISCA, pp. 2435–2439. pages 71, 74, 76

[122] McAulay, R., and Malpass, M. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing 28*, 2 (April 1980), 137–145. pages 5

[123] Mikolov, T. *Statistical Language Models based on Neural Networks.* PhD thesis, Brno University of Technology, 2012. pages 17

[124] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. Recurrent neural network based language model. In *INTERSPEECH* (Makuhari, Japan, September 2010), ISCA, pp. 1045–1048. pages 17

[125] Mirsamadi, S., and Hansen, J. H. L. Multichannel feature enhancement in distributed microphone arrays for robust distant speech recognition in smart rooms. In *IEEE Spoken Language Technology Workshop (SLT)* (Dec 2014), pp. 507–512. pages 85

[126] Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., and Graciarena, M. Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In *INTERSPEECH* (2014), ISCA, pp. 895–899. pages 71

[127] Miyazaki, R., Saruwatari, H., Inoue, T., Takahashi, Y., Shikano, K., and Kondo, K. Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 7 (September 2012), 2080–2094. pages 5

[128] Mohamed, A., Hinton, G., and Penn, G. Understanding how Deep Belief Networks perform acoustic modelling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (March 2012), pp. 4273–4276. pages 71

[129] Mohamed, A. R., Dahl, G. E., and Hinton, G. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 1 (January 2012), 14–22. pages 16

[130] Mohammadiha, N., and Doclo, S. Single-channel dynamic exemplar-based speech enhancement. In *INTERSPEECH* (2014), ISCA, pp. 2690–2694. pages 6, 28, 59

[131] Mohammadiha, N., and Doclo, S. Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 24*, 2 (February 2016), 276–289. pages 83, 84, 85

[132] Mohammadiha, N., Gerkmann, T., and Leijon, A. A new linear mmse filter for single channel speech enhancement based on nonnegative matrix factorization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (October 2011), pp. 45–48. pages 6

[133] Mohammadiha, N., Smaragdis, P., and Doclo, S. Joint acoustic and spectral modeling for speech dereverberation using non-negative representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2015), pp. 4410–4414. pages 85, 88

[134] Mohammadiha, N., Smaragdis, P., and Leijon, A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing 21*, 10 (2013), 2140–2151. pages 27

[135] Moreno, P., Raj, B., and Stern, R. A vector Taylor series approach for environment-independent speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 1996), vol. 2, pp. 733–736 vol. 2. pages 11

[136] Mottaghy, F. M., Sparing, R., and Töpper, R. Enhancing picture naming with transcranial magnetic stimulation. *Behavioural Neurology 17*, 3-4 (2006), 177–186. pages 101

[137] Nam, J., Mysore, G. J., Ganseman, J., Lee, K., and Abel, J. S. A super-resolution spectrogram using coupled PLCA. In *INTERSPEECH* (Makuhari, Japan, 2010), ISCA, pp. 1696–1699. pages 28

[138] Nielsen, M. A. *Neural Networks and Deep Learning.* Determination Press, 2014. pages 38

[139] Patterson, R. D., Allerhand, M. H., and Gigur, C. Time-domain modeling of peripheral auditory processing- A modular architecture and a software platform. *Journal of Acoustical Society of America 98* (1995), 1890–1894. pages 72

[140] Picht, T., Krieg, S., Sollmann, N., Rösler, J., Niraula, B., Neuvonen, T., Savolainen, P., Lioumis, P., Mäkelä, J., Deletis, V., Meyer, B., Vajkoczy, P., and Ringel, F. A comparison of language mapping by preoperative navigated transcranial magnetic stimulation and direct cortical stimulation during awake surgery. *Neurosurgery 72*, 5 (2013), 808–819. pages 101

[141] Plack, C. *The sense of hearing.* Lawrence Erlbaum Associates Publishers, 2005. pages 32, 61, 71, 72

[142] Pobric, G., Jefferies, E., and Ralph, M. A. L. Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology 20*, 10 (2010), 964 – 968. pages 101

[143] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (December 2011), IEEE Signal Processing Society. pages 38, 49, 74

[144] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (February 1989), 257–286. pages 14

[145] Raj, B., Virtanen, T., Chaudhuri, S., and Singh, R. Non-negative matrix factorization based compensation of music for automatic speech recognition. In *INTERSPEECH* (Makuhari, Japan, 2010), ISCA, pp. 717–720. pages 27

[146] Ramírez, J., Segura, J. C., Benítez, C., de la Torre, A., and Rubio, A. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication 42*, 3–4 (2004), 271 – 287. pages 4

[147] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2001), vol. 2, pp. 749–752. pages 93

[148] Sadjadi, S. O., and Hansen, J. H. L. Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2011), pp. 5448–5451. pages 83

[149] Sainath, T., Mohamed, A.-R., Kingsbury, B., and Ramabhadran, B. Deep convolutional neural networks for LVCSR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2013), pp. 8614–8618. pages 16, 49, 50

[150] Sainath, T. N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., and Sethy, A. Sparse representation features for speech recognition. In *INTERSPEECH* (2010), ISCA, pp. 2254–2257. pages 48

[151] Sainath, T. N., Vinyals, O., Senior, A. W., and Sak, H. Convolutional, long short-term memory, fully connected deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2015), pp. 4580–4584. pages 16

[152] Sak, H., Senior, A. W., and Beaufays, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH* (Singapore, September 2014), pp. 338–342. pages 16

[153] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*, 1 (Feb 1978), 43–49. pages 13

[154] Saon, G., Padmanabhan, M., Gopinath, R., and Chen, S. Maximum likelihood discriminant feature spaces. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2000), vol. 2, pp. 1129–1132. pages 14, 38

[155] Schädler, M. R., Meyer, B. T., and Kollmeier, B. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *Journal of Acoustical Society of America 131*, 5 (2012), 4134–4151. pages 19, 72, 74

[156] Schmidt, M. N., and Larsen, J. Reduction of non-stationary noise using a non-negative latent variable decomposition. In *IEEE Workshop on Machine Learning for Signal Processing* (October 2008), pp. 486–491. pages 6

[157] Schmidt, M. N., and Olsson, R. K. Single-channel speech separation using sparse non-negative matrix factorization. In *INTERSPEECH* (2006), ISCA, pp. 2614–2617. pages 6

[158] Schreiner, C. E., and Urbas, J. V. Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF) . *Hearing Research 21* (1986), 227 – 241. pages 33, 72

[159] Schuhmann, T., Schiller, N. O., Goebel, R., and Sack, A. T. The temporal characteristics of functional activation in Broca's area during overt picture naming. *Cortex 45*, 9 (2009), 1111–1116. pages 101

[160] Schuurman, I., Schouppe, M., and Hoekstra, H. *CGN, an annotated corpus of spoken Dutch.* Budapest, 2003, pp. 101 – 108. Nella 21 nov (No. 369). pages 104

[161] Seide, F., Li, G., and Yu, D. Conversational speech transcription using context-dependent deep neural networks. In *INTERSPEECH* (Florence, Italy, 2011), ISCA, pp. 437–440. pages 16

[162] Seltzer, M., Yu, D., and Wang, Y. An investigation of deep neural networks for noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2013), pp. 7398–7402. pages 71

[163] Singh, R., Raj, B., and Smaragdis, P. Latent-variable decomposition based dereverberation of monaural and multi-channel signals. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (March 2010), pp. 1914–1917. pages 83

[164] Slaney, M. An efficient implementation of the patterson-holdsworth auditory filter bank. In *Technical Report 35* (1993), Apple Computer, Inc. pages 72

[165] Slaney, M. Auditory Toolbox Version 2. *Interval Research Corporation 10* (1998). pages 35, 63, 76

[166] Smaragdis, P. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation*, C. Puntonet and A. Prieto, Eds., vol. 3195 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, pp. 494–499. pages 9, 87

[167] Smaragdis, P. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 1 (Jan 2007), 1–12. pages 6, 84, 86

[168] Smaragdis, P., Shashanka, M., and Raj, B. A sparse non-parametric approach for single channel separation of known sounds. In *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2009, pp. 1705–1713. pages 6

[169] Snodgrass, J. G., and Vanderwart, M. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory 6*, 2 (March 1980), 174–215. pages 102

[170] Sohn, J., and Sung, W. A voice activity detector employing soft decision based noise spectrum adaptation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 1998), vol. 1, pp. 365–368. pages 4

[171] Sollmann, N., Hauck, T., Hapfelmeier, A., Meyer, B., and Ringel, F.and Krieg, S. M. Intra- and interobserver variability of language mapping by navigated transcranial magnetic brain stimulation. *BMC Neuroscience 14*, 1 (2013), 1–10. pages 20, 99, 101

[172] Sprechmann, P., Bronstein, A. M., Bronstein, M. M., and Sapiro, G. Learnable low rank sparse models for speech denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, BC, Canada, 2013), pp. 136–140. pages 8

[173] Sreenivas, T., and Kirnapure, P. Codebook constrained Wiener filtering for speech enhancement. *IEEE Transactions on Speech and Audio Processing 4*, 5 (1996), 383–389. pages 27

[174] Srinivasan, S. *Knowledge-based speech enhancement.* PhD thesis, KTH-Royal Institute of Technology, Stockholm, Sweden, 2005. pages 5

[175] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27.* Curran Associates, Inc., 2014, pp. 3104–3112. pages 18

[176] Talmon, R., Cohen, I., and Gannot, S. Relative transfer function identification using convolutive transfer function approximation. *Audio, Speech, and Language Processing, IEEE Transactions on 17*, 4 (May 2009), 546–555. pages 83, 85

[177] Tarapore, P. E., Findlay, A. M., Honma, S. M., Mizuiri, D., Houde, J. F., Berger, M. S., and Nagarajan, S. S. Language mapping with navigated repetitive TMS: Proof of technique and validation. *NeuroImage 82* (2013), 260–272. pages 101

[178] Töpper, R., Mottaghy, F., Brügmann, M., Noth, J., and Huber, W. Facilitation of picture naming by focal transcranial magnetic stimulation of wernicke's area. *Experimental Brain Research 121*, 4 (1998), 371–378. pages 101

[179] Varga, A. P., and Moore, R. K. Hidden markov model decomposition of speech and noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (April 1990), vol. 2, pp. 845–848. pages 11

[180] Veselý, K., Ghoshal, A., Burget, L., and Povey, D. Sequence-discriminative training of deep neural networks. In *INTERSPEECH* (2013), ISCA, pp. 2345–2349. pages 38, 74

[181] Vincent, E., Gribonval, R., and Fevotte, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing 14*, 4 (2006), 1462–1469. pages 37, 93

[182] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research 11* (2010), 3371–3408. pages 118

[183] Vipperla, R., Geiger, J. T., Bozonnet, S., Wang, D., Evans, N. W. D., Schuller, B. W., and Rigoll, G. Speech overlap detection and attribution using convolutive non-negative sparse coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2012), pp. 4181–4184. pages 9

[184] Virag, N. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing 7*, 2 (March 1999), 126–137. pages 5

[185] Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 3 (March 2007), 1066–1074. pages 6, 8, 10, 18, 121

[186] Virtanen, T., Singh, R., and Raj, B., Eds. *Techniques for Noise Robustness in Automatic Speech Recognition.* Wiley, 2012. pages 57

[187] Vitikainen, A., Mäkelä, E., Lioumis, P., Jousmäki, V., and Mäkelä, J. P. Accelerometer-based automatic voice onset detection in speech mapping with navigated repetitive transcranial magnetic stimulation. *Journal of Neuroscience Methods 253* (2015), 70 – 77. pages 102

[188] Wang, D., and Lim, J. The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing 30*, 4 (August 1982), 679–681. pages 3

[189] Wang, Z., and Sha, F. Discriminative non-negative matrix factorization for single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence, Italy, May 2014), pp. 3749–3753. pages 118

[190] Wassermann, E. M., Blaxton, T. A., Hoffman, E. A., Berry, C. D., Oletsky, H., Pascual-Leone, A., and Theodore, W. H. Repetitive transcranial magnetic stimulation of the dominant hemisphere can disrupt visual naming in temporal lobe epilepsy patients. *Neuropsychologia 37*, 5 (1999), 537 – 544. pages 101

[191] Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., and Rigoll, G. The Munich 2011 CHiME challenge contribution:

NMF-BLSTM speech enhancement and recognition for reverberated multisource environments. In *CHiME 2011 Workshop on Machine Listening in Multisource Environments* (September 2011). pages 57

[192] Weninger, F., Le Roux, J., Hershey, J. R., and Watanabe, S. Discriminative NMF and its application to single-channel source separation. In *INTERSPEECH* (Singapore, September 2014), ISCA, pp. 865–869. pages 118

[193] Wu, Z., Virtanen, T., Kinnunen, T., Chng, E., and Li, H. Exemplar-based unit selection for voice conversion utilizing temporal information. In *INTERSPEECH* (2013), ISCA, pp. 3057–3061. pages 28

[194] Wu, Z., Virtanen, T., Kinnunen, T., Chng, E. S., and Li, H. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *The Eighth ISCA Tutorial and Research Workshop on Speech Synthesis* (September 2013), pp. 201–206. pages 9

[195] Yao, K., Zweig, G., Hwang, M.-Y., Shi, Y., and Yu, D. Recurrent neural networks for language understanding. In *INTERSPEECH* (Lyon, France, August 2013), pp. 2524–2528. pages 17

[196] Yilmaz, E., Gemmeke, J. F., Van Compernolle, D., and Van hamme, H. Noise-robust digit recognition with exemplar-based sparse representations of variable length. In *IEEE Workshop on Machine Learning for Signal Processing (MLSP)* (Santander, Spain, September 2012), pp. 1–4. pages 57

[197] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Noise robust exemplar matching with alpha-beta divergence. *Speech Communication 76* (2016), 127–142. pages 8

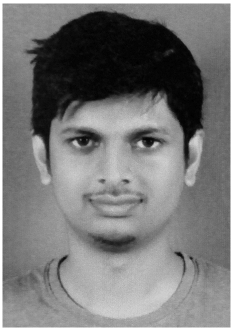
[198] Yoshioka, T., Nakatani, T., and Miyoshi, M. Integrated speech enhancement method using noise suppression and dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing 17*, 2 (February 2009), 231–246. pages 83

[199] Yoshioka, T., Nakatani, T., Miyoshi, M., and Okuno, H. Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 1 (January 2011), 69–84. pages 83

[200] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Noise robust exemplar matching using sparse representations of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 8 (Aug 2014), 1306–1319. pages 86

[201] Zhao, X., Wang, Y., and Wang, D. Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 4 (April 2014), 836–845. pages 83

Short Biography



Deepak Baby was born on 5 March 1988 in Thodupuzha, Kerala, India. He received the Bachelors degree in Electronics and Communication Engineering from College of Engineering, Trivandrum, India in 2009 and Masters degree in Communication and Signal Processing from Indian Institute of Technology, Bombay in 2012. He joined the Processing of Speech and Images (PSI) group, Department of Electrical Engineering, KU Leuven, Belgium as a doctoral student in July 2012.

His research interests include noise-robust automatic speech recognition, machine learning, speech enhancement, dereverberation, statistical models, neural networks and compressed sensing.

List of Publications

Articles in International Journals

[1] **Deepak Baby** and Hugo Van hamme. *Joint Denoising and Dereverberation using Exemplar-based Sparse Representations and Decaying Norm Criterion.* Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016.

[2] **Deepak Baby**, Laura Seynaeve, Patrick Dupont, Wim Van Paesschen and Hugo Van hamme. *An automatic evaluation routine for picture naming task with transcranial magnetic stimulation using machine speech recognition.* Submitted to Journal of Neuroscience Methods, 2016.

[3] **Deepak Baby**, Tuomas Virtanen, Jort F. Gemmeke and Hugo Van hamme. *Coupled Dictionaries for Exemplar-based Speech Enhancement and Automatic Speech Recognition.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 23, No. 11, pages 1788–1799, November 2015.

Articles in International Conferences

[1] **Deepak Baby** and Hugo Van hamme. *Supervised Speech Dereverberation in Noisy Environments using Exemplar-based Sparse Representations.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 156–160, Shanghai, China, March 2016.

[2] **Deepak Baby** and Hugo Van hamme. *Hybrid Input Spaces for Exemplar-based Noise Robust Speech Recognition using Coupled Dictionaries.* 23rd European Signal Processing Conference (EUSIPCO), pages 1676–1680, Nice, France, September 2015.

[3] **Deepak Baby** and Hugo Van hamme. *Investigating Modulation Spectrogram Features for Deep Neural Network-based Automatic Speech Recognition.* Proc. INTERSPEECH, pages 2479–2483, Dresden, Germany, September 2015.

[4] Emre Yılmaz, **Deepak Baby** and Hugo Van hamme. *Noise Robust Exemplar Matching for Speech Enhancement: Applications to Automatic Speech Recognition.* Proc. INTERSPEECH, pages 688–692, Dresden, Germany, September 2015.

[5] Emre Yılmaz, **Deepak Baby** and Hugo Van hamme. *Noise Robust Exemplar Matching with Coupled Dictionaries for Single-Channel Speech Enhancement.* 23rd European Signal Processing Conference (EUSIPCO), pages 874–878, Nice, France, September 2015.

[6] **Deepak Baby**, Jort F. Gemmeke, Tuomas Virtanen and Hugo Van hamme. *Exemplar-based Speech Enhancement for Deep Neural Network based Automatic Speech Recognition.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4485–4489, Brisbane, Australia, April 2015.

[7] **Deepak Baby**, Tuomas Virtanen, Jort F. Gemmeke, Tom Barker and Hugo Van hamme. *Exemplar-based Noise Robust Automatic Speech Recognition using Modulation Spectrogram Features.* IEEE Spoken Language Technology Workshop (SLT), pages 519–524, South Lake Tahoe, USA, December 2014.

[8] **Deepak Baby**, Tuomas Virtanen, Tom Barker and Hugo Van hamme. *Coupled dictionary training for exemplar-based speech enhancement.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2883–2887, Florence, Italy, May 2014.

Abstracts and Technical Reports

[1] **Deepak Baby** and Hugo Van hamme. *Coupled Dictionary-based Speech Enhancement with Adaptively Learned Atoms for the CHiME-3 Challenge.* Abstract, SPIRE Workshop, Groningen, The Netherlands, January 2016.

[2] **Deepak Baby**, T. Virtanen and Hugo Van hamme. *Coupled Dictionary-based Speech Enhancement for the CHiME-3 Challenge.* Technical report KUL/ESAT/PSI/1503, Leuven, Belgium, KU Leuven, ESAT. September 2015.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
CENTER FOR PROCESSING SPEECH AND IMAGES (PSI)
Kasteelpark Arenberg 10
B-3001 Heverlee
deepak.baby@esat.kuleuven.be
http://www.esat.kuleuven.be