Using deep learning methods for supervised speech enhancement in noisy and reverberant environments

Shadi Pir hosseinloo

Submitted to the graduate degree program in Electrical Engineering and Computer Science and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Prof. Shannon Blunt, Chairperson

Prof. Jonathan Brumberg, Co-Chair

Committee members

Prof. Erik Perrins

Prof. Sara Wilson

Prof. John Hansen

Date defended: December 11, 2019

The Dissertation Committee for Shadi Pir hosseinloo certifies
that this is the approved version of the following dissertation :

Using deep learning methods for supervised speech enhancement in noisy and reverberant environments

Prof. Shannon Blunt, Chairperson

Date approved: _____

Abstract

In real world environments, the speech signals received by our ears are usually a combination of different sounds that include not only the target speech, but also acoustic interference like music, background noise, and competing speakers. This interference has negative effect on speech perception and degrades the performance of speech processing applications such as automatic speech recognition (ASR), speaker identification, and hearing aid devices. One way to solve this problem is using source separation algorithms to separate the desired speech from the interfering sounds. Many source separation algorithms have been proposed to improve the performance of ASR systems and hearing aid devices, but it is still challenging for these systems to work efficiently in noisy and reverberant environments. On the other hand, humans have a remarkable ability to separate desired sounds and listen to a specific talker among noise and other talkers. Inspired by the capabilities of human auditory system, a popular method known as auditory scene analysis (ASA) was proposed to separate different sources in a two stage process of segmentation and grouping. The main goal of source separation in ASA is to estimate time frequency masks that optimally match and separate noise signals from a mixture of speech and noise. In this work, multiple algorithms are proposed to improve upon source separation in noisy and reverberant acoustic environment. First, a simple and novel algorithm is proposed to increase the discriminability between two sound sources by scaling (magnifying) the head-related transfer function of the interfering source. Experimental results from applications of this algorithm show a significant increase in the quality of the recovered target speech. Second, a time frequency masking-based source separation algorithm is proposed that can separate a male speaker from a female speaker in reverberant conditions by using the spatial cues of the source signals. Furthermore, the proposed algorithm has the ability to preserve the location of the sources after separation. Three major aims are proposed for supervised speech separation based on deep neural networks to estimate either the time frequency

masks or the clean speech spectrum. Firstly, a novel monaural acoustic feature set based on a gammatone filterbank is presented to be used as the input of the deep neural network (DNN) based speech separation model, which shows significant improvement in objective speech intelligibility and speech quality in different testing conditions. Secondly, a complementary binaural feature set is proposed to increase the ability of source separation in adverse environment with non-stationary background noise and high reverberation using 2-channel recordings. Experimental results show that the combination of spatial features with this complementary feature set improves significantly the speech intelligibility and speech quality in noisy and reverberant conditions. Thirdly, a novel dilated convolution neural network is proposed to improve the generalization of the monaural supervised speech enhancement model to different untrained speakers, unseen noises and simulated rooms. This model increases the speech intelligibility and speech quality of the recovered speech significantly, while being computationally more efficient and requiring less memory in comparison to other models. In addition, the proposed model is modified with recurrent layers and dilated causal convolution layers for real-time processing. This model is causal which makes it suitable for implementation in hearing aid devices and ASR system, while having fewer trainable parameters and using only information about previous time frames in output prediction. The main goal of the proposed algorithms are to increase the intelligibility and the quality of the recovered speech from noisy and reverberant environments, which has the potential to improve both speech processing applications and signal processing strategies for hearing aid and cochlear implant technology.

Acknowledgements

First, I would like to express my sincerest gratitude to my advisers Professor Jonathan Brumberg and Professor Shannon Blunt. This dissertation would not be possible without their guidance, support and involvement. I had the opportunity of working in speech labs and on interesting projects with exciting real-world applications. Throughout these years, Professor Brumberg taught me different aspects of scientific research like hardworking, creativity, consistency and technical writing.

Besides my advisers, I would like to thank Professors Sara Wilson, Professor Erik Perrins, Professor John Hansen for serving as my dissertation committee. Their questions and comments are extremely valuable throughout my PhD. I would also like to thank Professor Hansen for presenting our paper at Interspeech and arranging my visit to his lab at UTD. I would also like to thank Professor Holly Storkel for giving me a chance of teaching multiple courses in SPLH department.

I am fortunate to have two internships in industry at Apple and Knowles Intelligent Audio in Bay area. I want to thank Dr. Kuan Yen for supervising me on their research team at Knowles Intelligent Audio in 2017 and giving me the opportunity to work on noise reduction methods, which helped me to shape my research goals. I would also like to thank Dr. Jonathan Sheaffer for hosting me at Apple Acoustic team in 2019, which broadened my view of deep learning and machine learning in audio research field and give me perspectives on my career path.

Finally, I would like to extend my heartfelt gratitude to my parents. Without their sacrifices, supports, love and encouragement, I would have never made it this far through my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Speech plays an important role in human communication. In everyday listening situations, we hear a mixture of different acoustic signals and the signal reaching our ears is the combination of speech with the background noise and reverberation. It is very challenging to understanding speech when the signal-to-noise ratio is low, especially for hearing impaired listeners or people with hearing aid devices. However, the auditory system of a normal hearing listener has a remarkable ability to extract the individual desired sound and listen to a specific talker in a multi-talker and noisy condition. This capability of human auditory system is called auditory scene analysis (ASA) that analyze the receiving sound in two steps of segmentation and grouping [1]. First, it decomposes the signals into time-frequency units and assumes that each time-frequency unit belongs to one single source. Then it groups the units that belong to the same acoustic source [1].

While normal hearing listeners can easily separate the target speech from the interfering sounds, hearing impaired listeners have many challenges to understand speech in noisy environment. Therefore, it is necessary to develop methods for hearing aid devices and cochlear implants to separate target speech from the background interference. One of the main motivations of this dissertation is to improve the speech quality and speech intelligibility in noisy reverberant environments by using speech separation approaches.

In recent years, there have been a significant increase in human-computer interactions and people interact with devices such as Amazon Echo, Google Home, Apple HomePod and automobiles mainly by using speech. There is a huge demand for using these smart devices and the easiest way

to communicate with them is talking to them (using speech). Therefore, over the years, speech processing community has been studying and proposing many different approaches to improve the speech quality in noisy and reverberant environment. One common method to improve the speech quality in noisy environment is using source separation (in this specific case, speech separation).

Speech separation has many applications in telecommunication, voiced-controlled systems, automatic speech recognition system and speaker identification. Most of these speech interfaces need robust automatic speech recognition (ASR) system. Although there have been many progress in machine listening algorithms over the past decades, the performance of the ASR and speaker identification systems degrades in noisy and reverberant environments. Nonetheless, in clean acoustic conditions, these machines can even outperform human. On the other hand, human listeners are more robust when analyzing noisy and degraded signal. Therefore, another motivation of this work is to use speech separation algorithms as a pre-processing stage to clean the audio signal (speech) before using any human-computer interaction systems.

Over the years, many source separation algorithms were proposed to improve the performance of ASR system and hearing aid devices. The source separation algorithms can be divided based on the number of available microphones on the device to multi-microphone speech separation and monaural (one microphone) speech separation. The performance of the microphone array techniques depends on the number of microphones and array configuration. Furthermore, microphone arrays are not able to separate the sources when the sources are co-located or located near each other. Moreover, the performance of the microphone arrays degrades in reverberant conditions. Blind source separation is another approach for speech separation that make some assumption about the statistical independence of sources and it only works when the number of microphones are equal or greater than the number of sound sources. Therefore, the performance of these two categories depends on the structure of the microphone array, the number of microphones and the statistics of the signals. The other category for source separation is the auditory scene analysis, which is inspired by human auditory system that can also work when only one microphone is available. The utilization of single microphone on a device is more desirable in terms of imple-

mentation and configuration. The other motivation of this dissertation is to separate sources when limited number of microphones are available (monaural case) and make the source separation algorithm independent of the array configuration and assumption about signal statistics.

On the other hand, monaural recording lacks spatial information in comparison to multi microphone methods. Therefore, monaural speech separation algorithms need some prior knowledge about speech and noise. The monaural speech processing is divided to different categories of 1) speech enhancement techniques which make assumptions about the statistics of noise and speech, 2) model based approaches with the goal of learning models from speech and noise, 3) supervised methods (data-driven) such as deep learning that uses neural network to separate target speech from the interference.

The main focus of this dissertation is on source separation methods inspired by human auditory system which is capable of separating sources when the number of microphones are equal or less than the number of sources. The goal of ASA is to separate auditory streams from mixtures where each stream corresponds to an acoustic event [1]. Wang (2005) suggested that the ideal time-frequency (T-F) mask is the computational goal of ASA [2]. The ideal binary mask (IBM) identifies the dominant source in each time frequency unit which can be speech or interference. It has values of 1 and 0, where 1 indicates that the time-frequency unit is dominated by speech while the value of 0 represents the noise-dominant unit [2]. However, if the estimation of the binary mask has errors, the quality of separated speech degrades significantly which has negative effect on speech perception and speech intelligibility. In very noisy environments, most of the time frequency units of speech are masked by noise; therefore, mask estimation is not accurate, which leads to more removal of speech on those time frequency units. Therefore, other forms of time frequency maskings were proposed to alleviate this issue. Ideal ratio mask (IRM) [3] was proposed, which is a soft mask with the values in the range of 0 and 1 and it is defined based on the ratio of speech energy to noisy speech in each time-frequency unit [3]. Therefore, each time frequency unit in IRM represents the ratio of speech energy to the noisy mixture energy [3].

In recent years, supervised speech separation methods based on deep neural networks (DNN)

were proposed. There are multiple advantages of using deep learning for source separation. These machines can learn to map the noisy speech to time frequency masks and the generated masks could be used for source separation. Moreover, these networks can be trained to directly map the noisy speech to clean speech. These data-driven algorithms do not make any assumption about the statistics of signals and they do not depend on the microphone array configuration, which make them suitable for different source separation problems. Nonetheless, supervised speech separation methods using machine learning or deep learning techniques have generalization problem, which means that the the performance of supervised speech separation degrades if the testing condition is very different from the training condition. In real world, there are different types of noises, speakers and acoustic rooms. The supervised methods need to be trained with large amount of training data that includes all these variabilities (different speakers, various types of noise, different room simulations and different SNRs) to have robust separation performance.

1.2 Proposed research

The primary objective of this dissertation is to find methods inspired by human auditory system to separate target speech from interference when limited number of microphones (single microphone or 2 microphones) is available. Firstly, two algorithms are proposed to separate target speech from the interfering speech in reverberant condition using the information about spatial cues of interaural level difference (ILD) and interaural time difference (ITD). The main focus of this dissertation is to use supervised learning to separate target speech from interference such as reverberation and background noise. Particularly, deep neural networks are used as learning machines. Different sections of supervised speech separation such as neural network architecture and acoustic features are investigated for this work.

- *Interaural magnification algorithm for speech separation.* First, human auditory system and the cues that are used by human listeners to separate the sources are discussed. Consequently, the effect of spatial cues for separating the sources are explained. Inspired by remarkable

capability of human auditory system in using spatial cues such as ILD and ITD for source separation, a novel and simple algorithm is proposed that artificially manipulate these cues (enlarge the level difference of interfering source) to increase the discrimination between the interfering sound and target sound, which works directly on head related transfer functions (HRTFs).

- *Time frequency binary mask for blind source separation with preserved spatial cues*. The second proposed algorithm is a modified version of classical DUET algorithm that utilizes the spatial cues of ILD and ITD to define time frequency masks for speech separation in reverberant environment. The proposed algorithm has three stages of speech dereverberation, speech separation (classical DUET algorithm) and cue preservation. In dereverberation stage, the proposed algorithm exploits information about the interaural coherence of 2-channel reverberant mixtures to dereverberate the mixtures before separation. Furthermore, a novel cue preservation algorithm is added to the separation algorithm to preserve the location of sound sources after separation.

- *Exploring monaural acoustic features for supervised speech separation using deep neural networks*. Supervised speech separation based on DNN requires informative and robust acoustic features as an input. Therefore, an extensive feature study was done which includes both spectral acoustic features and gammatone based features for different noisy reverberant environments. A novel monaural acoustic feature based on gammatone filterbank (DMRCG) is proposed for a DNN masking-based speech separation in noisy-reverberant conditions with negative SNR, a variety of reverberation times and different noise types. DMRCG includes both local cochleagram information and spectro-temporal context and associated frequency dynamics. The evaluation of speech separation performance shows that the proposed DMRCG feature outperforms common acoustic features in terms of objective speech intelligibility and speech quality, suggesting a greater preservation of the target speech acoustics in noisy reverberant environment. In addition, our proposed feature is robust to the neural

network type and its performance does not degrade by using more complex neural network architecture.

- *Exploring binaural acoustic features for supervised speech separation using deep neural networks*. Furthermore, another feature study was done for binaural supervised speech separation, where common spatial features such as ILD and interaural phase difference (IPD) are usually used as an input acoustic feature. However, in noisy and reverberant environments, these features alone do not provide informative and discriminative information for the neural network. Therefore, a complementary feature set is proposed that includes the information about the partial cross-correlation between two channels, the energy difference between two channels and the ratio of energy difference to the total energy of channels. All the proposed feature sets are filtered with the 64 channel mel-filterbank to mimics human auditory frequency analysis. This complementary feature set combined with common binaural features of ILD and IPD provides significant improvement in objective speech intelligibility and speech quality in different room simulation with high reverberation times and background noise.

- *Improving the generalization of monaural supervised speech enhancement algorithm to different speakers and noise.* The supervised speech separation algorithms need to generalize well to different unseen conditions to be useful for real-world applications. Including different noises, room simulations and SNRs in training dataset alleviate the generalization issue in DNN based speech separation. However, including multiple speakers in training, DNN performance degrades [4]. Therefore, a novel convolution neural network architecture is proposed for both mapping based and masking based speech enhancement to capture both short-term context and long-term context in time and frequency by masking features of different levels to track the target speaker in noisy and reverberant environment. The proposed model aggregates multi-scale contextual information using dilated convolution with skip connections to integrate features of different levels to estimate final time frequency mask or

clean speech spectrum in unseen conditions where the speakers, background noise and room simulation are different in training and testing. The proposed model increases speech quality and speech intelligibility significantly compared to other models, while it is computationally more efficient by having fewer trainable parameters resulting to less memory requirement and shorter training time.

- *Improving the generalization of real-time monaural speech enhancement algorithm.* Furthermore, the real-time processing of these algorithms is very demanding in design of ASR system, hearing aid devices, headsets and mobile phones. Therefore, a new neural network structure is proposed that modifies the previous proposed model in this dissertation to operate in real-time, which means that the features can only be extracted from current and previous frames. Thus, the available contextual information becomes limited when using previous proposed model. Therefore, a recurrent layer is added to the network to leverage longer contextual information and compensate the limitation of contextual information extracted with convolutional layers. Moreover, the proposed model is speaker and noise independent and its performance is robust to different unseen conditions including challenging non-stationary background noise (Babble noise) and highly reverberant rooms.

1.3 Organization

This dissertation is organized in 7 chapters. The background review about different speech separation algorithms is provided in Chapter 2. The cocktail party problem was first explained and the most common types of speech separation such as beamforming, blind source separation, computational auditory scene analysis and time frequency masking were discussed. Moreover, the most recent speech separation methods using data-driven methods and supervised speech separation were briefly described.

In Chapter 3, first the human auditory system and binaural hearing will be explained. Based on the remarkable ability of human auditory system and binaural cues, two novel speech separation

methods were proposed. In the first section of chapter 3, a novel algorithm is presented that artificially manipulates the binaural cue (increases the interaural level difference) of the interfering sound which increase the separation ability between the target speaker and interfering speaker in both anechoic and reverberant rooms. The second proposed algorithm modifies a popular blind source separation based on time frequency masking (DUET algorithm [5, 6]) to work in reverberant condition while it can also preserve the location of sound sources after separation, which is very useful for detecting sound and understanding speech in challenging environments.

Chapter 4 explains the important sections of DNN-based supervised speech separation methods in more detail. Moreover, an extensive monaural acoustic feature study is presented for monaural speech separation using DNN and LSTM. Finally, a novel monaural acoustic feature based on gammatone filter bank is proposed that has significant improvement in objective speech intelligibility and speech quality for different testing conditions. Furthermore, the generalization of different acoustic features to unseen room simulations, background noise and speakers are examined. All of these features are used to train the neural network to generate an ideal ratio mask for speech enhancement and speaker separation in noisy-reverberant environment.

In chapter 5, the binaural acoustic features for DNN-based speech separation are explained. Common binaural features such as ILD, IPD and IC are usually employed for binaural source separation in non-deep learning techniques. However, in reverberant environment, most of these features are not robust and reliable to be used for training supervised speech separation models such as DNN. Therefore, a novel complementary binaural feature set is proposed to improve the performance of the DNN-based speech separation for estimating time frequency masks in noisy-reverberant environment. Different simulations such as unmatched-room simulation, matched-room simulation and unmatched-room-noise condition are examined to study the generalization of the proposed feature set to unseen conditions.

Chapter 6 describes the main challenges of generalization of supervised speech separation models to different noise type, room simulations and speakers. Ideally, the supervised speech separation model should be noise and speaker independent to be used for real-world applications. First,

LSTM, CNN and dilated CNN are explained in details. Then two novel neural network architectures are proposed for speech enhancement that are noise and speaker independent. Furthermore, these models provide significant improvement in terms of speech intelligibility and speech quality in comparison to other models. While the first model has significant improvement in speech intelligibility, it can only operate in offline applications; however, the second proposed network is capable of operating in real-time while having all the advantages of the first model in incorporating multi-resolution contextual information of the input for final output prediction. Moreover, these two networks are robust to different unseen conditions and they have fewer parameters which is more suitable for implementation and configuration on devices.

Chapter 7 summarizes the findings of this dissertation and explains the contribution of this work in speech processing research field and some future works are listed.

Chapter 2

Literature review

In this chapter, firstly, the cocktail party problem is explained. Over decades, many different methods for solving cocktail party were proposed and they will be discussed briefly in this chapter. Then, the topic of speech separation as a solution for cocktail party problem is explained. Eventually the most important approaches in this field such as spatial filtering, blind source separation and auditory scene analysis are discussed. Furthermore, the auditory scene analysis based speech separation method using time frequency masking is described in detail. In addition, Gammatone filterbank which is mainly used in auditory scene analysis methods is explained. Finally, the most recent supervised speech separation methods based on deep learning techniques are introduced which are the basis of this dissertation.

2.1 Cocktail Party Problem

In difficult listening situations where multiple signals from different sources are mixed together, human auditory system is capable of separating and focusing on one particular speaker (source). This capability of human auditory system is called "the cocktail party problem" [7]. This term can be express as "How can one recognize what another person is saying when other individuals are speaking at the same time?"[7]. Most of the normal hearing listeners do not notice this phenomenon unless they were in a cocktail party in a large room with high reverberation and background noise. In this situation, the speech intelligibility and target speech perception become difficult for normal hearing people. This scenario is even more challenging for hearing impaired listeners who have hearing problems in less noisy situations [8].

It should be noted that for this dissertation, the target speech is defined as the speech coming from the speaker that has the attention of the listener and other sounds such as background noise and reverberation are named as interference or masker. The intelligibility of degraded (noisy) speech could be improved by either separating the target speech from the interference (source separation), by enhancing the target speech (speech enhancement) or by reducing the interfering sounds (noise reduction). In this chapter, the main focus is on the speech separation methods that could improve the quality and the intelligibility of speech in noisy-reverberant environment.

2.2 Source separation

Over the years, many different algorithms have been proposed to solve the source separation problem. Source separation has a wide range of applications in acoustics, biomedical processing, image processing, digital communications and statistics. The most important applications of source separation in acoustics are speech separation, dereverberation, noise suppression, speech enhancement and automatic speech recognition (ASR) [9]. Furthermore, they have uses in mobile telephony, hands-free devices, human-machine interfaces, hearing aids, cochlear implants and airport surveillance [10, 11]. Source separation can be divided to different categories based on some criteria such as:

1. The type of the mixtures: instantaneous, convolutive, anechoic and reverberant mixtures.

2. Number of microphones (mixtures) and sources: under-determined (number of sources are greater than the number of mixtures), determined (equal number of sources and mixtures) and over-determined (number of mixtures are greater than the number of sources).

3. Domain of developing source separation algorithm: time domain, frequency domain and time-frequency domain.

4. Number of channels: single channel (mono) and multi-channel approaches.

5. Availability of a priori knowledge about sources or mixing process such as computational auditory scene analysis (CASA), semi-blind or blind source separation.

In general, source separation can be divided to three main classes: CASA methods, beamforming approaches and blind source separation algorithms.

2.2.1 Beamforming

Beamforming is a signal processing technique that has been used in radar, sonar and telecommunication fields for many years. Beamforming methods separate signals by using spatial filtering where array of sensors are used to allow source signals coming from a specific direction while attenuating signals coming from other directions. Beamforming methods assume that the location of target signal is different than the interfering sources and they are divided to two classes of conventional beamforming and adaptive beamforming.

Initially, beamformers were designed to separate narrowband signals; however, microphone arrays were later adapted for wideband signals like speech. The simplest narrowband beamformer is the delay-and-sum beamformer (DS). DS estimates the time difference of arrival (TDOA) between the signal arriving at the reference microphone and other signals at other microphones. The signals are then delayed by these time differences and summed to cancel the uncorrelated noise.

The authors in [12] described DS as a filter-and-sum technique, where finite impulse response (FIR) filter is applied on each microphone for a given target direction to obtain appropriate frequency-dependent delay. These beamformers are dependent on the direction of target signal and array calibration. Moreover, adaptive beamforming methods were proposed to pass target signal and also reject other interfering signals while being less sensitive to calibration.

The minimum variance distortionless response (MVDR) beamformer (Capon beamformer) [13] was proposed to pass signal coming from a specific direction with no modification while minimizing the output power of beamformer under a single linear constraint on the response of the array [13]. However, MVDR needs matrix inversion which is computationally expensive. Furthermore, the noise covariance matrix needs to be known for MVDR algorithm. Several studies have

been proposed that employed additional linear constraints [14] on beamformers. The linearly constrained minimum variance (LCMV) beamformer is a generalized MVDR that employs multiple linear constrains [15].

A wideband generalization of MVDR adaptive beamformer was proposed by Frost that is based on constrained least-mean-square (LMS) where the transfer function for target signal determines the constraint [16]. An alternative to Frost's algorithm was proposed in [17] which is called generalized sidelobe canceller (GSC) that consists of two parts where the constrained adaptation of MVDR is removed. First part is a fixed beamformer while the second part has a blocking matrix to remove the desired signal and adaptive filters to minimize the output noise power. Finally the output of both parts are subtracted to give the output signal [17]. In GSC method, a single constraint like MVDR beamformer has been used while in another study [18], GSC using LSTM beamformer was developed. In another study [19], the adaptive microphone array system for noise reduction (AMNOR) has been proposed which employs a soft-constraint that allows some distortion on the target signal. The original GSC assumes that received signals at sensors are the delayed version of original signals; therefore, GSC's performance degrades in reverberant rooms. Consequently, the frequency domain GSC (TF-GSC) techniques were proposed to deal with acoustic transfer functions (ATFs) and reverberation [20, 21].

2.2.2 Blind source separation

Blind source separation (BSS) is based on statistical mathematical principles. BSS algorithms recover a set of unknown original signals blindly from a set of observed mixtures [22]. The blind term means that there is little or no prior knowledge about the sources or the mixing structure. However, BSS is based on the assumption that the sources must be mutually statistically independent [23]. BSS techniques could be classified based on the type of the mixture (instantaneous or convolutive) and also the domain where the algorithm works (time domain, frequency domain).

There are some ambiguities in BSS algorithms such as scaling and permutation. Scaling refers to the fact that the variance of the recovered sources can not be exactly calculated because both the

mixing structure and original sources are unknown [9]. Furthermore, the order of the recovered sources can not be determined which is called permutation [9].

In instantaneous mixing model, it is assumed that the original signals are time aligned and they are received at microphones without any delay; therefore, the acoustic impulse responses between sources and microphones are only single coefficient [9]. In other words, the mixtures were assumed to be linear combinations of input signals without any additive noise. Independent component analysis (ICA) was proposed to solve the BSS for instantaneous mixtures [23, 24]. The goal of ICA is to estimate a linear non-orthogonal transformation which maximize the statistical independence between its components [23]. However, there are some limitations in ICA like the number of microphones should be equal or greater than number of sources, there should be no noise in mixtures and the components of the source vector should be mutually statistically independent at each time [9]. On the other hand, in real acoustic environment, the acoustic mixtures are not instantaneous and they include delays. The recorded signals by each microphone is the weighted sum of time delayed versions of original signals. In other words, the acoustic mixtures (convolutive mixtures) are the convolution of the original signals with the acoustic transfer function of the room [9]. Thus, BSS for convolutive mixtures were proposed.

BSS for convolutive mixtures can be grouped to two classes of time domain and frequency domain. The time-domain approaches have high computational costs when the impulse responses are long. Therefore, the frequency domain methods were proposed to use Discrete Fourier transform (DFT) to convert convolution to simple multiplication. It should be noted that the convolution operation can be transformed to multiplication when the frame size of DFT is much longer than the channel filter length. Consequently, the separation task can be decomposed to smaller tasks for each frequency bins which results to efficiency in computation [25]. In [26], ICA has been applied to each frequency bands instead of time domain. However, applying a separate ICA in each frequency band, adds the permutation and scaling ambiguity that need to be solved [25]. Convolutive BSS approaches use both the spatial and signal statistics to separate signals. Nonetheless, both microphone array approaches and BSS methods can separate sources when the number of micro-

phones are equal or greater than the number of sources. On the other hand, CASA based methods works for under-determined condition when the number of sources are greater than number of microphones.

2.2.3 Computational auditory scene analysis (CASA)

The spatial filtering (beamforming) approaches have trouble tracking the moving target or switching between different sound sources. Furthermore, when multiple sound sources are coming from the same location (or located near to each other), spatial filtering can not separate sources. The ICA-based BSS methods make the assumption of statistical independence of sources and their performance on moving speaker is limited. They also have the same limitation of not separating the sources coming from the same location. In this section, another popular approach for source separation called CASA will be discussed that does not have these limitations.

Auditory scene analysis (ASA) is the process that explains how the auditory system forms the acoustic signals into perceptual streams corresponding to different acoustic events [1]. The study of ASA by computational means is defined as computational auditory scene analysis (CASA) [27]. The main idea of CASA is using the intrinsic acoustic features of sounds for source separation instead of employing statistical used in BSS methods. In other words, CASA works like the mechanism of human auditory system. The goal of ASA is to separate auditory streams from mixtures where each stream corresponds to an acoustic event [1]. The ideal time-frequency (T-F) mask is suggested as the computational goal of ASA [2]. In addition, CASA can be categorized to monaural and binaural approaches based on the number of microphones. CASA has applications in fields such as automatic speech recognition, speaker recognition, automatic music transcription, hearing prostheses and audio information retrieval.

2.2.3.1 CASA structure

CASA is a two stage process of segmentation and grouping. In the first stage (segmentation), the acoustic signal passes through peripheral analysis and is decomposed into time frequency regions

by either using gammatone filterbank or short-time Fourier transform (STFT). Then the acoustic features such as onset, offset, amplitude modulation, frequency modulation and periodicity are extracted. Consequently in mid-level representation, the segments are formed from these feature. In the second stage (grouping), the segments are combined into perceptual structure, called streams, based on coming from the same environmental source [27]. The grouping stage can be classified in two categories:

1. Schema-based ASA is a process based on prior information that uses trained models or patterns to group segments [1]. Therefore, features belonging to the same learned pattern (like syllable) are grouped together. Some examples of schema-based cues are knowledge of language and phonetic restoration. Schema-based ASA is a top-down process [1].

2. Primitive based ASA separated speech signals based on intrinsic cues which remain constant over different languages. Cues such as onsets, offsets, amplitude and frequency modulation, spatial cues, periodicity and proximity in frequency belongs to this category [1, 27].

Finally the resynthesis stage is performed on streams to convert the time frequency representation of speech to time domain.

2.2.3.2 Cochleagram- Gammatone Filterbank

Finding the time-frequency representation of audio mixtures is the first step in CASA segmentation stage. This can be achieved by modeling the peripheral auditory system to analyze frequency like human. Generally, the cochlear frequency selectivity is modeled as cascades of filters (filterbank) where each filter models the frequency response of a specific point in cochlea [27]. Gammatone filterbank was proposed to model the impulse responses of the auditory nerve which is estimated as the reverse correlation of spike patterns [28]. The gammatone filter bank models the human auditory system and consists of a series of bandpass filters. The impulse response of the filter is defined as (2.1) in the time domain which is the multiplication of gamma function and a tone[29]

$$g_{f_c}(t) = t^{N-1} e^{-2\pi t b(f_c)} \cos(2\pi f_c t + \phi) u(t)\,, \tag{2.1}$$

where N is the order of the filter, f_c is the center frequency of the filter, ϕ is the phase and $u(t)$ is the unit step function. The $b(f_c)$ determines the bandwidth for each center frequency. The frequency response of the gammatone is approximated by (2.2) for large values of $f_c/b(f_c)$ as [29]

$$G(f) \approx \left[1 + \frac{j(f - f_c)}{b(f_c)}\right]^{-N} . \tag{2.2}$$

Patterson showed that for $N = 4$, the proposed gammatone filter bank fits the experimentally derived human auditory filter shape [30]. The bandwidth of the filter is derived from the Equivalent Rectangular Bandwidth (ERB) of human auditory filters as [31]

$$ERB(f) = 24.7 + 0.108f . \tag{2.3}$$

The center frequencies of the filter are distributed based on their bandwidth. For the fourth-order filter, the $b(f)$ is given as [30]

$$b(f) = 1.019\, ERB(f) . \tag{2.4}$$

In low frequencies, the gammatone filters have narrow bandwidths which leads to having peaks in impulse responses at later time than high frequency filters [27].

For auditory periphery processing, first the input signal is passed through the gammatone filterbanks and decomposed into time-frequency domain. Then the responses of the filter are half-wave rectified and then a square root is applied to the resulting signals. The simulation of the firing activity and the saturation effects of auditory nerve are done with these two steps [32]. Finally, the frames of the processed signals with length of 20 *ms* and 10 *ms* frame shift are generated. This process is referred to cochleagram [27].

2.2.4 Time Frequency Masking

In a room where there is a mixture of multiple signals, it is hard to identify which part of the hearing signal belongs to the target signal and which belongs to the interferers. However, it has

been shown that speech signal is sparse in Time-Frequency (TF) domain [33, 34]. In other words, different speech utterances do not overlap in TF units and they are orthogonal. Therefore, many source separation approaches in time-frequency domain were proposed. However, it should be noted that sparsity property is not valid for speech babble [27].

The Time-Frequency masking is a popular method to separate sources in time frequency domain. This masking implements frequency dependent and time-varying gains to TF units of signal and separates the target signal from the interfering signals [8, 35]. When these frequency dependent and time-varying gains have only the values of 1 and 0, the TF mask is called the binary mask [8, 35].

The STFT or windowed auditory filter band in form of the cochleagram represents the signal in TF domain [35]. The TF masking can be seen as a filter which is placed on the mixture to keep some TF units and remove the other units. Therefore, the TF masking system includes three stages: first the STFT or gammatone filter is applied to the signal to have a TF representation. Secondly, a TF mask is defined based on different separation algorithms and applied to the signal in TF domain. In the final stage, the signal is synthesized and converted back to the time domain [35]. The TF masking can be estimated based on CASA or ICA. Masks based on CASA use the principles of human auditory scene analysis to separate signals based on cues such as pitch, onset/offset, harmonicity and temporal continuity [8, 1]. On the other hand, ICA assumes that signals are statistically independent and the algorithm estimates a demixing matrix to solve separation problem [36].

Based on the number of audio mixtures, TF masking can be divided into two groups of monaural and binaural masking. Monaural TF masking algorithms typically use intrinsic sound properties such as harmonicity, onset and offset, amplitude and frequency modulation and temporal continuity[35] for source separation. A model proposed in [37] used a gammatone filterbank and segmented the signal based on frequency modulation, pitch and onset/offset while grouping based on pitch contour. In another study [38, 39], auditory segmentation was performed based on cross-channel correlation and temporal continuity and grouping by dominant pitch. Most monaural TF

masks are estimated based on CASA, while the binaural TF masks are developed in both CASA and ICA field [35].

In binaural TF masking, the masks are estimated based on the time, phase or level differences of mixtures recorded by the two microphones. The proposed separation system in [32], estimated the binaural cues of interaural level difference (ILD) and interaural time difference (ITD) for each TF pair and used the classification based on a maximum a posteriori (MAP) decision rule to estimate the binary mask at each ear.

On another study, the degenerate unmixing estimation technique (DUET) algorithm was proposed based on the orthogonality principle of speech and used the binaural cues of ILD and ITD with an unsupervised clustering algorithm to construct a binary mask for separating signals [5, 6]. Furthermore, there were several studies that combined the TF masking with ICA [40, 41, 42] or use the combination of beamformer with TF masking [43, 44].

An algorithm was proposed in [45] to separate one source from a mixture with binary mask and then separate the remaining signals with ICA . In another study [46], first the ICA was applied on the mixtures and then a binary mask was used to separate the remaining signals. In study [43], an adaptive beamformer was employed to provide the basis of the binary mask. The target signal from a known direction was canceled and the ideal binary mask was constructed by comparing the mixture signal and the output of the beamformer. It should be noted that the binary mask adds a distortion to the processed signal which shows strong fluctuations in TF domain [47]. This distortion is called musical noise.

2.2.4.1 Binary Masking

In the TF domain, the binary mask assigns the signals as either target or interferer. When the mask value equals 1 in a TF unit, it means that the acoustic energy of that TF unit corresponds mostly to the target signal and the mask should keep that unit. On the other hand, the mask value of zero shows that the energy of that unit belongs to the interferer and it should be removed. In other words, the TF regions assigned to the target signal are kept while the TF regions corresponding to

the interferer signals are removed [8]. Over the past decade, several time frequency masks were presented such as ideal binary mask (IBM) [35], ideal ratio mask (IRM), target binary mask (TBM) [8], spectral magnitude mask (SMM) [3] and phase sensitive mask (PSM) [48] and they will be discussed in more detail in the next chapters.

2.2.4.2 DUET algorithm

The DUET algorithm separates the sources by clustering different time-frequency points based on their interaural parameters using a 2-channel audio mixtures. This algorithm is based on W-disjoint orthogonality which states that every time-frequency unit in the mixture is dominated by only one source which can be expressed as

$$\hat{s}_j(\tau,\omega)\hat{s}_k(\tau,\omega) = 0 \qquad \forall \tau,\omega, \qquad \forall j \neq k\,, \tag{2.5}$$

where the $\hat{s}_j(\tau,\omega)$ is the STFT of $s_j(t)$ signal [5, 6].

At the first step, the time-frequency representation of both audio mixtures are generated by STFT. Then the interaural parameters of ILD and ITD are estimated based on the ratio of the two mixtures and the phase differences. This method essentially constructs a two-dimensional smoothed histogram of interaural parameters, with points weighted by their respective energy, and then selects each prominent peak in the histogram as the interaural parameters of each source [5, 6]. Next, the time-frequency binary masks are reconstructed based on the interaural parameters and are applied to the mixtures to calculate the original source estimates. Finally, the estimated sources are converted back to the time domain with invserse short time fourier transform (ISTFT) [5, 6].

The DUET algorithm is capable of recovering the original signals blindly from two mixtures even when the number of sources are more than the number of microphones. However, the performance of this algorithm degrades significantly in reverberant condition.

2.2.5 Speech Separation based on machine learning

Over the last decade, many techniques based on machine learning were introduced in the field of speech processing and speech recognition, which categorize speech separation algorithms into two categories: Model based speech separation for monaural case and supervised speech separation which can be used in both monaural and binaural conditions.

2.2.5.1 Model based separation

In model-based methods, machine learning techniques were used to build structures for speech and interference. One of the most popular model-based methods for speech separation is non-negative matrix factorization (NMF) [49, 50]. In NMF, the power spectrum or the magnitude of training data is decomposed into two matrices, a basis matrix and weight matrix. It is assumed that both basis and weight matrices are non-negative. In testing, the trained basis matrix is used to re-estimate the weight matrix and consequently reconstruct the speech signal from the noisy speech. The speech model and weight matrix can be used either as signal estimate or Wiener mask to be applied to the noisy speech. Several studies show the effectiveness of NMF in speech separation [51, 52, 53] and automatic speech recognition [54]. However, recent studies show that DNN based source separation outperforms NMF when the interference is a broadband non-stationary noise because the modeling of this noise with statistical modeling is hard [3, 55, 56]. Moreover, NMF needs a separate model for each of the sources in observed mixture and it has high computational cost which is a problem in real-time implementations.

Recently, some methods were proposed to use NMF as non-negative factorial hidden Markov model (N-FHMM) [57] and non-negative HMM (N-HMM) [58] to model speech. In N-FHMM, the Wiener mask is generated to separate the speech from the noise [57]. In N-HMM, several small dictionaries were used to model spectral structure of speech while HMM was employed for modeling the temporal dynamics of speech [58]. However, these methods require separate dictionaries for different sound sources.

2.2.5.2 Supervised speech separation

The supervised speech separation can be treated as a classification or regression problem to either find a binary decision or estimate a continuous value. In early research of supervised learning using neural network, multi-layer perceptrons (MLPs) was proposed to map the noisy speech to clean speech in time domain [59, 60]. In another study, the log-power spectra of clean speech was estimated by MLP [61].

In other studies [62, 63], Gaussian mixture models (GMMs) were proposed as classifiers. In [62], a Bayesian classifier was used to estimate a mask for ASR system while in [63], GMMs and Bayesian classifier were used to estimate Ideal binary mask (IBM) where the input features are AMS and its deltas. This method demonstrates improvement in speech intelligibility for normal hearing listeners; nonetheless, these improvements were achievable only when the noise in training and testing stage matches.

Another IBM estimation method was proposed in [64, 65] where Support vector machines (SVMs) and rethresholding techniques were employed to solve the noise mismatch problem. In this work, the training features were AMS and pitch. In other studies [66, 32, 67], maximum a posterior (MAP) classifier was developed to estimate the IBM for binaural speech separation where the input features for the classifier were the interaural level difference (ILD) and interaural time difference (ITD). These studies show a significant increase in speech intelligibility when the training and testing condition matches.

Furthermore, deep neural network (DNN) based supervised source separation methods have been proposed in recent years that demonstrate better performance over SVM for IBM estimation [68, 69]. It should be noted that DNN refers to any neural network that has at least two hidden layers. In supervised DNN speech separation, training target (label) is important in training and generalization and it is classified to two different categories: masking based targets and mapping-based targets [70]. The mapping-based targets are the spectral representations of the clean speech, while the masking based targets show the time frequency representation of clean speech to interference [70]. In other words, in mapping based method, DNN estimates the clean speech while

in masking based method, the time-frequency masks were estimated and then applied to the noisy speech to separate the speech from the interference.

In recent years, many DNN based speech separation systems have been proposed. In [71], the DNN-based IBM estimation provide large improvements in speech intelligibility for normal-hearing and hearing-impaired listeners. Moreover, the DNN based ratio ideal mask (IRM) has been proposed in [72] that outperforms DNN based IBM for ASR system. In addition, masking based estimation works better in speech denoising than mapping based estimation using DNN based speech separation algorithms [3].

DNN based speech separation can be used in both monaural and binaural separation. In [73], binaural features such as ILD and ITD with monaural features were employed as an input to DNN for binaural separation. This system has good generalization for unseen spatial configuration. DNN based separation has become a successful method for speech separation in real-time implementation since they have fast processing. However, the main problem for DNN is generalization. In the following chapters, the most important parts of DNN-based speech separation system such as acoustic features, DNN structure and training target will be discussed in detail. New acoustic features for binaural speech separation and monaural speech separation will be introduced. Finally, two new neural network architecture will be presented as noise and speaker independent models to separate target speech from background noise in noisy-reverberant environment.

Chapter 3

Speech separation based on human auditory system

In this chapter, the human auditory system will be explained and the important binaural features that human use for hearing are mentioned. Two new algorithms are proposed to separate target speaker from the interfering speaker. These algorithms employs the useful information from binaural features to separate the target speech in anechoic and reverberant condition. The first algorithm has prior knowledge about the head related transfer function of the sound sources and artificially manipulate them to make the interfering sound far from the target sound. The second algorithm uses a blind approach to separate two speakers from a 2-channel reverberant signal using time frequency masking. Moreover the proposed method is capable of preserving the information about the location of sources after separation.

3.1 Human auditory system

The human auditory system acts as a transducer that converts vibrations caused by sound to action potentials in the auditory nerve. This system is divided to three areas: outer, middle and inner ear. The outer ear consists of pinnae, ear canal and eardrum where the ear canal connects the outerear to eardrum [74]. The pinnae acts like an auditory filter that provides localization cues of sounds. When sound enters ear canal, it causes the eardrum to vibrate. These vibrations are transferred to cochlea by middle ear with three tiny bones(ossicles) where these bones act as impedance matching from air pressure to fluid in cochlea [74].

Cochlea is the main part of inner ear which has a fluid-filled tube that is divided by two membranes. These two membranes are called Reissner's membrane and basilar membrane. The basilar

membrane's mass and stiffness varies along its length; therefore, different parts of it resonates at different frequencies [74]. The base end of cochlea is connected to two windows of oval window and round window. So when the sound comes from the middle ear, the last bone (stapes) press the oval window; therefore, the fluid in cochlea starts moving and the wave travels through the basilar membrane. Furthermore, the round window is pushed outwards [74]. It should be noted that the basilar membrane is stiff and narrow at the base while it is wider at apex. Thus, the maximum displacement of basilar membrane happens at the apex for low frequency while, for higher frequencies, there is more vibration at the base of this membrane [74].

Finally, the inner hair cells transduce the movement of basilar membrane to neural activity which will be send to the brain. In other words, when the basilar membrane moves, the hair cells begin to bend back and forth and create action potentials [74]. It should be noted that the auditory nerve uses the half-wave rectification of sound because the movement of hair cells in only one direction generates action potentials [74].

3.2 Binaural Hearing

When we listen to different sounds in an environment, we can easily separate these sounds and focus on a specific sound by hearing with two ears. This binaural process (hearing with two ears) has important role for sound separation and sound localization. Furthermore, binaural hearing improves speech intelligibility in noisy and reverberant environment. In comparison to microphone array localization and separation which are dependent on the number of microphones and array configuration, human auditory system has remarkable ability to analyze complex acoustic scene, localize target sound and separate the desired sound based on binaural signals [75, 76]. Furthermore, this system is robust when there are multiple talkers in a room.

Human auditory system employs two major cues called interaural time difference (ITD) and interaural level difference (ILD) for acoustic source localization and source separation. The sound coming from the source arrives at closer ear before it arrives to the further ear. The time difference of arrival is called ITD. Furthermore, the loudness (level) of the sound at nearer ear is also higher

than the sound reaching the farther ear and this difference between the level is referred to ILD. If both the ILD and ITD are zero, we can conclude that the sound source is in front of the listener, while a ILD and ITD greater that zero indicate that the sound source is on one side of the listener. It should be noted that the human auditory system calculates these binaural cues to find the location of the sound [75].

One of the classical methods in ITD measurement is finding the main peak in the generalized cross-correlation (GCC) function between the left and the right ear signal [77]. The other technique to calculate ITD is applying FFT on binaural signals and find the interaural phase difference (IPD). The shadowing effect of the head creates interaural level difference. In other words, head produces an acoustic shadow for the ear on the opposite side of the sound source; therefore, the energy of the sound arriving the farther ear is reduced compared to the sound arriving to the near ear. When the sound reaches the head, it is diffracted and its energy attenuates. This is called head shadow and the amount of energy reduction is dependent on the sound's wavelength.

In higher frequencies, the wavelength is smaller than the diameter of head which leads to more reflection of sound; therefore, the ITD becomes ambiguous [75, 76]. For frequencies less than 800 Hz, only ITD is reliable; nonetheless, ILD is more reliable at frequencies higher than 1.5 KHz [75, 76]. The value of ITD depends on the angle of arrival of the source, frequency and distance from the sound source. The ability to localize sounds in vertical plane and solve the front-back ambiguities is provided mainly by the outer ear (pinnae). The frequent model for representing the auditory nerve response to the sound is a bank of linear bandpass filters (like cochlear frequency selectivity) with nonlinear operations such as half-wave rectification and nonlinear compression. Therefore, the auditory system extract ILD and ITD of sound based on the output of these bandpass filters instead of using the original signal [27].

3.3 Room Impulse Response

The binaural signals arriving at human ears can be created by convolving the source signal with the impulse responses corresponding to the source position. Assuming that the propagation of

sound from the fixed source to ear is linear and time invariant, the impulse response can be used to model the system [78]. In anechoic rooms, the impulse responses only contain the information about human head and torso; therefore, they are called head related impulse responses (HRIR). In addition, the Fourier transform of HRIR is called head related transfer function (HRTF). In binaural condition, the transfer function between the sound source and the human ears are called binaural room impulse response (BRIR) [78]. This transfer function includes the information about the room, the head, the torso and the human shoulders. Moreover, it contains the information about the position of the sound source relative to listener's ear (or microphone) position. Therefore, the BRIR for left ear and right ear can be written as follows

$$h^{\mathrm{L}}(t) = h^{\mathrm{L}}_{\mathrm{dp}}(t) + h^{\mathrm{L}}_{\mathrm{e}}(t) + h^{\mathrm{L}}_{\mathrm{rev}}(t) \quad \text{and} \quad h^{\mathrm{R}}(t) = h^{\mathrm{R}}_{\mathrm{dp}}(t) + h^{\mathrm{R}}_{\mathrm{e}}(t) + h^{\mathrm{R}}_{\mathrm{rev}}(t)\,, \tag{3.1}$$

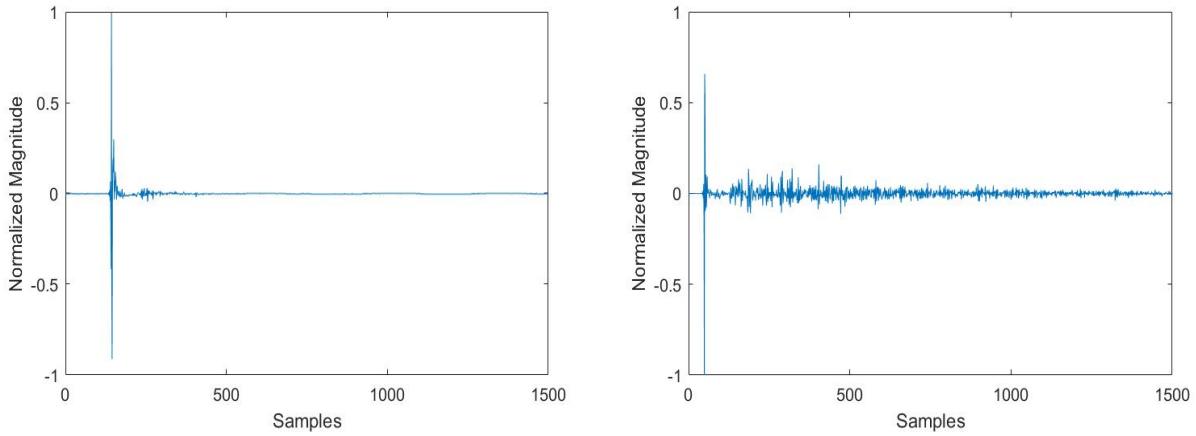
where $h^{\mathrm{L}}_{\mathrm{dp}}(t)$ is the direct path impulse response of left ear while $h^{\mathrm{L}}_{\mathrm{e}}(t)$ and $h^{\mathrm{L}}_{\mathrm{rev}}(t)$ are the early reflection and late reflection parts of impulse response respectively. The pressure of sound at the left ear $y_{\mathrm{L}}(t)$ and the right ear $y_{\mathrm{R}}(t)$ can be written as

$$\begin{aligned} y_{\mathrm{L}}(t) &= h_{\mathrm{L}}(t) * x(t) \quad or \quad Y_{\mathrm{L}}(\omega) = H_{\mathrm{L}}(\omega)\, X(\omega) \\ y_{\mathrm{R}}(t) &= h_{\mathrm{R}}(t) * x(t) \quad or \quad Y_{\mathrm{R}}(\omega) = H_{\mathrm{R}}(\omega)\, X(\omega) \end{aligned}, \tag{3.2}$$

where $H_{\mathrm{L}}(\omega), H_{\mathrm{R}}(\omega)$ are the HRTFs for the left and right ear and $x(t)$ is pressure of the sound source. It should be noted that the position of the sound source depends on three variables of azimuthal angle θ, (on the left or right side of the listener), the elevation angle ϕ (on vertical axis) and the distance between the listener's head and sound source. The BRIR is composed of three parts: the direct path, early reflections and late reflections (reverberations), which can be seen on the right section of Fig. 3.1.

Direct-path

The sound coming directly from the source is called direct path. In an anechoic room, BRIR only



(a) Anechoic impulse response

(b) Reverberant impulse response

Figure 3.1: The impulse response captured in an anechoic room and the reverberant room (office) for the left ear using a dummy head and Oldenburg HRTFs [78] when the sound source is in front of the head.

has direct path part. The direct path impulse response has the information about the location of source (direction of source); therefore, it is used for source localization [79]. Direct-path depends on the azimuth and elevation of source and ears. In addition, it depends on the human body (head,torso,pinnae and shoulders) [80].

Early reflection

The second part of impulse response is early reflections which come immediately after direct-path signal. They usually arrive within 1 to 50 ms after the direct-path signal and they contain information about the geometry of the room (walls and ceiling) [79]. Researchers in [81, 82] have shown that these early reflections increase the energy of speech arriving to the listener and therefore improve the speech intelligibility. It should be noted that early reflections depends less on source/listener movement compared to direct-path signal. Moreover, they are also less dependent on the distance between the source and the listener.

Late Reflection

Late reflections come after early reflections and they are made of multiple reflection and scattering

signals from the walls and the objects in the room. Therefore, they contain the information about the size of the room and the materials in the room and walls. The energy of late reflections has a uniform distribution and stays relatively constant. Late reflections are very sensitive to the motion of source or listener. It should be noted that the late reflections have negative effect on speech intelligibility [83].

3.4 Interaural magnification algorithm for enhancement of naturally-occurring level differences

In many everyday listening situations, a listener's goal is to hear out a specific sound of interest (target) from amongst a mixture of other interfering sounds. Despite the fact that all of these individual sounds are summed up into a single acoustic waveform, the binaural hearing system can very efficiently separate between different voices in a noisy environment and solve what has been coined as the binaural (or two-eared) cocktail-party problem [7]. This separation is accomplished by the use of binaural cues, such as interaural differences in level and time. Interaural level differences (ILDs) are the differences in the overall intensity or level of the signals received at the two ears. A signal with higher intensity at the left ear is perceived as a sound source located to the left of the listener. Interaural time differences (ITDs) refer to the different arrival times of signals at each ear due to the spatial separation of the two ears. A signal that reaches the left ear earlier than the right ear will be perceived as a sound source located to the left of the listener. Typically, ILDs are more informative regarding azimuth locations at frequencies above 3 kHz and ITDs below 1.5 kHz [76].

In order to better describe these effects, consider the typical scenario where the target source is located right in front of the listener, whereas another source (interferer) is located to the right of the listener. In this scenario, the sounds reaching each ear are transformed in a directionally dependent manner through filters associated with each ear. These linear time-invariant transfer functions can accurately capture the direction dependent effects of the head on the signals received at the two ears and are commonly known as the head-related transfer functions (HRTFs). To generate a binaural signal, the HRTFs are convolved with an input acoustic signal, generating a stereo signal with binaural cues associated with a source from a specific azimuth relative to the listener. For a source in front of the listener, there is very little difference in either the magnitude or the phase responses for both ears. For another source placed to the right of the listener, the magnitude of the source in the right ear is greater than the one on the left, while the time delay in the left ear is longer (i.e.,

the sound arrives at the right ear sooner than in the left ear).

In order to obtain better speech intelligibility even at poor signal-to-noise ratios (SNRs), normal-hearing listeners often take advantage of these perceived differences in magnitude and the fact that in most cases the target and competitors are spatially separated. This benefit, known as spatial release from masking (SRM), is fairly robust in normal-hearing listeners and has been well-established in the literature with many researchers demonstrating that speech perception is markedly better when the speech source is spatially separated from the interfering noise rather than co-located (e.g., see [84]).

To increase discriminability between two competing sound sources, an elegant approach is to artificially increase the deviation of the competing sound source from the midline by frequency scaling of the space filters or the head-related transfer functions [85]. The rationale is that this processing would ultimately enhance one's ability to use auditory spatial cues in psycho-physical tasks and in understanding speech in a noisy environment. This processing algorithm is referred to as the interaural magnification (IM) approach [85]. Theoretically, this interaural magnification procedure is equivalent to artificially enlarging the diameter of the listener's head. Such an enlarged head would in principle magnify both naturally-occurring interaural amplitude differences and interaural time differences [86].

A variant of this approach, has been previously applied to the spectra of the signals received by the two ears instead of the spatial filters [87]. The authors demonstrated a significant increase in binaural masking level difference particularly in listeners with hearing impairments. Other similar studies in human auditory perception, have shown that listeners can eventually adapt to such unnatural (altered or re-mapped) auditory spatial cues, which can, in theory, provide better than normal localization ability (e.g., see [88], [89]).

In Section 3.4.1, the proposed interaural magnification algorithm [90] is discussed for separating target speaker from the interfering speaker in anechoic and reverberation environment. The effect of interaural magnification on speech quality is evaluated for both anechoic and reverberant environment. The complex acoustic mixture perceived binaurally is processed by magnifying the

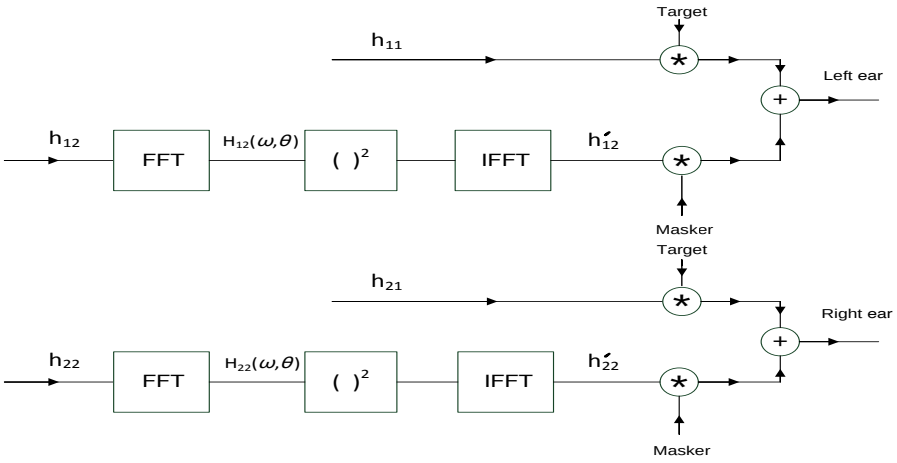


Figure 3.2: Block diagram of the proposed interaural magnification algorithm.

interaural level difference cues corresponding to the interfering sound source.This leads to a lateral spreading of the interfering source, which ultimately increases speech perception [91].

3.4.1 Interaural magnification algorithm

The IM algorithm for a two input-signal and two-output system configuration is shown in Fig. 3.2. First, the mixture signals perceived in each ear are generated by convolving the target source and the interferer with the binaural direction-dependent HRTFs corresponding to a particular azimuth location. The frequency-domain inputs $Y_L(\omega,\theta)$ and $Y_R(\omega,\theta)$ representing the left ear and right ear signal, respectively, at frequency ω and azimuth angle θ, can be written as follows

$$\begin{aligned} Y_{\mathrm{L}}(\omega,\theta) &= H_{11}(\omega,\theta)S(\omega,\theta)+H_{12}(\omega,\theta)N(\omega,\theta) \\ Y_{\mathrm{R}}(\omega,\theta) &= H_{21}(\omega,\theta)S(\omega,\theta)+H_{22}(\omega,\theta)N(\omega,\theta) \end{aligned}, \tag{3.3}$$

where $S(\omega,\theta)$ denotes the source signal and $N(\omega,\theta)$ is the interferer. Furthermore, $H_{11}(\omega,\theta)$, $H_{12}(\omega,\theta)$, $H_{21}(\omega,\theta)$ and $H_{22}(\omega,\theta)$ denote four linear-time invariant filters corresponding to the HRTFs in our experiment. Note that in (3.3) the convolution operations are transformed into efficient multiplication operations by setting the frame size of the fast Fourier transform (FFT) to be much longer than the filter length. Focusing on the interfering source, we define the ratio of

$H = H_2/H_1$ as the interaural transfer function (ITF), which is equal to [78]

$$ITF(\omega,\theta) = \frac{H_{22}(\omega,\theta)}{H_{12}(\omega,\theta)} \,. \tag{3.4}$$

Eventually, the interaural level difference can be extracted from the ITF as follows

$$\mathrm{ILD}(\omega,\theta) = 20\log_{10}(|\mathrm{ITF}(\omega,\theta)|)\,. \tag{3.5}$$

As depicted in Fig. 3.2, the filters h_{ij} corresponding to the impulse responses between the j^{th} source and the i^{th} ear, which are first converted to the frequency-domain. Secondly, the HRTFs describing the acoustic transfer functions are magnified by power of n and are then converted back to the time-domain. Finally, the enhanced outputs are calculated by the convolution of the target signal and the interferer signal with the processed or interaurally-magnified HRTFs. The magnified HRTFs are estimated according to

$$\tilde{H}_{12}(\omega,\theta) = [H_{12}(\omega,\theta)]^n \quad , \quad \tilde{H}_{22}(\omega,\theta) = [H_{22}(\omega,\theta)]^n \,, \tag{3.6}$$

where $\tilde{H}_{12}(\omega,\theta)$ and $\tilde{H}_{22}(\omega,\theta)$ are the modified HRTFs in the frequency-domain at frequency ω and azimuth angle θ and exponent n denotes the magnification power, which is equal to 2 in this work. The processed (enhanced) outputs can be estimated by the interaurally-magnified HRTFs as follows

$$\begin{aligned} \tilde{Y}_{\mathrm{L}}(\omega,\theta) &= H_{11}(\omega,\theta)S(\omega,\theta) + \tilde{H}_{12}(\omega,\theta)N(\omega,\theta) \\ \tilde{Y}_{\mathrm{R}}(\omega,\theta) &= H_{21}(\omega,\theta)S(\omega,\theta) + \tilde{H}_{22}(\omega,\theta)N(\omega,\theta) \end{aligned} \,, \tag{3.7}$$

where $\tilde{Y}_{\mathrm{L}}(\omega,\theta)$ and $\tilde{Y}_{\mathrm{R}}(\omega,\theta)$ are the modified signals for the left and right ear, respectively.

The modified interaural level difference $\tilde{\mathrm{ILD}}(\omega,\theta)$ is defined as:

$$\tilde{\mathrm{ILD}}(\omega,\theta) = n\,[20\log_{10}(|\mathrm{ITF}(\omega,\theta)|)]\,. \tag{3.8}$$

According to (3.8), the interaural level difference corresponding to the noise source, is multiplied by a factor of n, which is expected to increase the lateral spreading of the interfering source and improve the overall benefit due to spatial release from masking [90].

3.4.2 Experimental results on IM algorithm

The performance of the proposed interaural magnification algorithm was evaluated on a test set of 10 speech signals comprised of a single randomly selected male-spoken sentence. A female interferer was used with a root-mean-square value equal to the target source, such that the input SNR $= 0$ dB. The duration of each speech signal was approximately 3 s. All signals were recorded at a sampling rate of 22,050 Hz. To generate the speech test stimuli, IEEE database was used, which consists of phonetically balanced sentences, with each sentence being composed of approximately 7 to 12 words [92]. All signals had the same onset and were normalized to their maximum amplitude before convolving with the HRTFs.

Anechoic head-related impulse responses were used to simulate a non-reverberant listening condition. To simulate a more realistic scenario, a second set of reverberant head-related impulse responses were measured inside a typical office with reverberation time equal to RT_{60} = 300 ms, which is a typical value for a moderately reverberant environment. Both sets of impulse response measurements were conducted in the University of Oldenburg [78]. For each listening scenario, a total of four sound source locations were calculated for sound sources located 1 m away from the center of the listener in the azimuthal plane for every angle from 0^0 (i.e., straight ahead) to $+90^0$ to the right of the listener in 30^0 increments. In all cases, the target source was placed directly in front of the listener at 0^0, such that the ITF corresponding to the source is 1.

3.4.3 Perceptual evaluation of speech quality

The overall quality of the enhanced output binaural signals described in (3.7) was assessed with the perceptual evaluation of speech quality (PESQ) score [93]. The PESQ employs a sensory model to compare the original (unprocessed) with the enhanced (processed) signal, which is the output of

Table 3.1: PESQ input and output values for each azimuth location averaged over 10 IEEE sentences. The standard errors of the mean are inside the parentheses.

PESQ (anechoic)	0°	+30°	+60°	+90°
Left ear input	2.84 ($\pm$0.21)	3.11 ($\pm$0.17)	3.28 ($\pm$0.16)	3.17 ($\pm$0.17)
Right ear input	2.84 ($\pm$0.21)	2.60 ($\pm$0.15)	2.49 ($\pm$0.30)	2.62 ($\pm$0.17)
Left ear output	2.83 ($\pm$0.19)	4.23 ($\pm$0.09)	4.44 ($\pm$0.06)	4.30 ($\pm$0.06)
Right ear output	2.83 ($\pm$0.19)	3.39 ($\pm$0.11)	3.33 ($\pm$0.12)	3.49 ($\pm$0.09)
PESQ (reverberant)	0°	+30°	+60°	+90°
Left ear input	2.98 ($\pm$0.17)	3.12 ($\pm$0.16)	3.23 ($\pm$0.16)	3.15 ($\pm$0.18)
Right ear input	2.88 ($\pm$0.18)	2.79 ($\pm$0.19)	2.63 ($\pm$0.27)	2.69 ($\pm$0.27)
Left ear output	2.98 ($\pm$0.17)	4.03 ($\pm$0.08)	4.16 ($\pm$0.06)	4.05 ($\pm$0.08)
Right ear output	2.88 ($\pm$0.18)	3.42 ($\pm$0.10)	3.36 ($\pm$0.12)	3.37 ($\pm$0.12)

the IM algorithm, by relying on a perceptual model of the human auditory system. The PESQ score has been shown to exhibit a high correlation coefficient (Pearson's correlation) of $r = 0.91$ with subjective listening quality tests [94]. The PESQ measures the subjective assessment quality of the dereverberated speech rated as a value between 1 and 5 according to the five grade mean opinion score (MOS) scale. Table 3.1 compares the performance of the proposed algorithm in terms of PESQ, relative to the performance of the unprocessed binaural inputs for each ear. The score for the anechoic (unprocessed) sound source when it is co-located with the masker, averaged across 10 different sentences is equal to 2.84 (left and right), which suggests that a relatively high amount of degradation is present. In contrast, after processing the binaural signals with the proposed IM algorithm, the average scores in the left ear increase to 4.23, 4.44 and 4.30 for azimuths of 30^0, 60^0 and 90^0, respectively. In the reverberant conditions, after processing the binaural signals with the IM algorithm, the average scores in the left ear increase to 4.03, 4.16 and 4.05 for spatial locations corresponding to 30^0, 60^0 and 90^0, respectively. The estimated PESQ scores in both the anechoic and reverberant scenarios, suggest that the proposed algorithm improves the speech quality of the signals considerably, while keeping signal distortion to a minimum.

3.4.4 Discussion

An interaural magnification algorithm is developed and tested that can be used for binaural speech enhancement in noise and reverberation [90]. The proposed algorithm operates by magnifying the interaural level differences corresponding to a spatially separated interfering sound source. Experiments carried out with speech signals masked by a single interfering source in both anechoic and reverberant scenarios indicate that the proposed technique is capable of improving the speech quality of the signals considerably, while keeping signal distortion to a minimum. A limitation of the proposed technique is that prior knowledge of the head related transfer functions is available, which listeners use to understand and localize incoming sounds. Thus, for a practical implementation, pre-measure personalized HRTFs will be needed.

3.5 Time frequency binary mask for blind source separation with preserved spatial cues

A new algorithm is proposed to address the problem of blind speech separation (using no prior knowledge about the sources) by relying on time-frequency binary masks to segregate binaural mixtures. Moreover, the proposed algorithm is capable of separating target speaker from reverberant mixtures and extract the original sound sources while preserving their original spatial locations. The performance of the proposed algorithm is evaluated objectively and subjectively, by assessing the estimated interaural time differences versus their theoretical values and by testing for localization acuity in normal-hearing listeners for different spatial locations in a reverberant room. Experimental results indicate that the proposed algorithm is capable of preserving the spatial information of the recovered source signals while keeping the signal-to-distortion and signal-to-interference ratios high.

3.5.1 DUET algorithm

In realistic listening situations, human listeners excel at hearing out a specific sound of interest (target) from amongst a mixture of other interfering sounds. Inspired by this robust performance, research has been devoted to build speech separation systems that incorporate the known principles of auditory perception [1]. According to this theory, listeners perform segregation of a binaural mixture in a two-stage process. In the first stage, the acoustic input is analyzed to form time-frequency (T-F) segments, while in the second stage, listeners group sound elements based on whether those originate from the same locations (likely to come from a common source) or spatially distributed locations (likely to come from two different sources) [75]. To facilitate this latter process, listeners may rely on the differences in the overall intensity or level of the signals received at the two ears, known as interaural level differences (ILDs) and the different arrival times of signals at each ear due to the spatial separation of the two ears, known as interaural time differences (ITDs) [75, 95].

The techniques of separating individual sound sources from a mixture are known as blind source separation and computational auditory scene analysis (CASA). Both fields, have become popular in the recent decades and a number of methods have emerged from the study of this problem, most of which perform well for certain types of sources, such as speech (e.g., see [6, 96]). Speech is sparsely distributed in the time-frequency domain and even in challenging listening environments which may consist of multiple competing speakers, speech streams remain intelligible. This is mainly due to the fact that speech energy is concentrated in isolated regions in time and frequency. Binary time-frequency masks exploit this underlying sparsity and disjointness of speech spectra in their short-time-frequency representations by creating a mask that only preserves the spectro-temporal regions where the target is dominant [2].

In the following section, a widely used T-F masking algorithm called degenerate unmixing estimation technique (DUET) [5, 6] is investigated for source separation in reverberant settings. Since the performance of the DUET is known to degrade significantly in reverberation, a preprocessing stage is implemented for signal dereverberation based on interaural coherence. A novel

cue preservation method is proposed and implemented as a post-processing stage that preserves the level and timing cues in the extracted sources. The proposed algorithm is a robust version of DUET algorithm with cue preservation that is called DUET-CP [97]. The efficiency of the proposed cue preservation stage is evaluated in a challenging scenario, while the preservation of binaural cues is measured using a realistic localization test.

3.5.2 DUET-CP algorithm

Our second algorithm for speech separation based on CASA source separation with spatial cue preservation is presented as DUET-CP algorithm [97]. According to [1], listeners perform segregation of a binaural mixture in a two-stage process of segmentation and grouping. Listeners group sound elements based on whether those originate from the same locations (likely to come from a common source) or spatially distributed locations (likely to come from two different sources) [75]. Listeners may rely on ILD and ITD cues to group segments and separate the sources [75, 95].

DUET algorithm [5, 6] for source separation is modified to work in reverberant settings and preserve the location of sound sources. This algorithm employs the binary time-frequency masks to only preserve the spectro-temporal regions where the target is dominant and it is using the assumption that speech spectra in time-frequency domain is sparse and disjoint. Since the performance of the DUET is known to degrade significantly in reverberation, a pre-processing stage is employed for signal dereverberation based on interaural coherence. A cue preservation method implemented as a post-processing stage is shown to preserve level and timing cues in the extracted sources. Therefore, the DUET-CP algorithm includes three stages of dereverberation, speech separation using time freqeuncy masking and cue preservation.

Pre-processing stage : Interaural coherence (IC)

The dereverberation stage based on interaural coherence (IC) is employed as a pre-processing step for speech dereverberation before separating sources. First, the two reverberant mixture signals recorded from the left and the right channels $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ are transformed to the time-frequency

domain by using the STFT which produces the complex valued spectra $x_{\mathrm{L}}(\tau,\omega)$ and $x_{\mathrm{R}}(\tau,\omega)$ with τ as time frame and ω as frequency band. From the time-frequency representation of the left channel mixture and the right channel mixture, the IC is estimated using the normalized cross-correlation function calculated as (e.g., see [98, 99])

$$\mathrm{IC}_{\mathrm{LR}}(\tau,\omega) = \frac{|\Phi_{\mathrm{LR}}(\tau,\omega)|}{\sqrt{\Phi_{\mathrm{LL}}(\tau,\omega)\Phi_{\mathrm{RR}}(\tau,\omega)}}, \tag{3.9}$$

where the $\Phi_{\mathrm{LL}}(\tau,\omega), \Phi_{RR}(\tau,\omega)$ and $\Phi_{\mathrm{LR}}(\tau,\omega)$ are the exponentially weighted short-term auto-correlation and cross-correlation functions defined as

$$\begin{aligned}
\Phi_{\mathrm{LL}}(\tau,\omega) &= \alpha\, \Phi_{\mathrm{LL}}(\tau-1,\omega) + (1-\alpha)\, |x_{\mathrm{L}}(\tau,\omega)|^2 \\
\Phi_{\mathrm{RR}}(\tau,\omega) &= \alpha\, \Phi_{\mathrm{RR}}(\tau-1,\omega) + (1-\alpha)\, |x_{\mathrm{R}}(\tau,\omega)|^2 \quad , \\
\Phi_{\mathrm{LR}}(\tau,\omega) &= \alpha\, \Phi_{\mathrm{LR}}(\tau-1,\omega) + (1-\alpha)\, x_{\mathrm{R}}(\tau,\omega) x_{\mathrm{L}}^{*}(\tau,\omega)
\end{aligned} \tag{3.10}$$

where $0 \leqslant \alpha \leqslant 1$ denotes a smoothing factor. The IC describes the coherence of the left channel signal and the right channel signal which has a range of [0,1], where 1 indicates that both signals x_{L} and x_{R} are perfectly coherent. Consequently, a binary mask $\mathrm{M}_{\mathrm{IC}}(\tau,\omega)$ is derived from the estimated interaural coherence as

$$\mathrm{M}_{\mathrm{IC}}(\tau,\omega) = \begin{cases} 1, & \text{if} \quad \mathrm{IC}_{\mathrm{LR}}(\tau,\omega) > \text{threshold}(\omega) \\ 0, & \text{otherwise} \end{cases}, \tag{3.11}$$

where the threshold value is determined adaptively for each frequency band as $\text{threshold}(\omega) = \max\{0.8, Q_3(IC_{LR})\}$, where Q_3 is the 3^{rd} quartile for each frequency band. Fig. 3.3 illustrates the construction of the binary mask based on the histogram of the interaural coherence. The derived binary mask detects the bins where reverberant energy is dominant and retains the bins with IC close to 1. The estimated binary mask is applied to both channels of $x_L(\tau,\omega)$ and $x_R(\tau,\omega)$ producing the processed (dereverberated) mixtures. The dereverberated signals $\hat{x}_L(n)$ and $\hat{x}_R(n)$ are reconstructed using the inverse STFT and the dereverberant mixtures are then processed with the

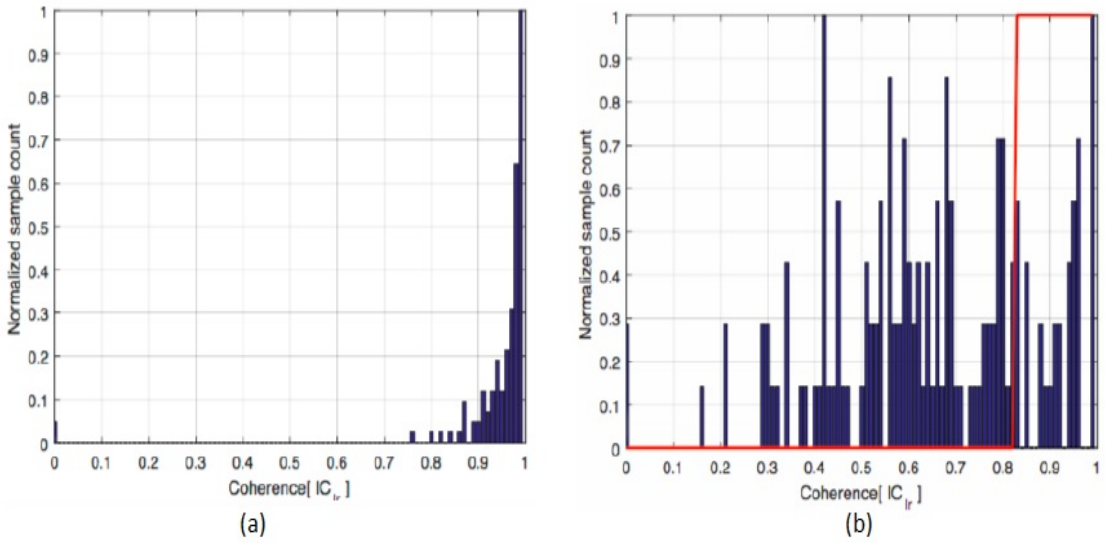


Figure 3.3: The IC histogram of the 700 Hz frequency channel for an anechoic (a) and a reverberant signal (b) and the binary mask (red line).

source separation algorithm described in the following section.

DUET source separation with cue preservation stage

In this section, the DUET algorithm is analyzed for N sources in a two-microphone system configuration [5, 6]. The windowed Fourier transforms of the left $\hat{x}_L(\tau,\omega)$ and right $\hat{x}_R(\tau,\omega)$ dereverberated mixtures can be written as

$$\begin{bmatrix} \hat{x}_L(\tau,\omega) \\ \hat{x}_R(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-j\omega\delta_1} & \dots & a_N e^{-j\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ \vdots \\ \hat{s}_N(\tau,\omega) \end{bmatrix}, \tag{3.12}$$

where N is the number of sources and a and δ represent the mixing parameters for each source for every different time-frequency point (τ,ω). The DUET separates the sources by clustering different time-frequency points based on their interaural parameters. The method essentially constructs a histogram of interaural parameters, with points weighted by their respective energy, and then

selects each prominent peak in the histogram as the interaural parameters of each source. It then creates a mask for each source that retains only time-frequency points with interaural parameters near the selected peak [5, 6]. The mixing parameters for each time-frequency point are calculated based on the mixtures as follows

$$\tilde{a}(\tau,\omega) = \left| \frac{\hat{x}_R(\tau,\omega)}{\hat{x}_L(\tau,\omega)} \right| \quad , \quad \tilde{\delta}(\tau,\omega) = -\frac{1}{\omega} \angle \left(\frac{\hat{x}_R(\tau,\omega)}{\hat{x}_L(\tau,\omega)} \right). \tag{3.13}$$

Instead of using the attenuation parameter $\tilde{a}(\tau,\omega)$, the symmetric attenuation $\tilde{\alpha}(\tau,\omega)$ is often utilized as shown below

$$\tilde{\alpha}(\tau,\omega) = \left| \frac{\hat{x}_R(\tau,\omega)}{\hat{x}_L(\tau,\omega)} \right| - \left| \frac{\hat{x}_L(\tau,\omega)}{\hat{x}_R(\tau,\omega)} \right|. \tag{3.14}$$

After obtaining $\tilde{\alpha}(\tau,\omega)$ and $\tilde{\delta}(\tau,\omega)$, a two-dimensional smoothed weighted histogram is calculated. The peak centers of the histogram represent the mixing parameters. Next, the time-frequency binary masks are reconstructed based on the mixing parameters and are applied to the mixtures to calculate the original source estimates [5, 6].

The DUET algorithm is capable of recovering the original signals blindly from two mixtures. However, this algorithm does not preserve the necessary spatial cues of the original source signals. In order to overcome this problem, a binaural cue preservation strategy that can be realized as an additional post-processing step added to the original DUET is described below. The peaks are located based on the DUET smoothed histogram and the peak centers determine the mixing parameters of $(\tilde{\alpha}_k, \tilde{\delta}_k)$ of the k^{th} source [5, 6]. The attenuation parameter is calculated based on the following equation [5]

$$\tilde{a}_k = \frac{\tilde{\alpha}_k + \sqrt{\tilde{\alpha}_k^2 + 4}}{2}. \tag{3.15}$$

Assuming that only one source is active so that the sources are disjoint, after obtaining both $\tilde{a}_k$ and $\tilde{\delta}_k$ for each separated source, we can recover the spatial cues corresponding to that source in the

time-frequency domain according to the (3.16), which defines the DUET with cue preservation as:

$$\begin{bmatrix} \tilde{s}_{k_{\mathrm{L}}}(\tau,\omega) \\ \tilde{s}_{k_{\mathrm{R}}}(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ \tilde{a}_k e^{-j\omega\tilde{\delta}_k} \end{bmatrix} \tilde{\hat{s}}_k(\tau,\omega)\,, \tag{3.16}$$

where $\tilde{\hat{s}}_k(\tau,\omega)$ is the k^{th} separated signal defined as the output of the DUET and $\tilde{s}_{k_{\mathrm{L}}}(\tau,\omega)$ and $\tilde{s}_{k_{\mathrm{R}}}(\tau,\omega)$ are the time-frequency representations of the k^{th} cue preserved signals corresponding to the left and right channel, respectively. By comparing (3.12) with (3.16), it becomes apparent that the mixing parameters $\tilde{a}_k$ and $\tilde{\delta}_k$ are equivalent to the interaural level differences and the interaural time differences of the spatially separated original sources.

3.5.3 Experimental results on DUET-CP

The proposed algorithm was evaluated on a test set of speech signals comprised of a single randomly selected male spoken sentence. A female interferer was used as the masker with a root-mean-square value equal to the target source, such that the input SNR was equal to 0 dB. The duration of each speech signal was approximately 3 s. All signals were recorded at a sampling rate of 22,050 Hz. To generate the speech test stimuli, sentences from the IEEE database were selected, which consists of phonetically balanced sentences, with each sentence being composed of approximately 7 to 12 words [92]. All signals had the same onset and were normalized to their maximum amplitude before convolving with the HRTFs. Office head-related impulse responses measured in the University of Oldenburg were used to simulate a reverberant listening condition with RT_{60} = 0.3 s [78].

For each listening scenario, a total of seven different azimuthal sound source locations were calculated for sound sources located 1 m away from the center of the listener in the azimuthal plane for every angle from -90^0 left of the listener to +90^0 to the right of the listener in 30^0 increments. In all cases, the interference source was placed directly in front of the listener at 0^0 and the target was allowed to virtually rotate around the listener in the presence of competing speech.

The performance of the proposed algorithm encompassing the IC and CP stages (IC-DUET-CP) was systematically evaluated according to three different performance outcomes: (1) degree of separation, (2) degree of dereverberation, and (3) preservation of binaural cues. To measure the degree of separation, the popular metrics of the signal-to-distortion-ratio (SDR) and signal-to-interference-ratio (SIR) [100] were used. To compute the amount of speech dereverberation achieved, the speech to reverberation modulation energy ratio (SRMR) [101] and the segmental signal-to-reverberation ratio (segSRR) [102, 103, 104] were utilized. To assess the effectiveness of the algorithm in retaining the necessary spatial cues of the output signal estimates, the correlation coefficient between the theoretical and estimated interaural time differences was calculated. Additionally, the localization listening tests were performed with normal-hearing listeners and measured the subjective sound identification responses using the root-mean-square localization error (RMSLE).

3.5.3.1 Reverberation suppression with SRMR and SegSRR

In order to evaluate the performance of the reverberation suppression stage, two metrics, the speech to reverberation modulation energy ratio (SRMR) [101] and the segmental signal-to-reverberation ratio (segSRR) [102, 103, 104] were used. For the SRMR evaluation, the processed signal was passed through a 23-channel gammatone filterbank and the temporal envelope of each filter output was calculated using the Hilbert transform. Second, the extracted envelopes were multiplied by a 256-ms Hamming window and then for each critical band, the modulation spectral energy was calculated. Next, the modulation frequency bins were grouped into eight bands. Finally, the ratio of the average modulation energy for the first four bands over the average modulation energy of the last four bands was calculated as SRMR [101]. For each channel the SRMR measurements were calculated separately and averaged over all mixtures. The ΔSRMR shows the effect of processing and is expressed as

$$\Delta\text{SRMR} = \text{SRMR}_{\text{processed}} - \text{SRMR}_{\text{reverberant}} . \tag{3.17}$$

The segSRR estimates the energy of the direct signal compared to the reverberant energy which is equivalent to the signal-to-noise ratio (SNR) when reverberation is considered as noise [103, 104]. Therefore, the segSRR for each frame m is calculated as

$$\text{segSRR}(m) = 10\log_{10}\left[\frac{\sum_{n=mR}^{mR+N-1} s_d^2(n)}{\sum_{n=mR}^{mR+N-1}(s_d(n)-\bar{s}(n))^2}\right], \tag{3.18}$$

where $s_d(n)$ is the direct signal, $\bar{s}(n)$ is the reverberant or the processed signal. The R and N values are the frame rate in samples and the total number of signal samples respectively. Finally, the segSRR was calculated as the average of segSRR(m) over all non-silence frames and the improvement of SRR is calculated as

$$\Delta\text{segSRR} = \text{segSRR}_{\text{processed}} - \text{segSRR}_{\text{reverberant}}, \tag{3.19}$$

where the $\text{SRR}_{\text{processed}}$ metric was calculated after the two signals were processed through the dereverberation stage and $\text{SRR}_{\text{reverberant}}$ was due to the unprocessed signals [103, 104].

3.5.3.2 Separation performance with SDR and SIR

In order to measure the separation performance, the SDR and SIR criteria were used [100]. The SDR calculates the ratio of the energy in the original signal to the summation of the energy of interference, artifacts and distortion, while SIR calculates the ratio of the target energy to the interferer energy, defined as [100]

$$\text{SDR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}+e_{\text{noise}}+e_{\text{artif}}||^2} \tag{3.20}$$

$$\text{SIR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2}. \tag{3.21}$$

3.5.3.3 Localization error (RMSLE)

To evaluate the cue preservation efficiency of the proposed algorithm, the theoretical and estimated ITDs were compared. The theoretical ITD was derived based on the peak location of the cross-correlation between left and right HRTFs. The ITD was taken to be the lag at which the largest peak occurred in the cross-correlation. The ITD of the cue preserved outputs, was calculated based on the peak location of the cross-correlation between the left and right channel of the output signals. The similarity between the theoretical and experimental ITD, was assessed with the Pearson's correlation coefficient. The proposed algorithm illustrates a high correlation coefficient of $\rho = 0.97$ between the theoretical and estimated ITDs.

To measure localization accuracy, six undergraduate students from the University of Kansas with American English as their first language were recruited for course credit. All listeners gave informed consent prior to testing. All listeners recruited had normal hearing and reported normal cognitive function. The listeners were presented with IEEE sentences processed in 3 different experimental conditions: (1) clean signal (target only), (2) DUET and (3) IC-DUET-CP through headphones. Each target sentence was presented in random order from any of the seven different locations for a total of two presentations from each location. The listeners were instructed to identify the perceived virtual location of the male spoken sentence as accurately as possible by verbally indicating the number 1 to 7. A diagram showing the location of the subjects' head and the seven possible virtual locations of the male spoken sentence relative to their position was mounted on the wall directly in front of the subjects. The locations were numbered from 1–7 with -90^{o} corresponding to location 1 going clockwise in 30^{o} steps to $+90^{\text{o}}$ represented by location 7. Each subject produced a total of 14 responses (7 locations x 2 repetitions) per condition. To quantify the ability of the algorithm in retaining spatial cues in the signal estimates, the RMSLE between the azimuth of the presented stimulus location and the actual responses was calculated [105].

3.5.4 Discussion

As shown in Table 3.2, the DUET and DUET-CP algorithms yield high values of SDR and SIR in the anechoic condition (ANE) tested. However, in reverberation (REV), there is a significant decrease in both criteria. Therefore, employing the dereverberation stage is essential before source separation. According to Table 3.2, when compared to the unprocessed reverberant mixtures, the DUET-CP increases SDR by 2.46 dB and SIR by 6.28 dB. Moreover, the IC-DUET-CP produces an SDR equal to 2.40, which is 2.96 dB greater than the SDR of the unprocessed mixtures and 0.5 dB greater than the SDR calculated after DUET was applied. The proposed algorithm retains the quality and adds no distortion to the output signals. Furthermore, the IC-DUET-CP algorithm produced higher SIR scores than the scores corresponding to the unprocessed mixtures and the DUET outputs. The cue preservation stage also maintains the SDR and SIR values, while preserving the location of the sources. The proposed algorithm retains the quality and adds no distortion to the output signals. Furthermore, the IC-DUET-CP algorithm produced higher SIR scores than the scores corresponding to the unprocessed mixtures and the DUET outputs. The cue preservation stage also maintains the SDR and SIR values, while preserving the location of the sources. The pre-processing dereverberation stage increases the SRMR and segSRR by 1.36 dB and 2.30 dB, respectively.

Fig. 3.4 shows bubble plots for the six listeners who participated in the localization task. The most striking feature of the localization responses is the relatively high rate of azimuth confusions observed when the target source estimates were generated without the cue preservation post-processing stage. In this case, the observed RMS localization error averaged across all listeners was 63^{o}. In contrast, when the source estimates were extracted using cue preservation, most responses fell on the diagonal indicating correct sound source identification by the listeners. In this case, the observed RMS localization error was 32^{o}, while the error when the target speech signal was presented alone was equal to 23^{o}. In conclusion, a blind source separation algorithm based on T-F masking is proposed that recovers the original sources in reverberation and preserves the spatial location of the sources. Preserving binaural cues increases the applicability of T-F binary

masking strategies as it not only ensures improved speech intelligibility in reverberation but also enhances localization performance.

Table 3.2: SDR and SIR values averaged over 70 mixtures.

	SDR (dB)		SIR (dB)	
	ANE	REV	ANE	REV
Mixtures	3.07	-0.56	3.07	1.85
IC	–	-0.81	–	1.71
DUET	10.66	1.90	16.70	8.13
DUET-CP	10.66	1.90	16.70	8.13
IC-DUET-CP	–	2.40	–	8.90

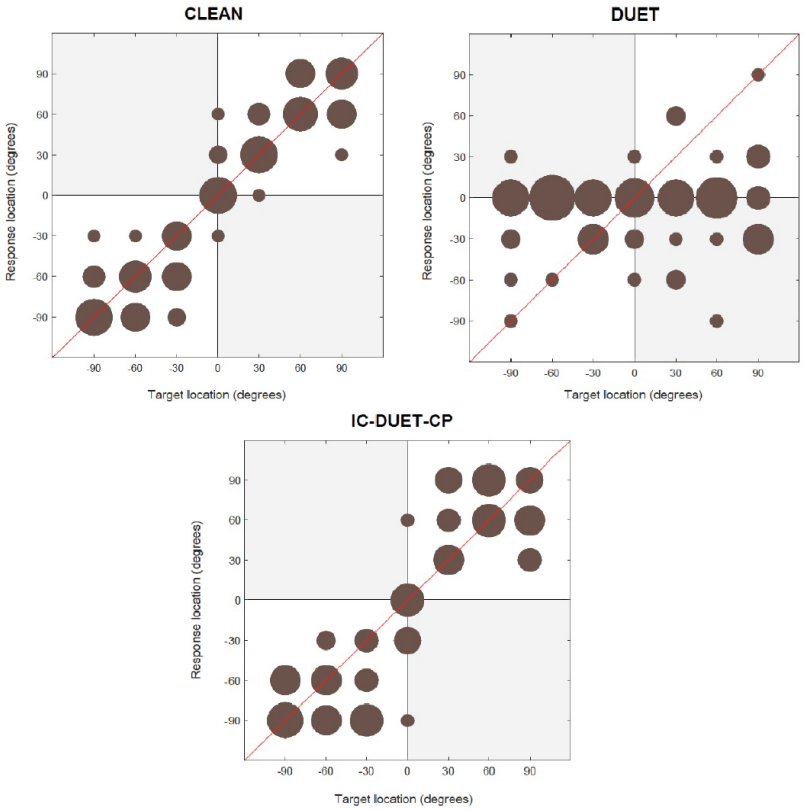


Figure 3.4: Localization patterns for six listeners tested with the clean signal, DUET and IC-DUET-CP. The response location is plotted as a function of the target location. The area of each circle is proportional to the number of responses.

Chapter 4

Supervised speech separation using deep learning techniques

This chapter explains the supervised speech separation methods using deep learning in detail. First, three important sections of supervised speech separation algorithm (learning machine, acoustic features and training target) are discussed . Then, different monaural acoustic features are explained for speech separation and examined in various room conditions, background noise and multiple speakers. Moreover, a novel monaural feature set is presented that leads to significant improvement in speech separation performance in terms of objective speech intelligibility and speech quality.

4.1 Introduction

In real world environments, speech signals that reach listeners' ears are usually corrupted by other, non-target sound sources and their surface reflections, or reverberations. The presence of such acoustic interference and reverberation has a negative effect on speech intelligibility for normal hearing and hearing-impaired listeners as well as speech processing applications such as automatic speech recognition (ASR) and speaker identification systems. Speech separation is the task of separating the target speech from its background interference such as nonspeech noise, speech noise and room reverberation. The separation of target speech from nonspeech noise is usually called as speech enhancement or speech denoising, while separation of target speech from other speech or multiple voices is called speaker separation [70]. Speech separation has many applications in hearing aid devices, mobile communication, speech recognition and speaker recognition. Depending on the number of recording microphones, two categories of speech separation are defined as monaural

speech separation (single microphone) and array based speech separation (multi-microphone).

In monaural case, the speech enhancement methods exploit the statistics of speech and noise to estimate the clean speech from the noisy mixture. In addition, CASA based methods inspired by perceptual principles of auditory scene analysis use grouping cues for monaural speech separation [27]. On the other hand, multi-microphone based algorithms such as beamforming and spatial filtering separate the desired signal by boosting the signal coming from a specific direction and attenuating signals coming from other directions. The performance of the spatial filtering approaches depends on the configuration of the array, number of microphones and array length, and co-located sources can not be separated. Furthermore, the separation ability of beamformers in reverberant condition is limited [70].

Recently, separating the target speech from background interference, which is the goal of speech separation algorithms, has been formulated as a supervised learning problem. In supervised learning problems like speech separation, different discriminative patterns of speech, speakers and background noise can be learned from the training data. Over the past decade, deep learning methods have been introduced to the field of speech separation and resulted in significant improvement in separation performance [70].

Supervised speech separation was originally inspired by time frequency masking in CASA where a weighting mask is applied to the time frequency representation of noisy mixture to separate the target source (speech) from the interference [27]. The use of ideal binary mask (IBM) [2] for speech separation makes the separation problem a binary classification task. With the representation of speech separation as a classification problem, the data driven methods have become more popular in solving challenges in speech processing field [70]. Over the last decade, supervised speech separation performance has increased significantly by employing larger training dataset, and exploiting greater computing resources. In the following sections, the three important components of supervised speech separation algorithms (learning machines, training target and acoustic features) will be discussed in details. Moreover, extensive study of acoustic features will be presented and a new set of monaural acoustic feature is proposed to improve the performance

of supervised speech separation method.

4.2 Learning Machines

Over the last decade, deep neural networks (DNNs) have been extensively used in many supervised learning problems such as pattern recognition, image classification, speech recognition, machine translation, natural language processing, and speech separation, and they have shown significant improvements compared to classical machine learning models. In the following section, different types of neural networks as learning machines will be discussed.

4.2.1 Neural network types

The simplest and the most popular form of neural network is the feedforward neural network, which consists of fully connected layers with feedforward connections between the consecutive layers [106]. In this type of network, there are connections between the nodes of the network in adjacent layers where each connection has its own associated weight. The feedforward network has three different types of layers such as input layer, hidden layer and output layer. The input layer passes the information to hidden layers and there is no computation on this layer, while the hidden layers transform the information from the input layer to the output layer. The feedforward network only has one input layer and one output layer but it can have multiple hidden layers. It should be noted that the information in feedforward network moves only in the forward direction and there is no cycle or loop in this network [106]. Multi layer perceptron (MLP) is a feedforward network that has one or more hidden layers. During training, the weights of the network are adjusted with backpropagation algorithm [107] to minimize the error between the predicted output and the desired output through gradient descent algorithm.

Convolutional neural networks [108] are another class of feedforward networks that had significant success for pattern recognition in computer vision and image processing field. CNN usually includes the cascade of convolutional layers and subsampling layers. The convolutional layers

consist of feature maps that extract the spatial and temporal features of its input. The subsampling layers reduce the resolution and sensitivity of the local variations by averaging or maximizing over a kernel of nodes. The subsampling (pooling) layers reduce the dimension of the representation and make the network invariant to shifts and distortions of its input. During training, the weights (filterbanks) are adjusted by the the backpropagation algorithm.

Another class of neural networks are recurrent neural networks (RNNs) that have feedback connections between hidden units. RNNs can model the changes of the data over time; therefore, it can model sequential data. For signals that have temporal structure where the current frame of the signal depends on the previous frames, RNN is a desirable model [109]. The backpropagation through time [110] is typically used to train the recurrent connections. Speech signals have temporal structure; therefore, recurrent network may be a suitable learning machine to model the dependency of speech frames. RNNs are designed to solve sequence prediction problems and they use the information of previous frames (samples) to understand the current sample. They can also model sequential and time dependencies simultaneously. Recurrent models are also suitable for language modeling, machine translation, text and audio processing and generative models. However, vanilla RNN has a major problem of vanishing gradient [111], which means that the network only memorize the recent samples and make prediction based on these samples and leads to inaccurate predictions.

Another kind of recurrent neural network is Long Short Term memory (LSTM) [112], which works better than vanilla RNN, does not have the vanishing gradient issue, and is capable of modeling long-term dependencies. LSTM model has memory cells such as input gate, output gate, forget gate and internal state, and it uses these memory cells to mitigate short term memory. These gates are important to keep or throw away the previous information. Gated recurrent unit (GRU) [113] is another type of RNN which is very similar to LSTM but it has two gates of reset and update. The update gate is like the forget and input gate of LSTM that decides how much information to keep or throw away, while the reset gate determines how much of past information to forget.

4.2.2 Parameters of neural network

In this section, the learning procedure of the neural network are discussed including backpropagation algorithm using gradient descent, the cost function in deep neural network and common activation functions are discussed. Furthermore, an efficient technique of batch normalization will be briefly mentioned that improves the generalization of neural network models to unseen conditions.

4.2.2.1 Training algorithm

Each of the deep learning models has a cost function which defines how well the model performs given different values for each of its parameters. During training, it is important to select the parameters that result in the lowest error between the predicted output and the desired output. A neural network propagates the input signal forward through its layers and then backpropagates the error in reverse through the network to adjust its parameters via backpropagation[107].

The backpropagation algorithm has two stages, a forward pass and backward pass. In forward pass, the data is fed to the the network and the network predicts an output based on its initial parameters. Then a loss function (error) is measured based on the predicted output and desired output. In the backward stage, the error is backpropagated through the network to adjust tunable parameters [107]. One of the most popular optimization algorithms for optimizing neural networks is gradient descent. Gradient descent is an iterative algorithm that is used to find the minimum value of the cost function (loss function) by finding its derivative and adjust the parameters of the network. The backpropagation algorithm needs the gradient of the cost function at each iteration and uses the partial derivative of the cost function to the weights and bias of the neural network to update these parameters. It should be noted that the cost function needs to be continuous and differentiable. For the feedforward neural network with sigmoid activation, the output of each layer is calculated as follows

$$a_n^{\mathrm{m}} = \sigma\left(\sum_k w_{nk}^{\mathrm{m}} a_k^{\mathrm{m}-1} + b_n^{\mathrm{m}}\right), \tag{4.1}$$

where $w^{\mathrm{m}}_{\mathrm{nk}}$ and $b^{\mathrm{m}}_{\mathrm{n}}$ are the weights and bias of the layer and $a^{\mathrm{m}}_{\mathrm{n}}$ is the activation of the n^{th} neuron of the m^{th} layer that is summed over all neurons k. The activation function on this sample layer is sigmoid, which makes the output value restricted between 0 and 1, and adds a non-linearity between the input and the output of the layer. Furthermore, the common cost function for a regression problem (predict a value) is the mean squared error (MSE) on output layer M that is defined as

$$C = \frac{1}{2}\sum_{n}(y_n - a_n^{\mathrm{M}})^2 \,. \tag{4.2}$$

Using the gradient descent algorithm, the weights of the neural network are updated as

$$w_{t+1} = w_t - \eta \nabla C(w) \,, \tag{4.3}$$

where η is the learning rate and $\nabla C(w)$ is the gradient of the cost function with respect to the weights of the neural network. It should be noted that the parameters of the network are updated in the opposite direction of the gradient of cost function with respect to the model's parameters.

One of the most important parts of the gradient descent algorithm is selecting the proper learning rate, which determines the speed of the algorithm's convergence. Selecting the learning rate is a challenging task. If the learning rate is too small, the algorithm converges very slowly which means higher training time, while a large learning rate results in fluctuation of the cost function and even may lead to its divergence. Therefore, different types of optimizers for gradient descent were proposed such as Adagrad [114], RMSProp [115], Adadelta [116] and Adam [117]. Adaptive moment estimation (Adam) was proposed as a first-order gradient-based optimizer for stochastic objective functions that uses low order moments. Furthermore, it is a very efficient optimizer and requires little memory. It is also invariant to diagonal rescaling of the gradients which is suitable for problems with large data and parameters [117]. Adam uses the estimates of the first and second moments of gradients to compute individual adaptive learning rates.

4.2.2.2 Activation functions

Another important part of neural networks is the type of activation function applied on each layer. Through many years, different types of activation functions were proposed to introduce non-linearity to neural network structure. The most common activation functions are mentioned below:

1. **Sigmoid**: Sigmoid function makes the output restricted to the range of [0-1] and it is mostly used in predicting the probability as an output. In speech processing, the ideal ratio mask can be represented by sigmoid activation function which is given as

$$f(x) = \frac{1}{1+e^{-x}}\,. \tag{4.4}$$

2. **Tanh**: Hyperbolic tangent is like a sigmoid function (S-shaped) with the range of [-1,1] which is differentable and monotonic as

$$f(x) = \frac{2}{1+e^{-2x}} - 1\,. \tag{4.5}$$

3. **ReLU** : The Rectified linear unit (ReLU) [118] function rect(x) = max(0; x) is one-sided and the response for negative values of the input is zero; therefore, it does not enforce sign symmetry or anti-symmetry and it allows the network to obtain sparse representations. The model with ReLU activation function does not have the problem of vanishing gradient due to activation non linearities (like sigmoid and tanh) because the derivative of ReLU with respect to the parameters of the model is always one. The ReLU computation is cheaper compared to tanh and sigmoid because there is no need to compute exponential function in activation. ReLU is defined as

$$f(x) = \begin{cases} x, & \text{if } \; x \geqslant 0 \\ 0, & \text{if } \; x < 0 \end{cases}\,. \tag{4.6}$$

4. **ELU**: Exponential linear unit (ELU) [119] is another type of activation function which improves learning compared to other activation functions. ELU has negative values which pushes the mean unit activation closer to zero. Shifting mean toward zero speeds up learning by making the normal gradient closer to the unit natural gradient. With smaller inputs, ELU saturates to a negative value and this saturation decreases the forward propagated variation and the information [119]. Therefore, the ELU is both noise-robust and low complex compared to other activation functions, which is defined as

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{if } x \leqslant 0 \end{cases} . \tag{4.7}$$

5. **Other rectified linear unit activation**: There are recently three new proposed rectified linear unit activation function such as Leaky rectified linear unit [120], scaled exponential linear units (SELU) [121] and parametric rectified linear unit [122]. Leaky rectified linear unit is defined as

$$f(x) = \begin{cases} x, & \text{if } x \geqslant 0 \\ \frac{x}{a}, & \text{if } x < 0 \end{cases} , \tag{4.8}$$

where a is a fixed parameter in the range of $(1, +\infty)$. Parametric Rectified Linear Unit is similar to leaky ReLU but the parameter of a is learned via back propagation during training. The scaled exponential linear units (SELU) was proposed to normalize the output of activation function as

$$f(x) = \lambda \begin{cases} x, & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{if } x \leqslant 0 \end{cases} . \tag{4.9}$$

SELU is very similar to scaled leaky RELU but the parameters of a and λ are not trained through backpropagation and they are not hyper parameters. It is parameterized by a and λ, which control the mean and variance of the output distribution. If we want the distribution

with 0 mean and unit variance, a needs to be around 1.67326 and λ needs to be around 1.0507 [121].

4.2.2.3 Batch normalization

Batch normalization [123] was proposed to reduce the internal covariate shift, which leads to faster training. The internal covariate shift in training is defined as the change in the distribution of network activations due to the changes of the network parameters. Batch normalization fixes the mean and variances of layer inputs. If the inputs of the network are whitened (linearly transformed to have zero mean and unit variance and decorrelated), network converges faster in training [123]. The fixed distributions of inputs are achieved when the inputs to each layer are whitened. Moreover, it reduces the dependency of the gradients to the scale of the parameters or their initial values [123]. Therefore, higher learning rates can be used for training the network without divergence risk.

4.3 Training Target

In supervised speech separation, two main categories of training targets were defined as masking based and mapping based training targets. In mapping based methods, the spectral representation of clean speech is the training target, while the time-frequency mask (time-frequency relationships of clean speech to interference) is the training target in masking based approach. Common masking based training targets are ideal binary mask, ideal ratio mask, target binary mask, spectral magnitude mask and complex ideal ratio mask, which are described briefly in the following section. Moreover, different training targets for a noisy speech signal were illustrated on Fig. 4.1.

4.3.1 Masking based training target

4.3.1.1 Ideal binary mask (IBM)

The target and interferer energy in each time frequency unit are required for ideal binary mask (IBM) construction and IBM [2] is defined as follows

$$\mathrm{IBM}(\tau,\omega) = \begin{cases} 1, & \text{if} \quad \mathrm{SNR}(\tau,\omega) > \mathrm{LC} \\ 0, & \text{otherwise} \end{cases}, \tag{4.10}$$

where τ and ω are the time frame and the frequency index, respectively. The $\mathrm{SNR}(\tau,\omega)$ is the SNR for each time frequency (TF) unit in dB, while the local SNR (LC) is the threshold for classifying the TF units as target or interferer.

4.3.1.2 Target Binary Mask (TBM)

The target energy in each TF unit is compared with a fixed interference like speech-shaped noise (SSN) to construct the target binary mask (TBM) [8]. This mask compares the target speech with the long-term average spectrum of the target speech as follows

$$\mathrm{TBM}(\tau,\omega) = \begin{cases} 1, & \text{if} \quad \frac{\mathrm{T}(\tau,\omega)}{\mathrm{r}(\omega)} > \mathrm{LC} \\ 0, & \text{otherwise} \end{cases}, \tag{4.11}$$

where $\mathrm{r}(\omega)$ is the long-term average of the energy in each frequency channel which is a function of frequency and $\mathrm{T}(\tau,\omega)$ is the target sound.

4.3.1.3 Ideal Ratio Mask (IRM)

Ideal ratio mask [3] is the smooth version of IBM with the output value in the range of [0-1] and related to Wiener filtering when the speech and noise are uncorrelated. The IRM is defined as

$$\mathrm{IRM}(\tau,\omega) = \sqrt{\frac{S^2(\tau,\omega)}{S^2(\tau,\omega)+N^2(\tau,\omega)}}, \tag{4.12}$$

where $S^2(\tau,\omega)$ and $N^2(\tau,\omega)$ are the speech and noise energy for each TF unit. The IRM is the square-root of Wiener filter [3].

4.3.1.4 Spectral Magnitude Mask (SMM)

Spectral magnitude mask (SMM) or FFT-MASK is defined based on the magnitude of the short-time Fourier transform (STFT) of clean speech and noisy speech as

$$\mathrm{SMM}(\tau,\omega) = \frac{|S(\tau,\omega)|}{|Y(\tau,\omega)|}, \tag{4.13}$$

where $|S(\tau,\omega)|$ and $|Y(\tau,\omega)|$ denote the spectral magnitudes of the clean speech and noisy speech [3].

4.3.1.5 Phase-Sensitive mask (PSM)

Most of the speech separation algorithms enhance the magnitude response of the noisy speech without modifying the phase response. However, researchers found out that by enhancing the phase spectrum of a noisy speech, there are some improvements in the quality of speech [124]. Furthermore, in negative SNRs, the phase of the noisy speech includes more of the phase of the background noise than the speech; therefore, using the phase of the noisy speech with predicted magnitude spectrum in speech synthesis step causes more problems.

According to the research in phase improvements, including the phase information in generating the TF masks would be beneficial. Including the phase difference in PSM leads to better estimation of clean speech than amplitude-based masks [48] . A masking based training target that includes the information about the phase of the input in mask generation was proposed as phase sensitive mask as

$$\mathrm{PSM}(\tau,\omega) = \frac{|S(\tau,\omega)|}{|Y(\tau,\omega)|}\cos\theta, \tag{4.14}$$

where θ is the phase difference between the clean speech and noisy speech in each TF unit.

4.3.1.6 Complex Ideal Ratio Mask (cIRM)

The complex ideal ratio mask (cIRM) was proposed to both enhance the magnitude and phase responses of the noisy speech [125]. The cIRM is like the IRM derived in the complex domain. The goal of the cIRM is to recover the STFT of the clean speech $S(\tau,\omega)$ when applied to the STFT of the noisy speech $Y(\tau,\omega)$. The procedure is explained as follows

$$S(\tau,\omega) = M(\tau,\omega) \times Y(\tau,\omega)\,, \tag{4.15}$$

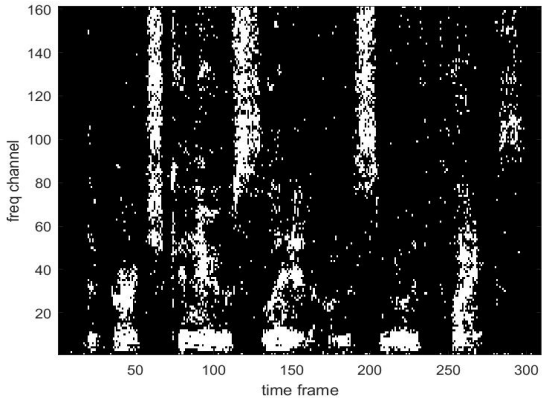
where $\times$ is the complex multiplication. Writing the complex values of $S(\tau,\omega)$ and $Y(\tau,\omega)$ to the real and imaginary part, the M is defined as :

$$\mathrm{M} = \frac{Y_\mathrm{r}S_\mathrm{r} + Y_\mathrm{i}S_\mathrm{i}}{Y_\mathrm{r}^2 + Y_\mathrm{i}^2} + \quad j \ \frac{Y_\mathrm{r}S_\mathrm{i} - Y_\mathrm{i}S_\mathrm{r}}{Y_\mathrm{r}^2 + Y_\mathrm{i}^2} \tag{4.16}$$

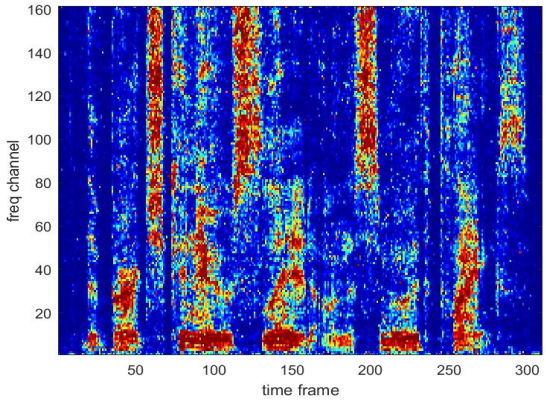
where Y_r and Y_i are the real and imaginary components of noisy speech while S_r and S_i are the real and imaginary components of clean speech. Unlike IRM which is bounded to [0,1], the cIRM is unbounded; therefore, some compression methods such as sigmoid function and tangent hyperbolic should be used to bound the mask values to decrease the complication in mask estimation [125].

4.3.2 Mapping based training target

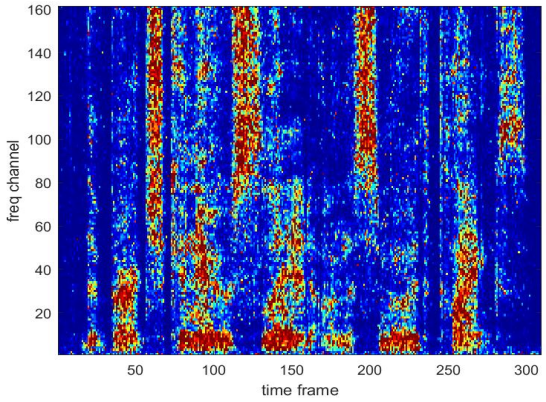
Target magnitude spectrum (TMS) is a mapping based training target, which is the magnitude of STFT of clean speech. In mapping based method, the neural network predicts the magnitude spectrogram of the clean speech from the noisy speech [126, 127]. Power spectrum and mel spectrum can also be used as the training target in mapping based method. Mean squared error is usually used in TMS estimation. Another type of mapping based training target is the gammatone frequency target power spectrum (GF-TMS) where the target is defined based on the gammatone filterbank. The target is estimated based on the power of cochleagram response to clean speech [3].



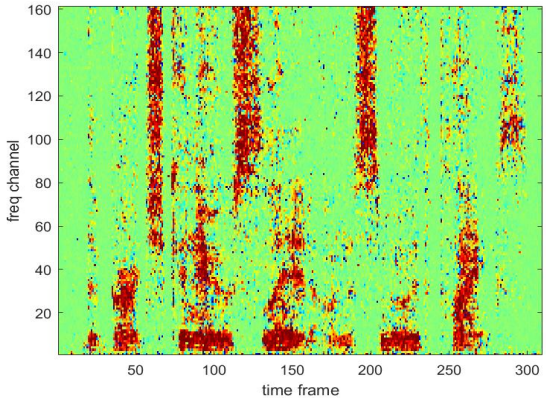
(a) IBM



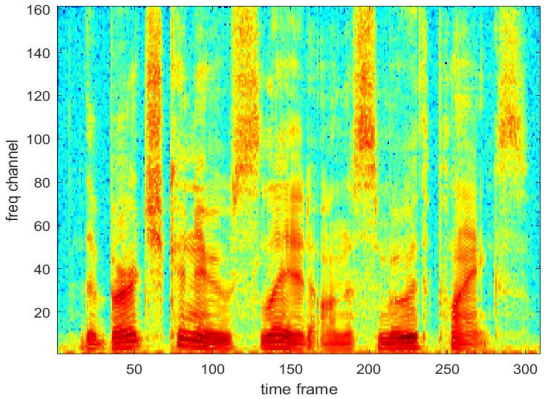
(b) IRM



(c) SMM



(d) PSM



(e) TMS

Figure 4.1: Visualization of different training targets for a IEEE clean speech mixed with factory noise at 0 dB SNR.

4.4 Acoustic features

Another important part of the supervised speech separation are the acoustic features that need to be fed to the neural network as an input. Over the past decades, different robust features were proposed for automatic speech recognition system, which could be used as acoustic features for speech separation. The common acoustic features used in ASR systems are mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) [128], relative spectral transform PLP (RASTA-PLP) [129] and amplitude modulation spectrogram (AMS) [63]. The features of MFCC, PLP, RASTA-PLP, AMS are spectral features that were derived from the short time fourier transform (STFT) of the noisy input signal. Recently, another category of acoustic features is presented as gammatone-domain features such as gammatone freqeuncy (GF) [27] , gammatone frequency cepstral coefficients (GFCC) [130] and multi-resolution cochleagram (MRCG) [131] for monaural supervised speech separation. These features are presented based on gammatone filterbank representation of input signals. In the next section, the existing acoustic features that have been successfully used in speech processing and speech recognition system will be discussed in detail and a novel set of acoustic features for monaural speech separation will be presented.

4.5 Monaural acoustic feature study

In case of monaural speech separation, two different categories of acoustic features were presented as spectral features and gammatone feature. The parameters of these features are selected based on the sampling rate of 16 KHz with the maximum frequency of 8 KHz in filterbank design.

4.5.1 Spectral features

- Mel-Frequency Cepstral Coefficients (MFCC) are a commonly used feature in ASR. First, the input signal is divided into frames of 20 ms using Hamming window with 50% overlap (10 ms). Then the FFT with the size of 320-point is applied to the time frames to find the power spectrum. The power spectrum is converted to mel scale and then log compressed.

Finally, the discrete cosine transform is applied to the compressed power to calculate 31 cepstral coefficients.

- The Log Magnitude spectral features (Log-Mag) is calculated by applying the log operation on the magnitude of STFT representation (320-point FFT) of the noisy signal [132].

- The Log-mel Filterbank Feature (Log-mel) computes a spectrogram of the input signal using a 40-channel mel filterbank and then log compressed [132].

- The Relative Spectral Transform Perceptual Linear Prediction feature (Rasta-PLP), first estimates the critical band power spectrum (similar to the PLP procedure), then the logarithm operation is applied to the magnitude spectrum and the RASTA filtering is applied to the log spectrum. Finally, the filtered log spectrum is decompressed and cepstral coefficients are obtained according to the perceptual linear prediction (PLP) feature extraction procedure [129].

- Amplitude Modulation Spectrogram (AMS) feature is computed as follows: the full-wave rectified envelope of the signal is decimated by the factor of 4. Then the frames of the signal with 32 ms and 10 ms overlap using Hamming window are extracted with 256-point FFT. Finally, the 15D feature is calculated by multiplying the FFT magnitudes with 15 triangular-shaped windows uniformly centered in the range of 15.6 to 400 Hz [63].

4.5.2 Gammatone domain features

- The gammatone frequency feature (GF) processes the input signal by a 64-channel gammatone filterbank [27] resulting in an estimate of the energy in each TF unit with 20 ms frame length and 10 ms frame shift. Finally, a cubic root compression is applied to the overall energy of each TF unit.

- The Multiresolution Cochleagram feature (MRCG) consists of 4 gammatone cochleagrams (4 CG) that represents the local and contextual information of the signal [131]. First, two

cochleagrams with the frame length of 20 ms and 200 ms are computed using a 64-channel gammatone filterbank. Next, the logarithm operation is applied to both cochleagrams and forms the CG1 and CG4 sub-features. The second and third cochleagram features (CG2 and CG3) are obtained by smoothing CG1 with a 11×11 window and a 23×23 window, respectively. Finally, all four cochleagrams (CG1, CG2, CG3, CG4) are concatenated to construct the full MRCG feature[131].

- Gammatone Frequency Cepstral Coefficient (GFCC) uses the same procedure as GF calculation. Then the discrete cosine transform (DCT) is applied to the GF features and 31 coefficients are derived as GFCCs.

The Waveform signal was also examined as a feature to directly train the neural network with no STFT processing or gammatone filtering. The time frames of 20 ms with the shift of 10 ms have been extracted directly from the wave signal which corresponds to 320 signal samples with 160 sample shift when the sampling frequency is 16 KHz.

4.6 New monaural gammatone-based acoustic feature

Recent studies have examined the effect of specific acoustic features (e.g., gammatone filters versus spectral decompositions) [131, 68, 132] on supervised learning of T-F masks for speech separation in anechoic and reverberant rooms with non-speech noise. The results of these studies found gammatone-domain acoustic features outperform spectral features in matched noise conditions where the same noise type is used in both training and test stage.

Inspired by these findings, a new gammatone-domain acoustic feature, the dynamic multi-resolution cochleagram (DMRCG), is proposed in this dissertation and evaluated for monaural supervised speech separation in noisy and reverberant environments [133]. DMRCG includes both local cochleagram information and spectrotemporal context and associated frequency dynamics. To evaluate the performance of this proposed feature, the speech separation objective intelligibility is tested with the target speech located in random, simulated room locations with different rever-

beration times and non-stationary background noise. Furthermore, the performance of DMRCG is examined in comparison to other spectral and gammatone-domain features used in current speech processing applications. Finally, the effectiveness of two neural networks for training the ideal ratio mask (IRM) [3] is compared. The results of these analyses show speech separation is improved after incorporating the proposed feature, both in terms of objective intelligibility and computational complexity. In addition, further improvements were found for LSTM neural network estimation of the IRM versus DNN-based estimation [133].

4.6.1 Dynamic multi-resolution cochleagram (DMRCG)

DMRCG is a gammatone based acoustic feature that has the same structure as multi-resolution cochleagram (MRCG) which includes 4 different cochleagrams. The first cochleagram (CG1) of our feature and the MRCG is obtained by analyzing the input audio signal using a 64-channel gammatone filterbank with 20 ms window and 10 ms overlap followed by computing the logarithm of the power for each time frame and frequency channel. The second cochleagram (CG2) is generated by smoothing CG1 with a square window of size 11×11 (100 ms and 500 Hz) time-frequency (TF) units. The third cochleagram (CG3) is also derived from the CG1, and is smoothed with a larger window size of 23×23 TF units to differentiate the spectrotemporal context at two resolutions. The final cochleagram (CG4) of our proposed feature is also based on CG1, but is filtered with a 11×11 TF units 2D Laplacian of Gaussian (LoG) to highlight rapid changes in the energy of each TF unit.

Laplacian filter operations are a common spatial filter in image processing for sharpening and edge detection that highlights regions of rapid change in intensity. By treating the cochleagram as an image, the 2D LoG filter provides more details about fast dynamics of the cochleagram spectrotemporal context, and has an added benefit of requiring less computation than the fourth MRCG cochleagram feature. Finally, all four cochleagrams (CG1,CG2,CG3,CG4; each dimension 64) are concatenated to create the final feature set with the dimension of 256. As shown in Fig. 4.2 for an example speech sample, CG1 is the original cochleagram, CG2 and CG3 capture the

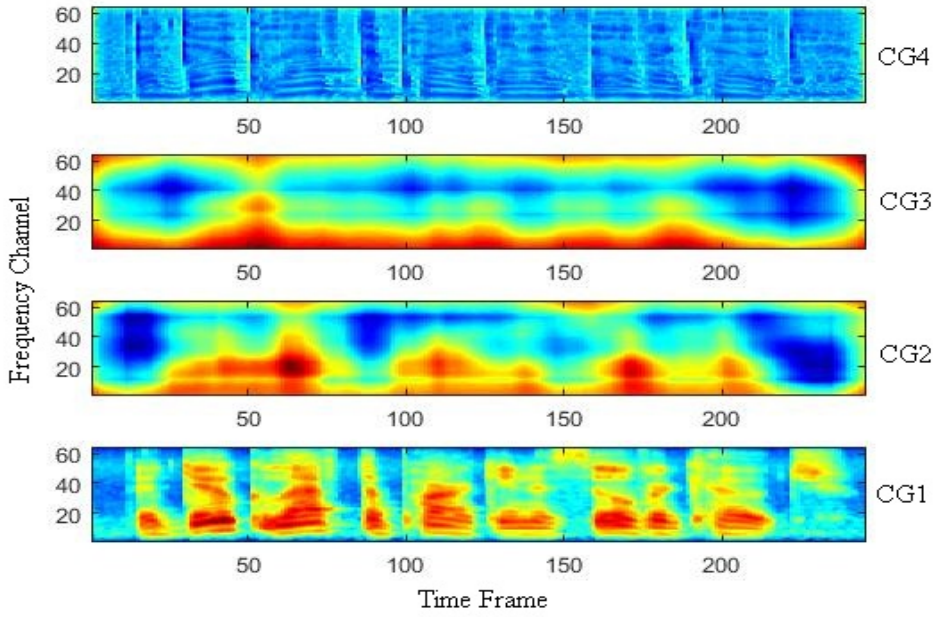


Figure 4.2: Visualization of the new feature set including four cochleagrams of CG1, CG2, CG3 and CG4 where each cochleagram has 64 frequency channels.

smaller and larger spectrotemporal context, respectively and CG4 detects the edges of the original cochleagram.

4.6.2 System description

The effects of our proposed acoustic feature on training of two common types of networks, DNN and long short-term memory (LSTM) for mask estimation are investigated. Both networks were designed using a mean squared error (MSE) cost function and the Adam algorithm [117] for back-propagation optimization. The DNN was constructed with three hidden layers each with 1024 hidden units using the exponential linear units (ELUs) activation function [119]. The output layer used a sigmoid activation function, and training was completed using mini-batch gradient descent with the size of 1024. The LSTM network [112] is a type of recurrent neural network (RNN), and was chosen to specifically model temporal dynamics of speech. This network consists of three hidden layers including one fully connected layer of 1024 units, two LSTM layers of 512 units with the time step of 128 (1.28 second) and the batch size of 256. A dropout rate of 0.2 was used

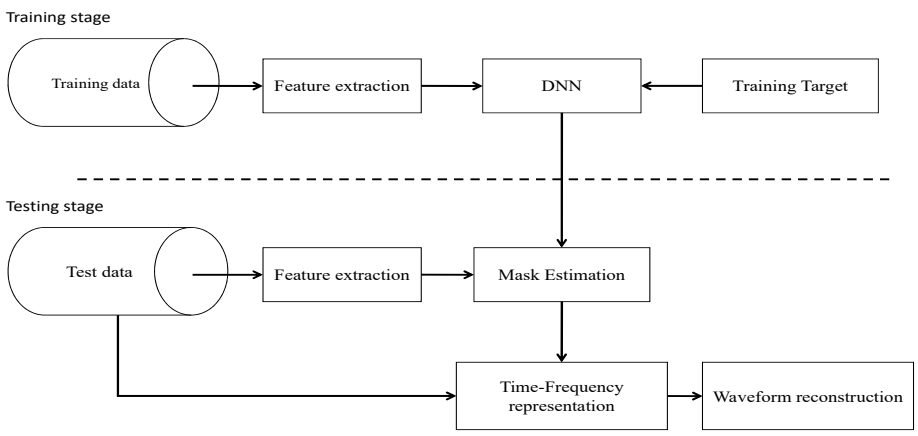


Figure 4.3: The block diagram of the DNN-based speech separation predicting time frequency masks.

on the hidden layers of both networks in order to avoid overfitting [134] and the number of training epochs is 50. Moreover, a five-frame context window (two on the left, two on the right) was fed to the networks to capture the temporal aspects of reverberation [132]. All features were normalized to have zero mean and unit variance on each frequency channel before feeding to the networks, and batch normalization was applied on hidden layers for faster training [123].

The DNN and LSTM both learn the relationship between training data and its associated IRM [3] for separating speech from noise. The IRM has been previously shown to have better performance than the ideal binary mask (IBM) [3]; therefore, we apply the IRM as the training target. The block diagram of DNN-based speech separation with its three important components is shown on Fig. 4.3.

4.6.3 Experimental setup

4.6.3.1 Speech enhancement task

Speech material from IEEE corpus [92], which includes 720 utterances spoken by a single male speaker, were used for both training and test datasets to evaluate performance of DMRCG for speech separation. The total number of 600 sentences for training and development sets were selected, while 120 sentences were used for testing. Moreover, the Factory, Operation room, Engine and Babble noises from the NOISEX repository [135] were used as additive background noise.

The noisy-reverberant mixtures (Total=15000; training=13500, development=1500) used for training the neural networks in this study were generated first by convolving the clean speech signal with a randomly selected room impulse response (IR) by simulating reverberant room conditions according to the image method technique [136] followed by adding a random selection of the different background noises. The room impulse response generator [137] was used to create IRs with different reverberation times (T_{60}) in the range of $\{0.3-0.9\}$ s and an anechoic condition with $T_{60} = 0.0$ s. For all simulations, the room dimension was defined as $(10 \times 9 \times 8)$ m and the microphone placement at $(3, 4, 1.5)$ m. The target speaker was simulated at a random position at distance in between 0.5–3 m from the microphone. A random signal-to-noise ratio in the range of [-5,0] dB was used for the training/development set and was based on the reverberant target speech instead of anechoic speech.

Two test sets were generated detailed below. Briefly, one set used real IRs recorded at the University of Surrey [138] and the other was based on simulated IRs from the CATT database [139].

- Test Set 1: 1200 mixtures were created from 120 IEEE utterances convolved with Surrey IRs and mixed with random segment of background noise. The Surrey database includes IRs recorded at 1.5 m between the sound source and the microphone from four reverberant rooms with T_{60} of 0.32 s, 0.47 s, 0.68 s and 0.89 s, respectively.

- Test Set 2: 1200 mixtures were created from 120 IEEE utterances convolved with CATT IRs and mixed with a random segment of background noise. The IRs from CATT database were

selected with T_{60} of 0.2s, 0.4s, 0.6s, 0.8s and a distance of 1.5 m between the sound source and the microphone.

The SNR for both test sets was in the range [-5,-2] dB, and no similar noise segment, speech utterance or IR were included in both training and testing data sets. Both the speech utterances and IRs were resampled to 16 KHz before generating noisy-reverberant monaural mixtures for training and testing.

Table 4.1 and Table 4.2 report the results for speech enhancement task using different acoustic features and different neural network structures.

4.6.3.2 Speaker separation task

Furthermore, the proposed feature was evaluated for speaker separation task when there are two active speakers. Two different models were trained for supervised speaker separation. One model was trained and tested when both the target speaker and interfering speaker are male, while the second model was trained and tested when target speaker is male and interfering speaker is female. The same dataset was used for training as described in Section 4.6.3.1 with the total number of 15000 noisy-reverberant mixtures (training=13500, development=1500); however, interfering speakers (female and male) are extracted from TIMIT corpus [140]. For training stage, 30 utterances from different male speakers and 30 utterances from different female speakers were selected from TIMIT database, while 10 sentences were used for testing. Therefore, for each speaker separation model, the total number of 1200 reverberant multi-speaker mixtures were generated as testing dataset. The same procedure was used to generate the reverberant training, development and testing dataset; however, the interfering speaker was convolved with the same IRs as the target speaker and then added to the reverberant target speech instead of background noise. In this case, only the recorded IRs [138] were used to generate the test dataset, while the same simulated IRs generated by room impulse response generator [136, 137] were used to generate the training and development dataset. The training SNRs were in the range of [-5,5] dB while the testing SNR is -5 and -2 dB. Table 4.3 shows the results for speaker separation task for different feature sets,

different network structure and interfering speakers.

4.6.3.3 Speech enhancement task for unseen condition

In addition, the performance of the proposed feature set was evaluated when there is a mismatch between the noises used in training and testing (unmatched condition). The model explained in Section 4.6.3.1 was used as our trained model and the new test dataset (1200 noisy-reverberant mixtures) was generated with unmatched noises. In this case, the unmatched noises were Cafe, Living room, Kitchen and Car that were derived from the DEMAND [141] repository. The results of speech enhancement for unmatched condition are shown on Table 4.4.

4.6.4 Evaluation criteria

The speech separation performance was evaluated with three different criteria: short-time objective intelligibility (STOI) [142], perceptual evaluation of speech quality (PESQ) [143] and, frequency-weighted segmental signal to noise ratio (SNR_{fw}) [144]. STOI calculates the correlation between the temporal envelopes of clean and processed signals and results in a score ranging from 0 to 1 with higher score indicating greater objective speech intelligibility [142]. The PESQ was used as an objective measure of speech quality of processed speech relative to clean speech, and results in a score between [-0.5,4.5] where higher values are reflective of greater speech quality [143]. Finally, SNR_{fw} was used to compute the weighted signal-to-noise ratio in each critical band [144]. The scores for all three criteria for different spectral and gammatone based acoustic features were reported in Table 4.1 and Table 4.2. Table 4.3 shows the results of speaker separation for different features for both female and male interfering speakers using DNN and LSTM. Furthermore, Table 4.4 reports the scores of all three criteria when there is a mismatch between the noises used in training and testing stage.

4.6.5 Results

4.6.5.1 Matched noise simulation

Table 4.1 shows the values of STOI (%), PESQ and SNR_{fw} for noisy mixtures and processed signals using common spectral and gammatone-domain features with DNN to learn the speech separation masks for four different background noises. In both cases of recorded IR and simulated ones, DMRCG has the highest improvement in terms of STOI and PESQ for all noise types. However, the SNR_{fw} of MRCG is slightly higher than DMRCG for most of the noise types and IRs. The highest improvement of STOI, PESQ and SNR_{fw} is for the Engine background noise with the increase of 18.5%, 0.51 and 5.14 (dB), respectively by DMRCG in recorded IRs. The same pattern can be seen when simulated IRs were used to generate reverberant condition.

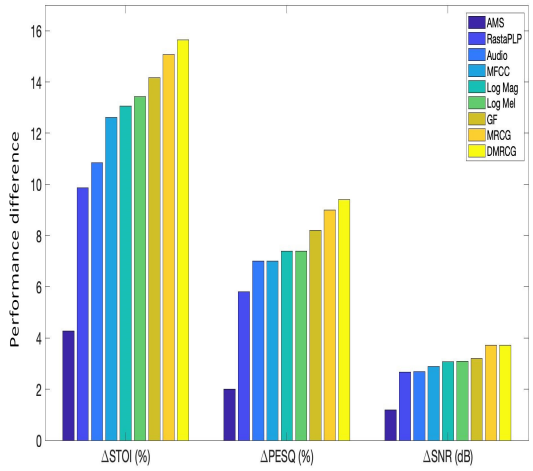
Table 4.2 demonstrates the average values of STOI (%), PESQ and SNR_{fw} over all noise types using two different neural network structures and two types of IRs. The average of STOI, PESQ and SNR_{fw} of unprocessed mixtures using recorded IRs are 56.61, 1.47 and 2.43 (dB), respectively, and 56.52, 1.40 and 2.13 (dB) for simulated IR scenario using DNN. For real room IRs, the DMRCG feature increases STOI, PESQ and SNR by 15.65%, 0.47 and 3.72(dB), respectively compared to the noisy test mixtures (see Table 4.2), and for simulated IRs, 14%, 0.39 and 2.99 (dB), respectively. In addition, the DMRCG feature outperforms the STOI for the best spectral feature (Log-Mel) and the best gammatone-based feature (MRCG) by 2.23% and 0.58% for real-room IRs and 2.76% and 0.47% for simulated IRs.

Our analysis is repeated using LSTM to evaluate whether the DMRCG feature is robust to neural network type. Table 4.2 compares the performance of the same spectral and gammatone-based features for mixtures generated using simulated and recorded IRs in terms of STOI, PESQ and SNR. Using LSTM to learn the speech separation mask, our DMRCG feature provides higher values in all three criteria in comparison to the DNN-trained masks. In addition, the rate of increase for STOI, PESQ and SNR are 15.81%, 0.46 and 3.79 (dB) for recorded IRs and 14.13%, 0.39 and 3.06 (dB) for simulated IRs, respectively. For all other features, the LSTM training im-

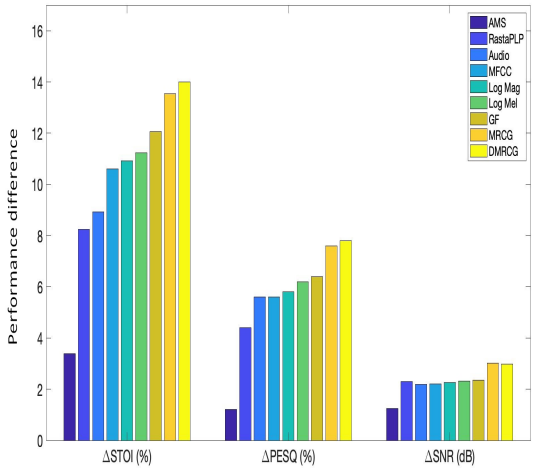
Table 4.1: STOI(%), PESQ and SNR_{fw} (dB) for different background noise using DNN with recorded and simulated IRs.

IR	Feature	**STOI**				**PESQ**				**SNR_{fw}**			
		Factory	*Babble*	*Engine*	*Opt.*	*Factory*	*Babble*	*Engine*	*Opt.*	*Factory*	*Babble*	*Engine*	*Opt.*
Recorded	Noisy mixture	54.57	55.63	57.65	58.51	1.36	1.52	1.54	1.44	1.89	2.43	1.68	3.72
	DMRCG	**69.84**	**67.58**	**76.15**	**75.31**	**1.87**	**1.76**	**2.05**	**2.07**	5.63	5.49	6.82	6.62
	MRCG	69.20	67.03	75.58	74.75	1.85	1.75	2.03	2.05	**5.78**	**5.54**	**6.89**	**6.74**
	GF	67.96	65.88	75.06	74.04	1.79	1.70	1.99	2.01	5.14	4.84	6.26	6.27
	Log-Mel	66.85	65.47	74.42	73.22	1.73	1.68	1.97	1.98	5.02	4.87	6.10	6.09
	Log-Mag	66.06	65.37	74.09	73.00	1.72	1.68	1.96	1.98	4.84	4.96	6.18	5.99
	MFCC	65.96	64.66	73.55	72.54	1.73	1.66	1.93	1.96	4.75	4.72	5.88	5.91
	RastaPLP	63.88	62.22	69.59	70.05	1.69	1.63	1.84	1.87	4.55	4.58	5.41	5.82
	Audio Frame	64.62	63.36	71.50	70.20	1.75	1.69	1.92	1.93	4.66	4.47	5.82	5.51
	AMS	57.38	57.69	64.75	63.55	1.47	1.58	1.59	1.62	2.89	3.11	3.74	4.70
Simulated	Noisy mixture	54.76	55.62	57.11	58.50	1.31	1.40	1.49	1.39	1.65	2.08	1.37	3.41
	DMRCG	**68.49**	**65.30**	**74.42**	**73.74**	**1.75**	**1.62**	**1.87**	**1.90**	4.68	4.56	5.60	**5.63**
	MRCG	68.06	65.00	73.89	73.10	1.74	1.60	1.87	1.89	**4.71**	**4.64**	**5.66**	5.57
	GF	66.27	63.56	72.60	71.73	1.66	1.56	1.82	1.85	4.09	3.81	4.90	5.11
	Log-Mel	65.18	63.27	71.82	70.64	1.64	1.55	1.81	1.82	4.10	3.79	4.89	5.02
	Log-Mag	64.80	63.20	71.23	70.54	1.62	1.53	1.80	1.81	3.80	4.00	4.90	4.93
	MFCC	64.64	62.58	71.12	70.04	1.62	1.52	1.77	1.80	3.96	3.81	4.64	4.92
	RastaPLP	62.75	60.82	67.11	68.24	1.58	1.49	1.68	1.71	4.03	3.93	4.59	5.11
	Audio Frame	63.41	61.71	68.07	68.48	1.65	1.55	1.71	1.79	4.14	3.51	4.75	4.83
	AMS	56.48	56.59	63.33	63.08	1.38	1.46	1.48	1.51	2.68	2.85	3.47	4.43

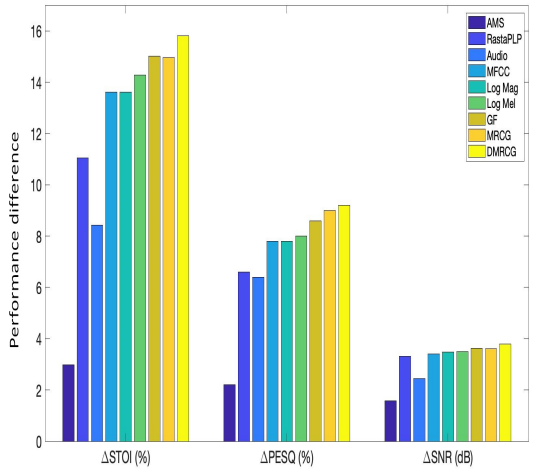
proves performance across all three criteria, except for MRCG, which slightly decreases compared to DNN-based training. Therefore, the proposed DMRCG feature, like all features but MRCG, is able to take advantage of recurrent learning from the LSTM training, is robust to network type, and



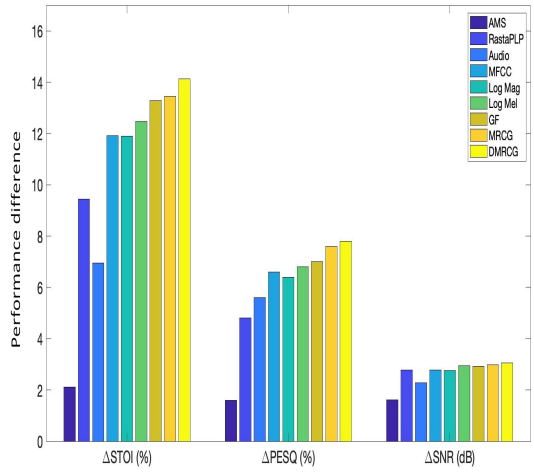
(a) Using recorded IRs - DNN model



(b) Using simulated IRs - DNN model



(c) Using recorded IRs - LSTM model



(d) Using simulated IRs - LSTM model

Figure 4.4: Improvements of STOI (%), PESQ (%) and SNR_{fw} (dB) for each feature set using recorded IRs and simulated IRs for DNN and LSTM based speech separation models.

generally improves speech separation performance relative to other common spectral and gammatone domain features.

The improvements in STOI (%), PESQ (%) and SNR_{fw} (dB) for different room simulations, different feature sets and neural network are illustrated on Fig. 4.4. From the given charts, it can be observed that the total amount of improvement in STOI is generally higher than PESQ and SNR for

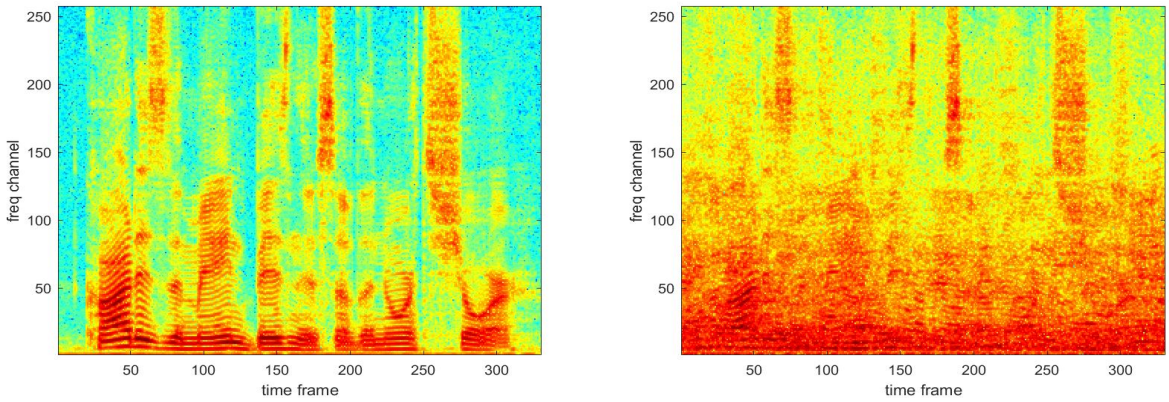
Table 4.2: STOI(%), PESQ and SNR_{fw} averaged over 1200 mixtures using different network types of DNN and LSTM for matched noise condition.

Network	Feature	**Recorded IR**			**Simulated IR**		
		STOI	***PESQ***	SNR_{fw}	***STOI***	***PESQ***	SNR_{fw}
DNN	Noisy mixture	56.61	1.47	2.43	56.52	1.40	2.13
	DMRCG	**72.26**	**1.94**	6.15	**70.52**	**1.79**	5.12
	MRCG	71.68	1.92	**6.24**	70.05	1.78	**5.15**
	GF	70.78	1.88	5.64	68.58	1.72	4.48
	Log-Mel	70.03	1.84	5.53	67.76	1.71	4.45
	Log-Mag	69.67	1.84	5.50	67.44	1.69	4.39
	MFCC	69.22	1.82	5.32	67.13	1.68	4.34
	RastaPLP	66.47	1.76	5.09	64.76	1.62	4.42
	Audio Frame	67.46	1.82	5.12	65.45	1.68	4.31
	AMS	60.88	1.57	3.62	59.91	1.46	3.37
LSTM	**DMRCG**	**72.42**	**1.93**	**6.22**	**70.65**	**1.79**	**5.19**
	MRCG	71.57	1.92	6.03	69.97	1.78	5.12
	GF	71.63	1.90	6.06	69.80	1.75	5.05
	Log-Mel	70.89	1.87	5.92	68.99	1.74	5.07
	Log-Mag	70.22	1.86	5.91	68.41	1.72	4.89
	MFCC	70.22	1.86	5.83	68.43	1.73	4.91
	RastaPLP	67.66	1.80	5.74	65.96	1.64	4.91
	Audio Frame	65.04	1.79	4.87	63.47	1.68	4.41
	AMS	59.60	1.58	4.00	58.63	1.48	3.74

all feature sets and testing conditions. Moreover, all feature sets have higher improvements in all criteria for the a recorded IR simulation than a simulated IR for both DNN and LSTM separation models. The spectrogram of the processed signals using different acoustic features with 2 different background noise and reverberation times were illustrated on Fig. 4.5 and Fig. 4.6.

4.6.5.2 Speaker separation

In Table 4.3, the results of the speaker separation task were reported. Two testing conditions were examined: 1) male target speaker and female interfering speaker, 2) male target speaker and male interfering speaker. It should be noted that new training/development datasets were generated for speaker separation task, that includes two active speakers with no background noise in reverberant rooms. Table 4.3 shows that DMRCG has better performance in increasing STOI, PESQ and SNR when the interfering speaker is female for both network structures. However, MRCG has slightly



(a) Clean

(b) Noisy-reverberant mixture

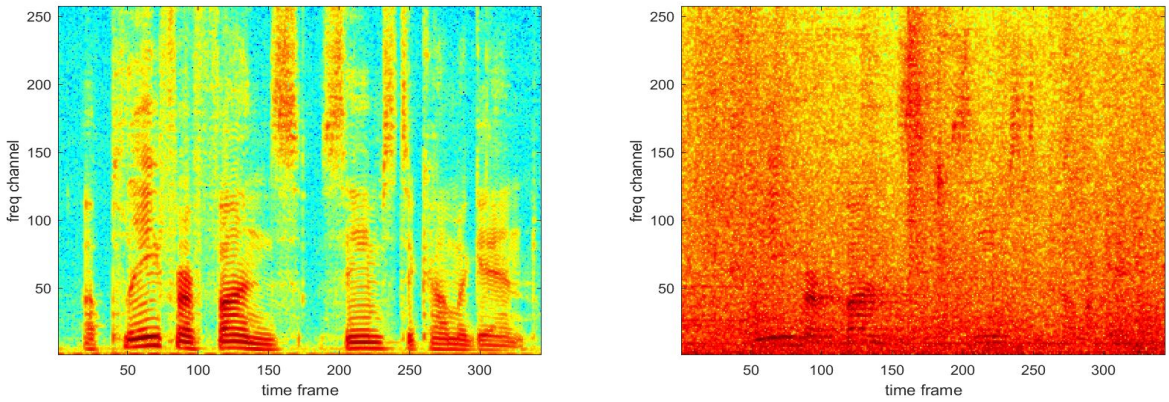
(c) Mel-Spectrum output

(d) GF output

(e) MRCG output

(f) DMRC output

Figure 4.5: Visualization of processed signals using different feature set for training DNN model for a IEEE clean speech mixed with babble noise at -5 dB SNR and RT_{60} of 0.89 second.



(a) Clean

(b) Noisy-reverberant mixture

(c) Mel-Spectrum output

(d) GF output

(e) MRCG output

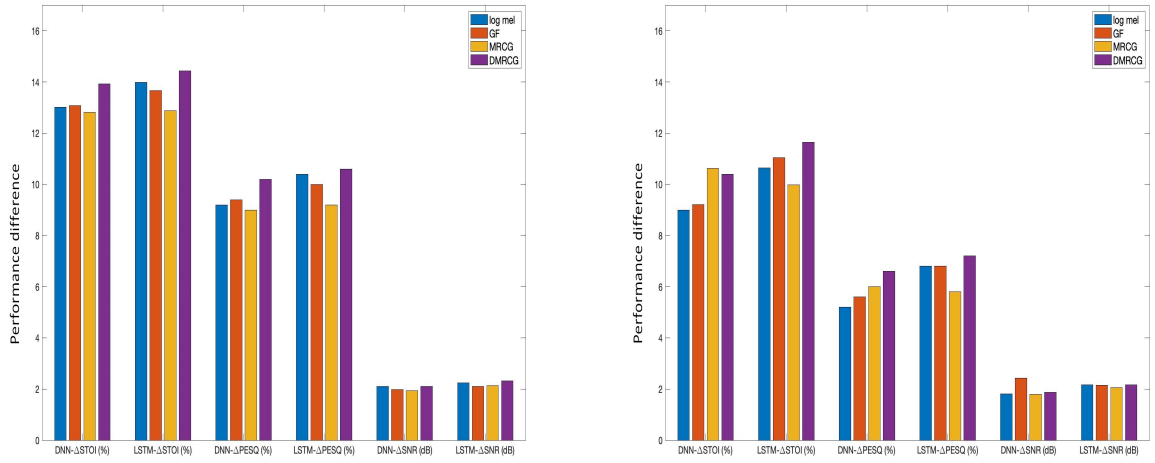
(f) DMRC output

Figure 4.6: Visualization of processed signals using different feature set for training DNN model for a IEEE clean speech mixed with factory noise at -2 dB SNR and RT_{60} of 0.89 second.

better performance in increasing speech intelligibility when the interfering speaker is male in DNN based model. Nevertheless, MRCG is not robust to the network type and its performance decreases slightly when the recurrent neural network was used. In addition, the rate of increase for STOI, PESQ and SNR are 13.91%, 0.51 and 2.11 (dB) for female speaker and 10.39%, 0.33 and 1.88 (dB) for male speaker, respectively using DMRCG with DNN. For LSTM model, the rate of increase for STOI, PESQ and SNR are 14.43%, 0.53 and 2.32 (dB) for female speaker and 11.64%, 0.36 and 2.16 (dB) for male speaker, respectively using DMRCG. The LSTM model has higher rates in STOI, PESQ and SNR for the top four features for both female and male speakers compared to the DNN based speaker separation model. However, the amount of improvement in all these three criteria is smaller than the speech enhancement model explained in Section 4.6.5.1. The increase of speech intelligibility and speech quality for the two different speaker separation models for different testing conditions were plotted in detail on Fig. 4.7. All the results are averaged only for target speaker and we have excluded the interfering speaker results in our report.

Table 4.3: STOI(%), PESQ and SNR_{fw} averaged over 1200 mixtures using different network types of DNN and LSTM for speaker separation task with female and male interfering speaker (target speaker is male).

Network	Feature	**Female speaker**			**Male speaker**		
		STOI	***PESQ***	$\boldsymbol{SNR_{fw}}$	***STOI***	***PESQ***	$\boldsymbol{SNR_{fw}}$
DNN	Noisy mixture	58.06	1.45	3.61	56.35	1.44	3.16
	DMRCG	**71.99**	**1.96**	**5.72**	66.74	**1.77**	5.04
	MRCG	70.88	1.90	5.54	**66.97**	1.74	4.96
	GF	71.14	1.92	5.59	65.56	1.72	**5.59**
	Log-Mel	71.08	1.91	5.71	65.35	1.70	4.97
	Log-Mag	71.07	1.91	5.58	65.15	1.70	4.78
	MFCC	70.09	1.85	5.61	64.92	1.69	4.89
	RastaPLP	65.71	1.70	5.01	61.80	1.57	4.41
	Audio Frame	69.16	1.84	5.02	62.48	1.65	4.39
	AMS	65.41	1.64	4.87	58.99	1.50	4.07
LSTM	**DMRCG**	**72.49**	**1.98**	**5.93**	**67.99**	**1.80**	5.32
	MRCG	70.94	1.91	5.74	66.33	1.73	5.22
	GF	71.73	1.95	5.72	67.39	1.78	5.31
	Log-Mel	72.05	1.97	5.85	66.99	1.78	5.33
	Log-Mag	71.92	1.96	5.74	66.40	1.77	5.14
	MFCC	71.10	1.91	5.79	67.32	1.76	**5.35**
	RastaPLP	66.92	1.75	5.14	62.56	1.60	4.69
	Audio Frame	68.16	1.79	5.35	61.93	1.64	4.73
	AMS	66.15	1.65	4.98	58.63	1.50	4.27



(a) Female interfering speaker

(b) Male interfering speaker

Figure 4.7: Improvements of STOI (%), PESQ (%) and SNR_{fw} (dB) for the top four feature set using DNN and LSTM based speaker separation models.

4.6.5.3 Unmatched noise simulation

Further analyses were done to examine the generalization of the neural network to unseen conditions. Table 4.4 reports the result of different feature sets for unseen noises . Both DNN and LSTM were trained with noisy mixtures where the background noise were Factory, Babble, Engine and Operation room, while the testing condition includes unseen (unmatched) noise of Cafe, Living room, Kitchen and Car. The results on Table 4.4 demonstrate that the proposed DMRCG feature set has better performance compared to other features for both recorded and simulated IR with both network structure.

The rate of increase for STOI, PESQ and SNR are 6.99%, 0.29 and 1.76 (dB) for recorded IRs and 6.93%, 0.25 and 1.11 (dB) for simulated IRs, respectively for DNN model. However, both MRCG and DMRCG's performance decreases slightly when we have recurrent neural network for unseen noises, which means that both of these features depend on the training condition while other spectral features are independent of the input data when recurrent neural network was used as training model. Furthermore, the increase in PESQ and SNR criteria is higher for recorded IRs

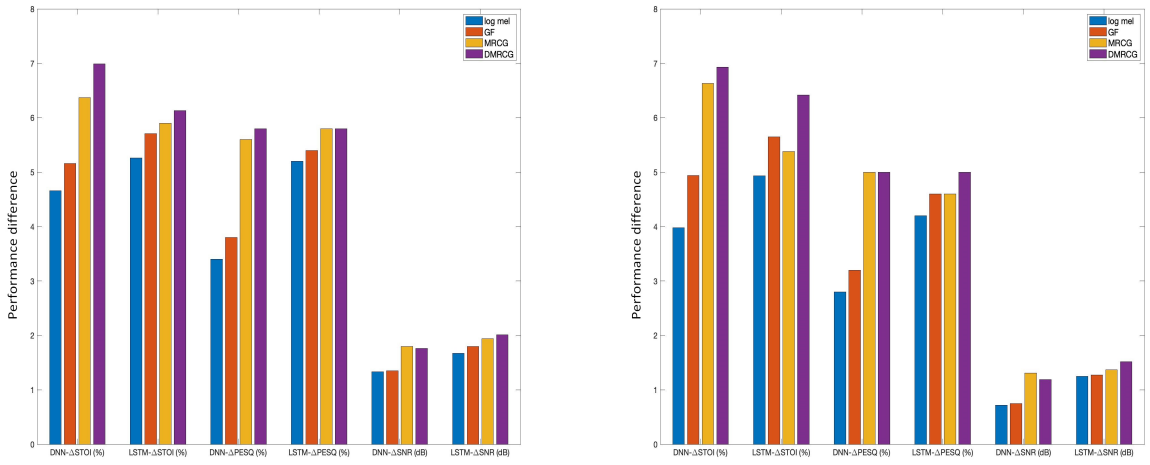
Table 4.4: STOI(%), PESQ and SNR_{fw} averaged over 1200 mixtures using different network types of DNN and LSTM for unmatched noise.

Network	Feature	**Recorded IR**			**Simulated IR**		
		STOI	***PESQ***	$\boldsymbol{SNR_{fw}}$	***STOI***	***PESQ***	$\boldsymbol{SNR_{fw}}$
DNN	Noisy mixture	69.37	2.01	5.05	67.68	1.84	4.16
	DMRCG	**76.36**	**2.30**	6.81	**74.61**	**2.09**	5.35
	MRCG	75.74	2.29	**6.85**	74.31	2.09	**5.47**
	GF	74.53	2.20	6.40	72.62	2.00	4.91
	Log-Mel	74.03	2.18	6.38	71.66	1.98	4.88
	Log-Mag	73.76	2.18	6.37	71.69	1.98	4.88
	MFCC	73.05	2.17	6.33	70.83	1.95	4.80
	RastaPLP	73.85	2.22	6.54	72.14	2.01	5.45
LSTM	**DMRCG**	**75.50**	**2.30**	**7.06**	**74.10**	**2.09**	**5.68**
	MRCG	75.27	2.30	6.99	73.06	2.07	5.53
	GF	75.08	2.28	6.85	73.33	2.07	5.43
	Log-Mel	74.63	2.27	6.72	72.61	2.05	5.41
	Log-Mag	74.66	2.26	6.76	72.50	2.04	5.36
	MFCC	74.46	2.24	6.67	72.86	2.01	5.40
	RastaPLP	73.89	2.23	6.76	71.53	2.05	5.62

compared to the simulated ones. Both recorded and simulated IRs in unseen noise condition, have less improvement compared to the matched noise condition. In conclusion, the improvements for all three criteria are smaller compared to seen noise that indicates that both networks have poor generalization to unseen conditions using any type of feature sets. The detailed results of unmatched noise simulation were presented on Fig. 4.8.

4.6.6 Discussion

This chapter discussed in detail supervised speech separation based on deep learning techniques. Three important sections of supervised speech separation were explained such as learning machines, training target and acoustic features. Furthermore, an extensive feature study was done to evaluate the performance of supervised speech separation models in different room simulations, various SNRs and background noise. Moreover, a novel monaural acoustic feature based on gammatone filterbank (DMRCG) was proposed for a DNN masking-based speech separation in noisy-reverberant conditions with negative SNR, a variety of reverberation times and different noise types. The evaluation of speech separation performance shows that the proposed DMRCG



(a) Using recorded IRs for unmatched noise

(b) Using simulated IRs for unmatched noise

Figure 4.8: Improvements of STOI (%), PESQ (%) and SNR_{fw} (dB) for top four feature set using DNN and LSTM based speech separation model for unseen noises.

feature outperforms common acoustic features in terms of STOI and PESQ, suggesting a greater preservation of the target speech acoustics in noisy reverberant environment. Furthermore, this novel feature requires less computation compared to MRCG while having the same dimension, and is more robust to the neural network type used to learn the optimal speech separation T-F masks.

Generalization is a critical issue in supervised learning algorithms. The generalization of different acoustic features were also examined to different unmatched conditions such as room simulations and background noise type. The gammatone-domain feature sets outperform other spectral features in all these experiments, which indicates that the features extracted based on human perception are better features for training the neural networks for speech separation.

In addition, the acoustic features were evaluated for speaker separation tasks. The proposed acoustic feature performs better than other common features in speaker separation task. The robustness of DMRCG was examined to two network types and two speaker types. DMRCG gives promising results in speaker separation task; however, when the target speaker and interfering speaker are same gender, the improvement in speech intelligibility and speech quality are limited

using any type of features.

Finally, the performance of DMRCG was examined for unseen noise conditions and the reports show that this feature has highest values in terms of speech quality and speech intelligibility to unseen noises, which can be useful in generalization of the neural network to unseen data and unseen environment; nonetheless, all these features have limited ability for generalization to unseen noise conditions. On the other hand, the proposed feature set is robust to the network type when there is a mismatch between the training noise and testing noise for both simulated rooms and real recording in rooms.

Chapter 5

Binaural acoustic features for DNN-based speech separation

In this chapter, different binaural acoustic features were examined to improve speech separation algorithm based on deep neural network (DNN) in noisy and reverberant conditions. First, the overview of the most common binaural features is given. Then, a new complementary binaural feature set is presented to improve the performance of the DNN-based supervised speech separation by estimating time-frequency masks from binaural features. The new complementary binaural features are based on the interaural coherence and the difference of energy between the binaural noisy-reverberant mixtures. The results of our analyses show that the proposed binaural feature set combined with commonly-used spatial features improves the objective speech intelligibility, speech quality and signal-to-noise ratio of the separated speech in noisy and reverberant environments.

5.1 Introduction

Despite considerable past research in binaural speech separation, it is still a challenging task to separate sources from a 2-channel recording in noisy and reverberant conditions. More recently, deep learning has emerged as a method for improving speech separation using a supervised learning approach. In particular, DNN can be constructed to utilize combinations of different acoustic features (binaural and monaural) to generate time frequency masks for speech separation.

Interaural level difference (ILD) and interaural time difference (ITD) have been identified as important binaural features and have been used for speech separation when the target speech is fixed at 0° azimuth [73, 145]. When the target and interfering sources are coming from different

spatial directions, both ILD and ITD can be used to estimate time frequency masks for speech separation. However, for co-located sources or sources near each other, binaural features are not useful for separation. In this case, the monaural features that are independent of the location of the sources can be used for source separation. In [145], spectral features obtained using a fixed beamformer were combined with binaural features to improve speech separation performance. Additional work by May [146] combined the additional features of interaural coherence (IC), interaural phase difference (IPD) and ILD with spectral features using Mel-filterbanks for neural network-based postfilter for speech dereverberation in diffuse babble noise.

In this chapter, a set of new spatial features for binaural supervised speech separation based on DNN in noisy and reverberant environments is proposed. The proposed spatial features combined with previously validated binaural features (e.g., ILD & IPD) provide complementary information for speech separation that results in improved objective measures of speech intelligibility and quality. Specifically, speech separation procedure was tested with the target speech located in different random positions with diffuse multi-talker babble as background noise. The DNN is trained to estimate the ideal ratio mask (IRM) [3] from binaural features to separate target speech from background noise.

It should be noted that the proposed complementary binaural feature was tested for a speaker dependent model where the same speaker was used for both training and testing. In this case, the utterances from a single male speaker from IEEE dataset [92] was used for both training and testing. Moreover, a diffuse babble noise was employed as additive background noise for both training and testing condition.

5.2 System description

A simulation was used to generate binaural signals with background multi-talker babble noise and reverberation by convolving the target speaker with head-related impulse responses (HRIRs) for anechoic rooms or binaural room impulse responses (BRIRs) for reverberant rooms. Next, binaural features used for speech separation are obtained using the short time Fourier transform

(STFT) followed by individual processing for each feature defined in Section 5.2.1. The DNN then learns the mapping from the binaural spatial features to the ideal ratio mask, which is then applied to the magnitude spectrum of the noisy-reverberant input mixtures to estimate the clean magnitude spectrum. Finally, the clean speech waveform is resynthesized using the phase of noisy-reverberant signal and the estimated magnitude spectrum.

5.2.1 Binaural feature extraction

In this section, the two binaural features previously identified for their importance in speech separation, IPD and ILD were explained. All binaural features were obtained initially from the STFT of the binaural signals using a hamming window of length of 32 ms with 8 ms time shift (75% overlap) with the sampling rate of 16 KHz. Then, a 512-point fast Fourier transform (FFT) is applied to each time frame to obtain the time-frequency representation of the binaural input signals. The STFT of the left and the right ear signal is given as $X_L(\omega,\tau)$ and $X_R(\omega,\tau)$ where τ and ω corresponds to the time frame and frequency bin, respectively. Following initial STFT analysis, IPD and ILD are obtained after processing with a 64-channel Mel-filterbank. Each channel of the filterbank c applies a frequency-dependent response of $G(c,\omega)$ over frequency range of 50 Hz to 8000 Hz with mel-frequency spacing [147].

The interaural phase difference (IPD) is calculated by the phase of the ratio between the left and the right ear signal in TF domain. Modified IPD can be determined by the phase of the ratio of $X_L(\omega,\tau)$ and $X_R(\omega,\tau)$ for each filterbank channel c according to [146]

$$\mathrm{IPD}(c,\tau) = G(c,\omega)\ \arg\left(\frac{X_L(\omega,\tau)}{X_R(\omega,\tau)}\right). \tag{5.1}$$

In order to remove spatial aliasing when estimating IPD, all phase differences are wrapped to the range of $[-\pi,\pi]$.

Interaural level difference (ILD) is defined as the ratio of the amplitude between the left and the right ear in TF domain that is expressed in dB. In other words, the ILD corresponds to the energy

ratio for each TF unit. The modified ILD is computed with filtering of ILD with 64-channel Mel-filterbank of $G(c,\omega)$ as follows [146]

$$\mathrm{ILD}(c,\tau) = G(c,\omega)\left(20\log_{10}\left(\left|\frac{X_L(\omega,\tau)}{X_R(\omega,\tau)}\right|\right)\right). \tag{5.2}$$

Another important binaural feature is interaural time difference (ITD), which is usually calculated by normalized cross-correlation function (CCF) between the left and the right ear signal. The ITD is estimated by the lag corresponding to the maximum in the CCF.

The interaural coherence (IC) is an important binaural cue that is usually used for speech dereverberation. IC can be important for enhancing speech separation in reverberant conditions. One of our proposed features is based on the IC, which is defined as the normalized cross-correlation function according to [99, 104]

$$\mathrm{IC}(\omega,\tau) = \frac{|\Phi_{LR}(\omega,\tau)|}{\sqrt{\Phi_{LL}(\omega,\tau)\Phi_{RR}(\omega,\tau)}}, \tag{5.3}$$

where the $\Phi_{LL}(\omega,\tau), \Phi_{RR}(\omega,\tau)$ and $\Phi_{LR}(\omega,\tau)$ are the exponentially weighted short-term auto-correlation and cross-correlation functions defined as:

$$\Phi_{LL}(\omega,\tau) = \alpha\,\Phi_{LL}(\omega,\tau-1) + (1-\alpha)\,X_L(\omega,\tau)X_L^*(\omega,\tau) \tag{5.4}$$

$$\Phi_{RR}(\omega,\tau) = \alpha\,\Phi_{RR}(\omega,\tau-1) + (1-\alpha)\,X_R(\omega,\tau)X_R^*(\omega,\tau) \tag{5.5}$$

$$\Phi_{LR}(\omega,\tau) = \alpha\,\Phi_{LR}(\omega,\tau-1) + (1-\alpha)\,X_L(\omega,\tau)X_R^*(\omega,\tau) \tag{5.6}$$

The smoothing factor of α is adjusted by the time constant of T as $\alpha = \exp(-STFT_{\text{shift}}/T)$ where T is 10 ms. The histograms of common binaural features of ILD and IPD for two clean speech signals (coming from different directions) were visualized on Fig. 5.1 and Fig. 5.2. It should be noted that the histograms show the original ILD and IPD (exclude processing with $G(c,\omega)$ filterbank).

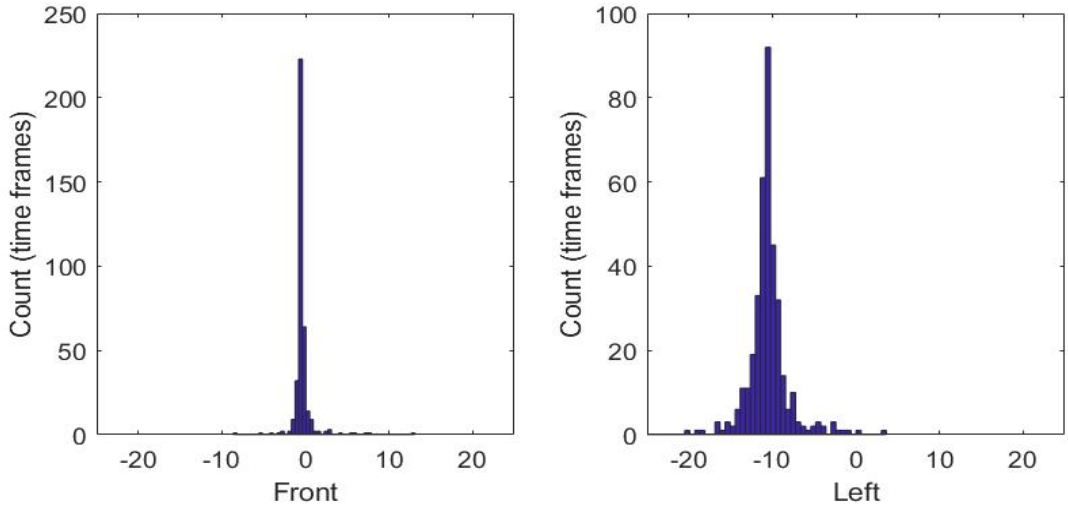


Figure 5.1: Histogram of ILD feature on high frequency channel for two clean speech signals where one is coming from the front side and the other coming from the left side.

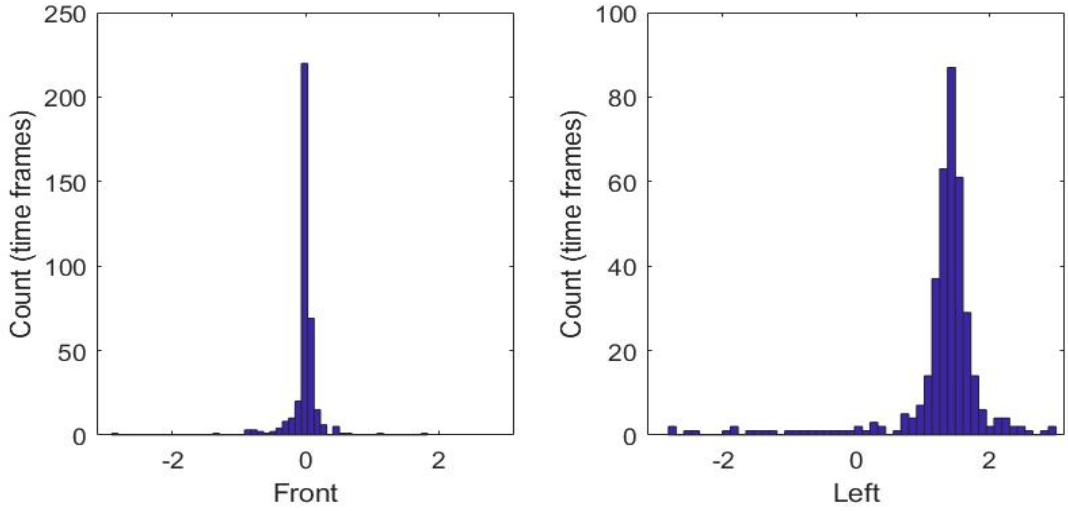


Figure 5.2: Histogram of IPD feature on low frequency channel for two clean speech signals where one is coming from the front side and the other is coming from the left side.

5.2.2 Proposed complementary binaural feature

The proposed feature set includes three new binaural features that complement the typical features of ILD and IPD, to improve the performance of the DNN-based speech separation algorithm. The first feature is the real part of IC, while the second and third feature depends on the difference of the energy between the left and the right ear signals in TF domain. The proposed features are then

concatenated with the ILD and IPD features for training the DNN to separate the target speech from the background noise.

1. The first feature is a partial component of IC where only the real part of cross-correlation is used for IC estimation. Therefore, the partial interaural coherence $\mathrm{IC_p}$ is defined as

$$\mathrm{IC_p}(c,\tau) = G(c,\omega)\frac{\mathrm{Real}\{\Phi_{LR}(\omega,\tau)\}}{\sqrt{\Phi_{LL}(\omega,\tau)\Phi_{RR}(\omega,\tau)}}, \tag{5.7}$$

   where the $\Phi_{LL}(\omega,\tau), \Phi_{RR}(\omega,\tau), \Phi_{LR}(\omega,\tau)$ are defined in Eq.5.4, Eq.5.5 and Eq.5.6. The partial IC feature is filtered by 64 channel Mel-filterbank of $G(c,\omega)$.

2. The second feature is defined as the difference between the left ear and the right ear energy according to

$$\mathrm{D}(c,\tau) = G(c,\omega)\left(10\log_{10}\left|X_L^2(\omega,\tau) - X_R^2(\omega,\tau)\right|\right). \tag{5.8}$$

   The difference of energy between the left and right ear signals is then filtered with the 64 channel Mel-filterbank of $G(c,\omega)$ as described in [146] .

3. The third spatial feature is defined as the energy difference between the left ear and the right ear energy normalized by the total energy of the left ear and the right ear signal on Mel-scaling that is expressed in dB

$$\mathrm{E}(c,\tau) = G(c,\omega)\left(10\log_{10}\left(\left|\frac{X_L^2(\omega,\tau) - X_R^2(\omega,\tau)}{X_L^2(\omega,\tau) + X_R^2(\omega,\tau)}\right|\right)\right). \tag{5.9}$$

   This feature set is also filtered with a 64-channel Mel-filterbank.

Finally, the proposed complementary feature set is combined with ILD and IPD and the total size of the resulting feature set is 320. The histogram of each of the complementary features sets are visualized for two clean signal coming from the front side and the left side in anechoic room simulation on Fig.5.3, Fig.5.4 and Fig.5.5, where the post processing with Mel-filterbank was excluded.

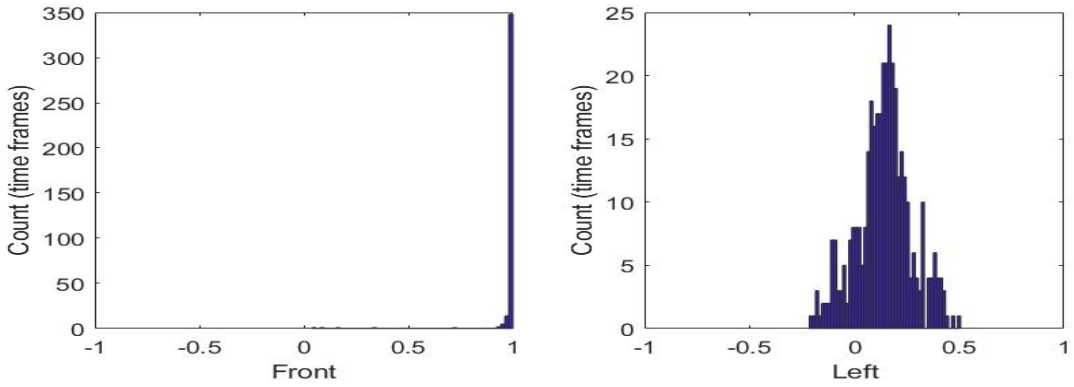


Figure 5.3: Histogram of IC_p feature on low frequency channel for two clean speech signals where one is coming from the front side and the other is coming from the left side.

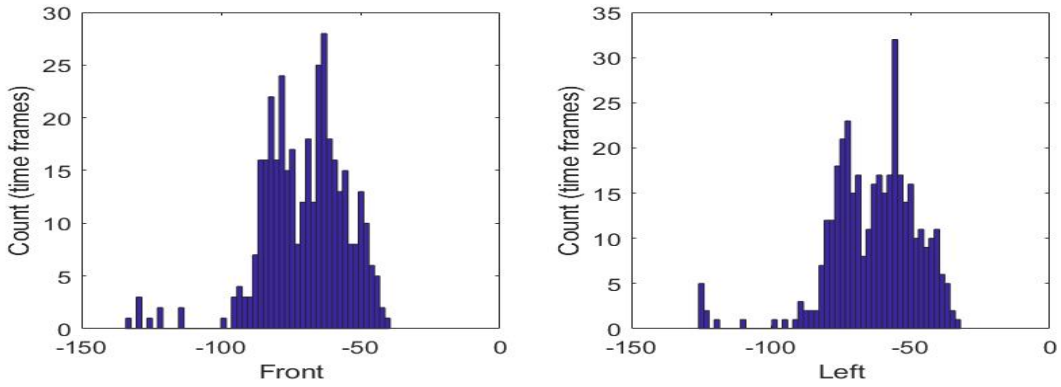


Figure 5.4: Histogram of D feature on high frequency channel for two clean speech signals where one is coming from the front side and the other is coming from the left side.

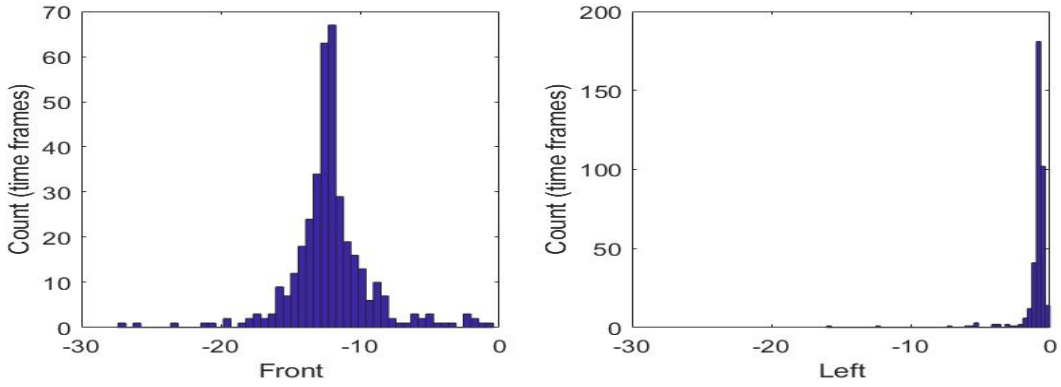


Figure 5.5: Histogram of E feature on high frequency channel for two clean speech signals where one is coming from the front side and the other is coming from the left side.

5.2.3 Neural network architecture

The neural network used in this study is a simple MLP that consists of 3 hidden layers with 1024 hidden units in each layer. The activation function for hidden layers is the ELU [119] while the sigmoid activation function is used for the output layer. Mean squared error (MSE) was selected as the cost function and the Adam algorithm [117] for backpropagation optimization. A hidden-layer dropout rate 0.2 was used in order to avoid overfitting [134] and the number of training epochs is 50. The mini-batch gradient descent with the size of 1024 is used for training. All features are normalized to have zero mean and unit variance on each frequency channel before DNN processing, and batch normalization is applied on hidden layers for faster training [123].

The DNN learns the relationship between training data and its associated IRM [3] for separating speech from noise. The IRM has been previously shown to have better performance than the ideal binary mask (IBM) [3]; therefore, we apply the IRM as the training target on each T-F unit as follows:

$$\mathrm{IRM}(\omega,\tau) = \sqrt{\frac{s^2(\omega,\tau)}{s^2(\omega,\tau)+n^2(\omega,\tau)}} \tag{5.10}$$

where $s^2(\omega,\tau)$ and $n^2(\omega,\tau)$ are the energy of speech and diffuse noise in each time-frequency unit, respectively, based on the STFT decomposition.

5.3 Experimental setup

5.3.1 Training dataset

Speech material from IEEE corpus [92], which includes 720 utterances spoken by a male speaker, was used for both training and test datasets. For training, 400 utterances were selected while 50 sentences were used for testing. Binaural mixtures were generated by convolving the clean speech signal with different HRIRs and BRIRs for anechoic and reverberant condition, respectively. Three sets of impulse responses (IRs) (Oldenburg [78], CATT [139] and Berlin [148]) were used for generating binaural training data.

1. The Oldenburg database includes both anechoic and reverberant IRs with T_{60} = 0.3 s, where the distance between the sound source and receiver is 0.8 m and 3 m for anechoic IRs and 1 meter for reverberant IRs.

2. The Berlin database has only HRIRs (anechoic IRs) with 4 different distances of 0.5, 1, 2, 3 meters between the sound source and receiver.

3. CATT database includes both anechoic and reverberant IRs with T_{60} between the range of 0.1 s to 1 s in 100 ms steps. Therefore, a total number of 10 reverberation times and 1 anechoic condition IRs were used, where the distance between the sound source and receiver is 1.5 meter.

The azimuth range for all IRs is between -90° to 90° with 5° steps; therefore, the total number of 37 azimuth directions were used from each three IRs databases.

A small set of noisy-reverberant mixtures (Total=4000; training=3600, development=400) was used for training the neural networks in this study. The training and development dataset were generated first by convolving the clean speech signal with a randomly selected IR from the three datasets of Oldenburg, Berlin and CATT (37 azimuth directions and different reverberation times), followed by adding a diffuse babble noise. A random signal-to-noise ratio in the range of [-5,5] dB was used for the training/development set and was based on the reverberant target speech instead of anechoic speech.

5.3.2 Testing dataset

5.3.2.1 Unmatched room simulation

Two test datasets were generated by using BRIRs from the Surrey [138] and Aachen databases [149], separately. The Surrey database includes IRs from 4 different rooms (A,B,C,D) with T_{60} times of 0.32 s, 0.47 s, 0.68 s and 0.89 s, respectively and 1.5 m distance between sound source and receiver. The Aachen database includes two BRIRs (Stairway and Aula) with T_{60} times of 0.9 s and 4.82 s, respectively and the distance of 3 meters between the receiver and the sound source.

The azimuth range for Surrey dataset is $[-90^{\circ}, 90^{\circ}]$ with 5° steps (37 azimuth directions) while Aachen database has 5 azimuth directions from -90° to 90° with 45° steps. It should be noted that the IRs used for training and testing are different; therefore, it is called an unmatched condition

1. Test set 1: A total of 750 reverberant binaural mixtures were generated by convolving 50 utterances with randomly selecting Surrey BRIRs from 4 different rooms and 37 directions. The diffuse multitalker babble was added to the reverberant speech with different SNRs of -5, 0 and 5 dB.

2. Test set 2: A total of 750 reverberant binaural mixtures were generated by convolving 50 utterances with randomly selecting Stairway and Aula IRs from Aachen database and 5 different azimuth directions. The diffuse multitalker babble was added to the reverberant speech with different SNRs of -5, 0 and 5 dB.

A diffuse multitalker babble [145] generated from TIMIT corpus [140] was used as the background noise for both training and testing data. Specifically, the background noise was added to the reverberant target speech to create noisy-reverberant binaural mixtures. No similar noise segment or speech utterances are included in both training and testing data sets. Both the speech utterances and IRs were resampled to 16 KHz before generating noisy-reverberant binaural mixtures. The results of this experiment were reported on Table 5.1.

5.3.2.2 Matched room simulation

The performance of the different binaural features and the proposed feature was examined in a matched room simulation, where the IRs for both training and testing stage are Oldenburg, Berlin and CATT. The model was trained with the training dataset explained in Section 5.3.1. However, a new test dataset was generated using 50 utterances from IEEE speech corpus that includes the total number of 750 reverberant and noisy mixtures. Both anechoic and reverberant IRs from 3 databases (Oldenburg, Berlin and CATT) were randomly chosen and convolved with the clean speech. Then the multi-talker babble noise was added as a background noise to the reverberant

mixtures. Both anechoic and reverberant room simulation were tested in this experiment. The SNRs were in three levels of -5, 0 and 5 dB. The performance of the proposed feature set was reported on Table 5.2.

5.3.2.3 Unmatched noise and unmatched IRs

The performance of the proposed binaural feature was also examined in a total unmatched condition that both the background noise and IRs used in testing stage are completely different from training stage. One set of noisy reverberant mixtures (total number of 750 mixtures) was generated by using the 50 utterances convolved with Surrey and Aachen IRs with different reverberation times. Furthermore, the background noise of Factory and Engine were chosen from NOISEX database [135] for unmatched noises. Table 5.3 reports the results of different binaural features and our feature set.

5.3.3 Evaluation criteria

The speech separation performance was evaluated with short-time objective intelligibility (STOI) [142], perceptual evaluation of speech quality (PESQ) [143] and, frequency-weighted segmental signal to noise ratio (SNR_{fw}) [144]. STOI calculates the correlation between temporal envelopes of clean and processed signals and results in a score ranging from 0 to 1 with higher score indicating greater objective speech intelligibility [142]. The PESQ was used as an objective measure of speech quality of processed speech relative to clean speech, that results in a score between [-0.5,4.5] where higher value means greater speech quality [143]. Finally, SNR_{fw} was used to compute the weighted signal-to-noise ratio in each critical band [144]. The score for all 3 criteria is reported for unprocessed noisy-reverberant binaural mixtures and different binaural features and the scores were averaged over the two channels.

5.4 Results

5.4.1 Unmatched room simulation

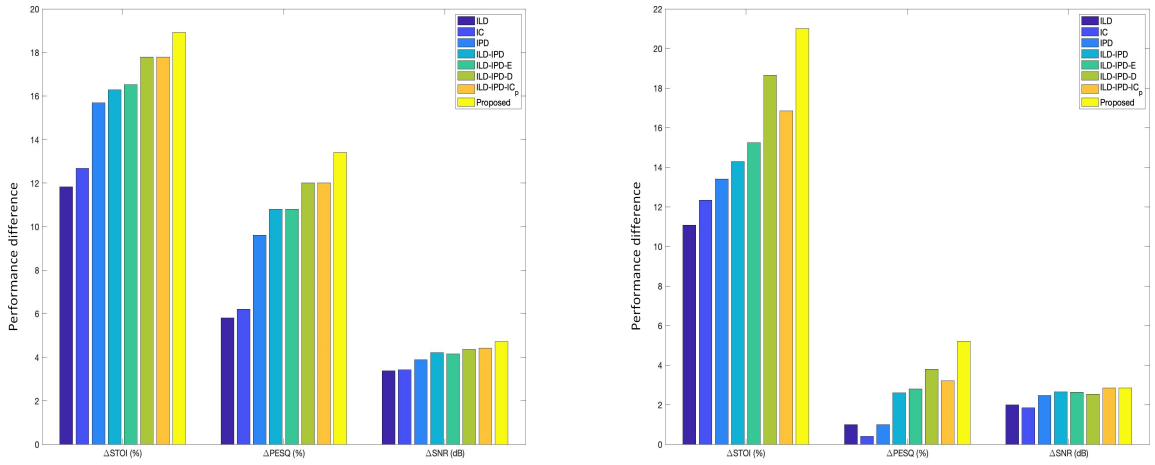
The performance of our DNN-based speech separation algorithm is shown in Table 5.1 for combinations of tested binaural features for unmatched room simulations in training and testing stage. The experimental results reflect greatest increase in all criteria scores for the IPD feature compared against all other individual features averaged over all processed signals of all room simulations. IPD improves STOI, PESQ and SNR_{fw} by the average of 15%, 0.34 and 3.41 dB compared to unprocessed mixtures. Furthermore, the addition of one complementary feature of either E, D or IC_p to the combination of conventional spatial features (ILD-IPD) slightly increases STOI, PESQ and SNR, with the D feature leading to the greatest increase compared to the other two complementary features in STOI and PESQ for all room simulations.

Furthermore, the full feature set including ILD, IPD, IC_p, E and D (ILD-IPD-IC_p-E-D) improves STOI by 19.03%, PESQ by 13.40% and SNR_{fw} by 4.72 dB compared to unprocessed noisy-reverberant mixtures for rooms of Surrey database, while the increase in STOI, PESQ and SNR_{fw} are respectively 21.01%, 5.20% and 2.85 dB for Aachen room simulation. Therefore, the proposed complementary feature set increases significantly the speech intelligibility in a very reverberant room, while it has more effect on improving speech quality in a less reverberant room than speech intelligibility. For instance, in highly reverberant room (Aula room in Aachen dataset), the proposed feature set improves the STOI score by 21.64 % compared to unprocessed mixtures and it is 8.9% better than ILD-IPD feature set. The spectrogram on processed signals with different feature sets are shown on Fig 5.7.

In conclusion, the combination of different feature sets significantly improve the STOI and PESQ scores in moderate reverberant rooms, while their ability in increasing SNR and speech quality is limited in highly reverberant rooms (such as Aachen room simulation). The improvements for different criteria and different room conditions are illustrated in detail on Fig.5.6.

Table 5.1: Results of STOI (%), PESQ and SNR_{fw} (dB) averaged over 750 mixtures for each testing set where ILD-IPD-IC_p-E-D represents our proposed feature set including 5 features of interaural level difference (ILD), interaural phase difference (IPD), partial interaural coherence (IC_p), normalized-energy (E) and energy difference (D).

Metric	Feature	Surrey				Aachen		Mean
		A	B	C	D	Stairway	Aula	
STOI (%)	unprocessed	62.05	60.64	60.55	48.58	48.26	40.82	53.48
	ILD	74.32	72.72	76.38	56.12	60.35	50.90	65.13
	IC	75.56	73.77	76.92	56.68	59.79	53.96	66.11
	IPD	78.29	76.41	80.53	59.78	62.28	53.61	68.48
	ILD-IPD	78.79	77.18	80.97	60.44	64.09	53.56	69.17
	ILD-IPD-E	79.16	77.62	81.11	60.43	64.43	55.11	69.64
	ILD-IPD-D	80.05	79.00	82.12	62.19	67.33	59.04	71.62
	ILD-IPD-IC_p	80.26	78.98	82.25	61.91	64.89	57.87	71.02
	ILD-IPD-IC_p-E-D	**81.21**	**80.30**	**83.71**	**62.71**	**68.64**	**62.46**	**73.17**
PESQ	Unprocessed	1.65	1.65	1.56	1.66	1.55	1.43	1.58
	ILD	2.09	1.94	1.84	1.82	1.61	1.48	1.79
	IC	2.13	1.99	1.84	1.82	1.57	1.46	1.80
	IPD	2.29	2.14	2.06	1.98	1.61	1.48	1.92
	ILD-IPD	2.34	2.19	2.11	2.05	1.73	1.52	1.99
	ILD-IPD-E	2.34	2.21	2.11	2.04	1.73	1.53	1.99
	ILD-IPD-D	2.41	2.29	2.14	2.10	1.78	1.58	2.05
	ILD-IPD-IC_p	2.40	2.28	2.15	2.10	1.75	1.56	2.04
	ILD-IPD-IC_p-E-D	**2.49**	**2.37**	**2.22**	**2.15**	**1.83**	**1.68**	**2.12**
SNR_{fw} (dB)	Unprocessed	2.85	3.17	3.22	2.35	2.28	1.94	2.63
	ILD	6.90	6.63	6.87	4.67	4.36	3.86	5.54
	IC	6.94	6.72	6.88	4.70	4.11	3.82	5.52
	IPD	7.28	7.20	7.60	5.03	4.71	4.43	6.04
	ILD-IPD	7.58	7.54	7.91	5.37	5.01	4.50	6.31
	ILD-IPD-E	7.58	7.57	7.74	5.29	4.99	4.50	6.27
	ILD-IPD-D	7.75	7.79	7.95	5.53	4.87	4.44	6.38
	ILD-IPD-IC_p	7.80	7.86	8.05	5.53	5.12	4.79	6.52
	ILD-IPD-IC_p-E-D	**8.14**	**8.23**	**8.40**	**5.69**	**5.21**	**4.72**	**6.73**



(a) Surrey IRs room simulation

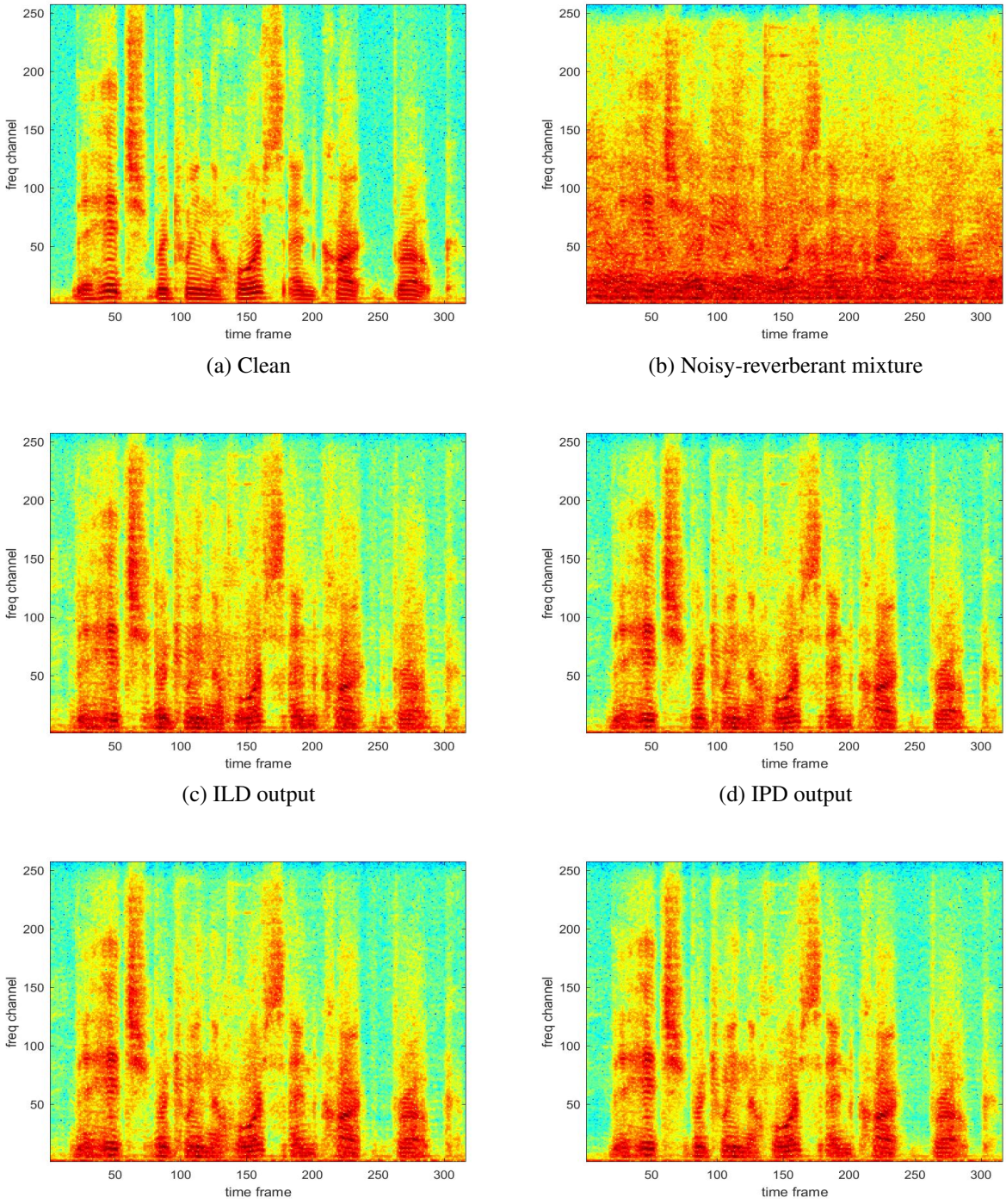
(b) Aachen IRs room simulation

Figure 5.6: The improvement in the performance of binaural speech separation model in terms of STOI, PESQ and SNR$_{fw}$ for different binaural feature sets in different room simulations.

5.4.2 Matched room simulation

In addition, the performance of the proposed feature set was examined for different matched rooms simulation. In the matched room condition, similar IRs (Berlin, Oldenburg and CATT) were used to simulate same rooms in generating training and test binaural mixtures. Table 5.2 demonstrates the improvements in STOI, PESQ and SNR$_{fw}$ for different room simulations (3 different sets of IRs including anechoic and reverberant conditions). On average over all matched rooms, the complementary feature set increase STOI by 22.44%, PESQ by 18.4% and SNR by 6.33 dB. It can be seen from the Fig. 5.8 that the proposed feature STOI score is 3.20 % higher than the ILD-IPD feature and 4.5 % higher than IPD feature. Therefore, for both seen rooms and unseen rooms, the proposed feature set improves the performance of the model significantly and this feature generalizes well to seen and unseen room simulations.

The increase of STOI in the anechoic (Ane) condition for matched rooms is 23.27%, while it is 21.23 % for reverberant (Rev) matched rooms. Therefore, the improvement in STOI is in the same range for both anechoic and reverberant condition of matched room simulation.



(a) Clean

(b) Noisy-reverberant mixture

(c) ILD output

(d) IPD output

(e) ILD-IPD output

(f) Proposed output

Figure 5.7: Visualization of processed signals using different binaural feature set for training DNN model for a IEEE clean speech mixed with babble noise at 0 dB SNR and RT_{60} of 0.68 second.

Table 5.2: The STOI (%), PESQ and SNR_{fw} (dB), averaged over 750 mixtures for each testing set including matching rooms condition where the proposed feature is ILD-IPD-IC_p-E-D.

Feature	STOI (%)					PESQ					SNR_{fw} (dB)				
	Berlin	Oldenburg		CATT		Berlin	Oldenburg		CATT		Berlin	Oldenburg		CATT	
	Ane	Ane	Rev	Ane	Rev	Ane	Ane	Rev	Ane	Rev	Ane	Ane	Rev	Ane	Rev
Unprocessed	67.54	63.79	61.94	67.74	54.90	1.61	1.53	1.51	1.65	1.65	3.74	3.39	2.97	3.35	2.30
ILD	84.70	82.00	76.42	83.68	68.43	2.25	2.17	2.08	2.12	1.93	8.68	8.57	6.98	7.50	5.62
IPD	88.37	85.04	78.05	87.41	67.88	2.52	2.42	2.21	2.42	2.01	9.70	9.53	7.40	8.69	5.94
ILD-IPD	88.63	85.45	79.26	88.07	70.73	2.54	2.45	2.27	2.45	2.06	9.79	9.63	7.70	8.83	6.20
Proposed	**91.02**	**88.15**	**83.01**	**89.70**	**76.26**	**2.69**	**2.58**	**2.46**	**2.57**	**2.25**	**10.97**	**10.69**	**8.81**	**9.63**	**7.33**

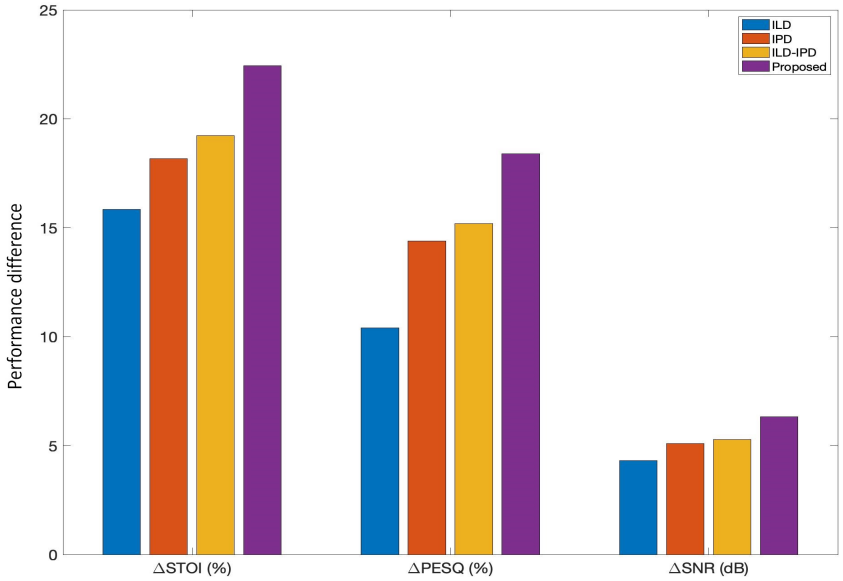


Figure 5.8: The improvement in the performance of DNN-based binaural speech separation model in terms of STOI, PESQ and SNR_{fw} for different binaural feature sets in matched room simulations.

5.4.3 Unmatched room simulation and unmatched noise

The trained model explained in Section 5.3.2.1 was used to estimate the time frequency mask for noisy reverberant binaural mixtures, when both the background noise and room simulation were unmatched to training condition. Table 5.3 and Figure 5.9 illustrate the results for a total mismatch condition for two different background noise. For Engine background noise, the improvements in STOI, PESQ and SNR_{fw} are 18.34 %, 12 % and 4.31 dB respectively using proposed feature

Table 5.3: The STOI (%), PESQ and SNR$_{fw}$ (dB), averaged over 750 mixtures for each testing set including unmatched rooms and unmatched noises where the proposed feature is ILD-IPD-IC$_{\text{p}}$-E-D.

	STOI (%)				PESQ				SNR$_{fw}$ (dB)			
	Surrey		Aachen		Surrey		Aachen		Surrey		Aachen	
	Engine	Factory	Engine	Factory	Engine	Factory	Engine	Factory	Engine	Factory	Engine	Factory
Unprocessed	61.97	61.44	47.75	44.35	1.49	1.26	1.48	1.09	2.33	2.41	1.31	1.66
ILD	72.68	73.57	55.25	53.71	1.91	1.87	1.72	1.46	6.23	6.44	3.64	4.41
IPD	77.01	77.90	57.77	56.58	2.12	2.10	1.80	1.49	6.85	7.10	4.16	4.88
ILD-IPD	77.09	78.23	57.85	56.60	2.15	2.15	1.73	1.51	6.99	7.31	4.17	4.89
Proposed	**80.29**	**81.18**	**66.11**	**64.72**	**2.34**	**2.33**	**1.82**	**1.74**	**7.70**	**7.79**	**4.56**	**4.82**

set, while they are 21.06 %, 17.2 % and 4.27 dB for Factory noise. Comparing these results with Table 5.1 shows that binaural features have better performance when the background noise is more diffuse and not similar to speech.

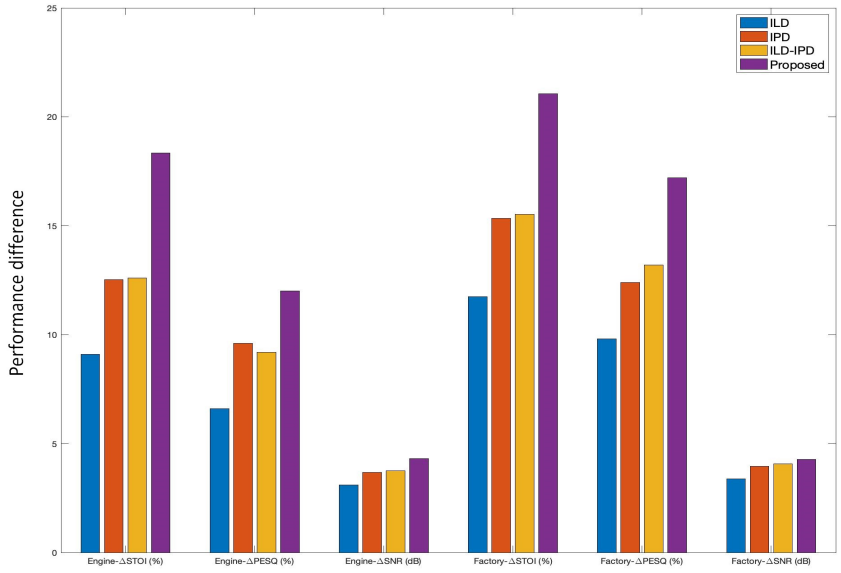


Figure 5.9: The improvement in the performance of DNN-based binaural speech separation model in terms of STOI, PESQ and SNR$_{fw}$ for different binaural feature sets in unmatched noises of Factory and Engine and room simulations.

5.5 Discussion

This chapter discussed in detail the binaural features for supervised speech separation. A feature study was done to evaluate the performance of the binaural speech separation model in different room simulations, various SNRs and background noises. Moreover, a novel complementary binaural spectral feature set was proposed for a DNN masking-based speech separation in noisy-reverberant conditions. The evaluation of speech separation performance shows that the proposed complementary feature outperforms common binaural features of ILD and IPD in terms of STOI and PESQ and SNR, suggesting a greater preservation of the target speech acoustics in noisy reverberant environment. Furthermore, this novel feature is robust to different room simulation and background noises. Using the information about the distribution of energy difference between the left and the right ear signals and normalized energy difference with partial coherence, the neural network has better ability to track the target speech in noisy-reverberant environment. Furthermore, combining these new features with common binaural features increase the performance of the network to predict better time frequency masks for speech separation. For future work, the performance of other types of neural networks with this feature set should be investigated and extend the work for a speaker separation task.

Chapter 6

Noise and speaker independent model for monaural speech enhancement

This chapter focuses on the generalization issue of supervised speech enhancement algorithms. The current solutions for this problem will be explained and finally two novel neural network architectures will be proposed to improve the performance of supervised models in unseen conditions. Furthermore, the performance of our models were evaluated in a variety of simulated environments having different reverberation times and signal to noise ratios and quantified using two objective measures of speech intelligibility and speech quality. Moreover, both the mapping based models and masking based models were examined and the performance of our models is compared with the current supervised speech enhancement models in terms of speech intelligibility and speech quality improvement and computational efficiency.

6.1 Introduction

For any supervised learning problem, the generalization to unseen conditions is a very challenging task. For supervised speech separation, the same issue of generalization exists. In supervised speech separation methods, there are three different aspects in generalization to unseen conditions that should be considered such as unseen noises, unseen speakers and unseen SNRs. This chapter will focus on these challenging issues and provide two models to solve generalization problems. Furthermore, the advantages of using convolutional neural network and recurrent neural networks in speech enhancement will be mentioned that helped us to tackle these problems. In speech enhancement based on data-driven approaches, it is important to generalize the model to different

speakers, noise types, SNRs and rooms. For SNR and noise generalization, the simple solution is to train the model with different SNRs and different noise types [4, 150].

For SNR generalization, including different SNR levels in training datasets has revealed that supervised speech enhancement is not sensitive to SNR levels. The main reason for it is that the local SNRs at the frame level and T-F unit have a wide range which provides the large variability needed for training to solve generalization problem [70].

For noise generalization, stationary and non-stationary noise may interfere with speech. In two studies [150, 4], Chen proposed that training the network with large scale training dataset with wide-variety of noises is the solution for noise generalization which improved speech intelligibility for both normal hearing and hearing impaired listeners in unseen noises. In [150], Chen showed that having enough training noises with a single speaker, the DNN has good generalization for unseen noises and the speaker-dependent DNN improves significantly the STOI scores for the seen speaker [150].

On the other hand, the speaker-dependent DNN is incapable of separating new speaker from background noise [150]. Moreover, increasing the number of speakers in training stage does not help the generalization to different speakers for a DNN based separation model. Furthermore, Chens' experiments showed that by increasing the number of training speakers, the performance of the speaker-dependent DNN degrades [150]. Specifically, when the background noise includes speech components (such as babble noise), the speaker-dependent DNN mistakes these interfering babble noise as the target speech [150].

For speaker generalization, prior work showed that training feedforward DNNs with different speakers did not achieve good performance because of the limited capacity of DNN in modeling the number of speakers [150]. A context window was proposed to provide temporal contexts for feedforward DNNs. However, when a DNN is exposed to large number of speakers, it mistakes noise segments as speech segments especially when the background noise is babble. The acoustic features presented as the input of DNN has limited temporal content; therefore, it is not sufficient to focus on the target speaker because of noise and target speaker fluctuations over time [4].

Chen suggested using sequence to sequence mapping as speech enhancement method to capture long-term context of signals. He proposed using a well known type of recurrent neural network (LSTM) for speaker generalization [150, 4]. The recurrent network is capable of modeling temporal dynamics of speech where it uses the long-term context to focus on target speaker. In [4], the author demonstrates that LSTM outperforms DNN when the network is trained on multiple speakers. Moreover, it has significant performance in comparison to DNN when tested on untrained speakers.

In recent years, convolutional neural networks (CNN), recurrent neural networks (RNN), and different combinations of two have been used for many applications such as noise and speaker independent speech enhancement and speech separation models. Since the time-frequency representation of speech signals is like an image including information in both time axis and frequency axis, many popular convolutional neural network architectures have been examined for speech enhancement and speech separation.

Two studies of gated residual networks (GRN) [151] and convolutional recurrent neural networks (CRN) with encoder-decoder structure [152] were proposed for a noise and speaker independent speech enhancement model, which resulted in greater generalization for untrained speakers and fewer trainable parameters compared to LSTM. However, most of the proposed methods were only trained and tested in noisy environments with no reverberation and they do not consider the acoustics of room in their models. Furthermore, GRN and CRN have large number of training parameters compared to the proposed networks in this dissertation, which increase the training time and computational cost that are not suitable for real-time speech enhancement.

Inspired by the success of CNNs in image processing and LSTM in temporal modeling, two novel neural network architectures are proposed in this chapter to improve the generalization of supervised speech enhancement models to unseen conditions. The first model incorporates dilated convolutions for tracking a target speaker through context aggregations, skip connections, and residual learning for both masking-based and mapping-based monaural speech enhancement, while the second proposed model, modifies the first model to operate in real-time processing with

LSTM layer and causal convolution layers to increase the capability of the network to track the target speaker in noisy-reverberant environments when the background noise is non-stationary and similar to speech. Furthermore, the number of training parameters in second model is smaller than the first model which makes the model computationally more efficient which can also be fitted for real-time applications.

In other words, the proposed models take advantage of the properties of dilated convolution [153] and recurrent network to create speaker independent and noise independent models that have limited number of training parameters compared to previous studies, which makes them computationally more efficient. The performance of our models were evaluated in a variety of simulated environments having different reverberation times and quantified using two objective measures. In the following sections, both the LSTM model and convolutional neural network (CNN) and dilated convolutions are explained in more detail. Moreover, the proposed two models are explained and examined for different conditions.

6.2 Long short-term memory (LSTM)

RNN is proposed to model temporal dependencies by training with back propagation through time (BPTT). However, vanilla RNN suffers from vanishing and exploding gradient through BPTT [111]. LSTM has been proposed to solve this issue by using memory cells and it has been used in many applications such as language modeling and acoustic modeling. LSTM has a memory cell and three gates of input gate, output gate and forget gate [112]. The input gate controls the information added to the cell, while the forget gate controls the amount of previous information to be deleted from the cell. These three gates are restricted to the range of [0-1] with the sigmoid activation function, while the output of the network is limited to [-1,1] by both sigmoid and tanh activation functions. Both the input gate and the forget gate are dependent on the input and previous hidden activation which make the memory cell sensitive to context. Therefore, the LSTM based supervised speech separation is sensitive to speaker information extracted from previous frames [4].

LSTM equations are as follows

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{6.1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{6.2}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{6.3}$$

$$z_t = \tanh(W_{zx}x_t + W_{zh}h_{t-1} + b_z) \tag{6.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot z_t \tag{6.5}$$

$$h_t = o_t \odot \tanh(c_t)\,, \tag{6.6}$$

where x_t, z_t, c_t, h_t are input, block input, memory cell and hidden activation at time t. Input gate, forget gate and output gate are represented by i_t, f_t, o_t. The operator $\odot$ denotes the element-wise multiplication and W, b are the linear transforms and biases, respectively [150]. Moreover, σ is the sigmoid activation function and g represents *tanh* activation function as

$$\sigma(s) = \frac{1}{1+e^{-s}} \tag{6.7}$$

$$g(s) = \frac{e^{s} - e^{-s}}{e^{s} + e^{-s}}\,. \tag{6.8}$$

6.3 Convolutional neural network (CNN)

Convolutional neural networks (CNNs) are very useful for machine learning projects and specifically in image processing and computer vision. However, there are some differences between the fully-connected layers as in DNN and convolutional layers as in CNN. CNN output shape is affected by the shape of the input, the kernel size, zero padding and strides, while in fully connected layers, the output shape is independent of the input shape. Furthermore, the CNNs can have pooling layers, which add more complexity in comparison to neural network fully connected layers.

CNNs have been used in image processing and speech recognition with images and sound clips usually stored as multi-dimensional arrays and represented by one or more axes (time frames and frequency bins) where the ordering of these axes are important. Therefore, convolutional layers can be useful for processing images and sounds.

The discrete convolution is a type of linear transformation that can preserve the ordering of the input data [154]. Moreover, convolution is sparse, which means that the output units of convolution is affected only by few input units. Furthermore, the convolution shares the parameters, which means that the same weights are applied to different locations in the input [154].

The process of convolution is as follows: The kernel slides across the input feature map and at each point (location), the overlapped input element is multiplied with the kernel element and the results are summed up to give the value for the output in that specific location called output feature maps [154]. On the other hand, there is a limitation in the resulting output feature maps of CNN layers. These feature maps memorize the precise location of features in the input. Therefore, if there are small changes in the input such as rotation, shifting or cropping, the feature maps will be completely different. Downsampling is a solution for this problem that can be done either by using strides or pooling layers in CNNs.

Another important capability of CNNs are the pooling layers. The pooling layers reduce the size of the feature maps by summarizing the sub-regions with either taking the average or the maximum value of those regions [154], and provide invariance to small translations of the input. For instance, in max pooling layers, the input is split to non-overlapping patches and then the maximum value of each patch is selected as the output. The average (mean) pooling layer takes the mean value of each patch and selects it as the output [155]. Selecting the maximum value of different features instead of their average is more informative because the features of the input are more likely encoding the spatial presence of some patterns over different tiles of the feature map [155].

6.3.1 Dilated convolution

Many image classification networks employ pooling and subsampling layers to integrate multi-scale contextual information. On the other hand, the dilated convolution procedure was proposed to aggregate multi-scale contextual information without losing resolution (like pooling layers) or analyzing rescaled images [153]. Dilated Convolution (Atrous convolution) was originally proposed for wavelet decomposition. In deep convolutional neural networks, the idea of dilation is to insert zeros between pixels in convolutional kernels to increase image resolution and receptive field, which leads to dense feature extraction. In one study [153], dilated convolution was proposed for semantic segmentation, which is a dense prediction problem. Semantic segmentation requires pixel-level accuracy with multi-scale contextual reasoning. Dilated convolution networks is designed for dense prediction and it uses exponential expansion of the receptive field (with no pooling or upsampling layers) and with no loss in resolution [153]. This approach employs serialized convolutional layers with increasing dilation rate to aggregate contextual information. A 2-D discrete convolution operator * convolves the F signal (discrete function) with the kernel (k) of size $(2m+1)\times(2m+1)$ as follows

$$(F*k)(p)=\sum_{s+t=p}F(s)k(t)\,, \tag{6.9}$$

where p,s $\in\mathbb{Z}^2$ and t $\in[-m,m]^2\cap\mathbb{Z}^2$ where $\mathbb{Z}$ is the set of integers [153]. The dilated convolution $*_l$ can be defined as

$$(F*_l k)(p)=\sum_{s+lt=p}F(s)k(t)\,, \tag{6.10}$$

where l is the dilation rate. The conventional convolution is the dilated convolution with the dilation rate of one. Therefore, the 1-D l-dilated convolution is defined similarly as the 2-D l-dilated convolution with p,s $\in\mathbb{Z}$ and t $\in[-m,m]\cap\mathbb{Z}$. The dilated convolutions expand receptive fields exponentially by increasing dilation rate without any resolution loss or convergence [153]. Having discrete functions of $F_0,F_1,...,F_{n-1}$ and kernels of $k_0,k_1,...,k_{n-2}$ with the size of 3×3 and applying

the filters with exponentially increasing dilation rate

$$F_{i+1} = F_{i*2^i} \, k_i \quad for \quad i = 0, 1, ..., n-2 \,, \tag{6.11}$$

the size of receptive field of each element p in F_{i+1} is $(2^{i+1}-1) \times (2^{i+1}-1)$.

6.4 Monaural speech enhancement with dilated convolution

In this section, a novel neural network structure will be presented for monaural speech enhancement that aims to separate target speakers from background noise and reverberation using stacked dilated convolutions [156]. By expanding the receptive field in convolutional neural network, the contextual information is augmented. There are three methods to increase the receptive field in CNNs. Increasing the network depth leads to expansion of receptive field; however, this method is not computationally efficient and deep neural networks have the vanishing gradient issue. The second method is to enlarge the kernel size, however that results in higher computations and more training time. The third method is using dilated convolutions, which has been proposed for multi-scale aggregation in image segmentation to expand the receptive field exponentially without losing any resolution [153].

Inspired by the dilated convolutions method [153], which aggregates multi-scale contextual information without losing resolution, a new fully-convolutional neural network is developed that utilizes 2D convolutional networks with pooling layers for feature extraction. Furthermore, the network includes two blocks of stacked dilated 1D convolutions that expand the receptive field (augment the contextual information) and generate a smooth mask on lower level features. Moreover, skip connections are used to incorporate different level features in final estimation of speech in noisy-reverberant environment. The results of our analyses show that the proposed dilated convolutional network has better performance and lower computational cost compared to LSTM and previous convolutional networks for recovering speech from noisy-reverberant environment. Furthermore, the proposed model generalizes well to different unseen conditions.

6.4.1 System description

The goal of our proposed monaural speech enhancement model is to estimate the target speaker from a single microphone noisy-reverberant mixture. Supervised speech enhancement can be treated as a regression problem that maps acoustic features from noisy-reverberant mixtures to a training target, which can be either time frequency mask or a spectral representation of a target speaker. The acoustic features and training target are passed to the neural network for training and in the testing stage the estimated output and the phase of noisy-reverberant mixture are then resynthesized into a time-domain speech waveform.

In this study, the noisy-reverberant mixtures were generated by convolving the target speaker audio with an impulse response (IR) for different room locations and reverberations, then mixed with different types of non-stationary background noises [90]. The signal was segmented into time frames with a 20 ms hamming window and 10 ms overlap (50%).The 320-point short-time Fourier transform (STFT) was used to calculate the magnitude spectra, which was then used as the input acoustic features with the dimension of 161.

6.4.2 Proposed dilated convolution network

Dilated convolutions were developed for semantic segmentation that aggregates multi-scale contextual information and supports exponential expansion of the receptive field without loss of resolution [153]. The 1-dilated convolution is the conventional convolution and its receptive field scale increases linearly with the layer depth in the CNN. However, the scale of receptive field in dilated convolution increases exponentially with the depth of the layer when the kernels are stacked with exponentially increasing dilation rates [153].

In this work, four layers of 2D-convolution were first stacked with exponential linear unit (ELU) activation function [119], batch normalization (BN) [123], and max pooling layer with the size of (1,2) for feature extraction to capture contextual information in both time and frequency domains. It should be noted that the max pooling layers are only applied on the frequency axis and the time axis is preserved so that the input and the output of the network have the same size on the

time axis. The kernel size are 5×5 (50 ms and 250 Hz) and 9×9 and zero padding is used on each layer. The learned features from 2D-convolution blocks are then reshaped and given as an input for 1D-convolution layers.

Our model includes two residual blocks that perform 1D-convolution with a kernel size of 3 and 256 output channels and 1D dilated convolutions with exponentially increasing dilation factors as 2,4,8,16,32,64,128 with the kernel size of 3 and 16 output channels. The dilated convolutions enable our model to deepen the network by increasing the dilation factor instead of increasing the length of the filter which is more efficient and does not reduce the resolution. Stacking a set of layers with dilation to the maximum dilation factor can create context stacks. Moreover, the output of each dilated block is given to another 1D-convolution with 256 output channels to make it the same size as its lower layer (size of 256) and finally passed through a sigmoid activation function that operates as a soft mask. This soft mask is then multiplied with the learned features from previous 1D-convolution layer. Furthermore, skip connections were used to allow the network to use features that are extracted from different hierarchical levels for final prediction [157]. Therefore, three different levels of extracted features are summed and passed through a ReLU function. These skip connections preserve and integrate the knowledge learned by each stacked layer. Finally, there is an output layer to estimate the desired clean signal or the time frequency mask. The activation function on the output layer is rectified linear unit (ReLU) [118] for mapping based speech enhancement models, while sigmoid is used as an activation function for masking based speech enhancement.

The proposed dilated convolutional network (DCN) [156] is shown on Fig. 6.1 and the detailed description of it is given in Table 6.1. The hyperparameters for 2D-convolution layers are shown as (kernel size, (strides), output channel), while the hyperparameters for 1D-convolution are given as (kernel size, dilation rate, output channels) in Table 6.1. Note that the zero-padding is applied to both 2D-convolution and 1D-convolution layers. Our network was compared with four common network architecture of DNN, LSTM, CRN and GRN, which are explained in the following section.

The time axis receptive field for 2D-convolution is $1+(5-1)\times 1+(9-1)\times 1+(5-1)\times 1+(9-1)\times 1=25$. Furthermore, the receptive field for the 1D dilated convolution is calculated as $2\times(2+4+8+16+32+64+128)\times 2=1016$ and $(3-1)\times 1\times 5=10$ for the rest of 1D-convolutions. The total receptive field is 1051 which means that the output layer depends on 1051 time frames of input, which is 10.51 s ($1051*0.01$ s) because of the frame shift of 10 ms in our calculation. Therefore, the network operates on 5.25 s of the past and 5.25 s of the future input. Thus, the proposed dilated network uses a large past and future information to predict the spectrum or the time frequency mask for speech enhancement and the causal model will be evaluated in the next section.

6.4.3 Comparison models

DNN: A simple feedforward neural network is used with 4 hidden layers of 1024 units. The activation layer of each hidden unit is ELU and droput rate of each hidden layer is 0.2. Moreover, batch normalization is applied on each layer before passing the output through the activation functions.

LSTM: The LSTM network includes four hidden layers with the size of 1024 units while the input and output layers have 161 units. The time step for these layers is 128 and the dropout rate of 0.2 is used on each layer. Each hidden layer is also batch normalized. The output layer of this model is a fully-connected layer (Dense layer) with the ReLU or sigmoid activation function for the mapping based speech enhancement and the masking based speech enhancement model, respectively.

CRN: The convolutional recurrent network (CRN) [152] has a total of ten 2D-convolution layers and two LSTM layers. CRN includes an encoder and a decoder with five convolutional layers and five deconvolutional layers, respectively. Moreover, the convolutional layers are causal and have a stride of 2 along frequency. In addition, 2 layers of LSTM are stacked between encoder and decoder to model temporal dynamics of speech and skip connections are used to concatenate the output of each encoder layer to the input of each decoder layer [152].

GRN: Gated residual network (GRN) is a 62 convolutional layer network, constructed with a stack of frequency-dilated convolutional layers and multiple gated residual blocks with different dilation

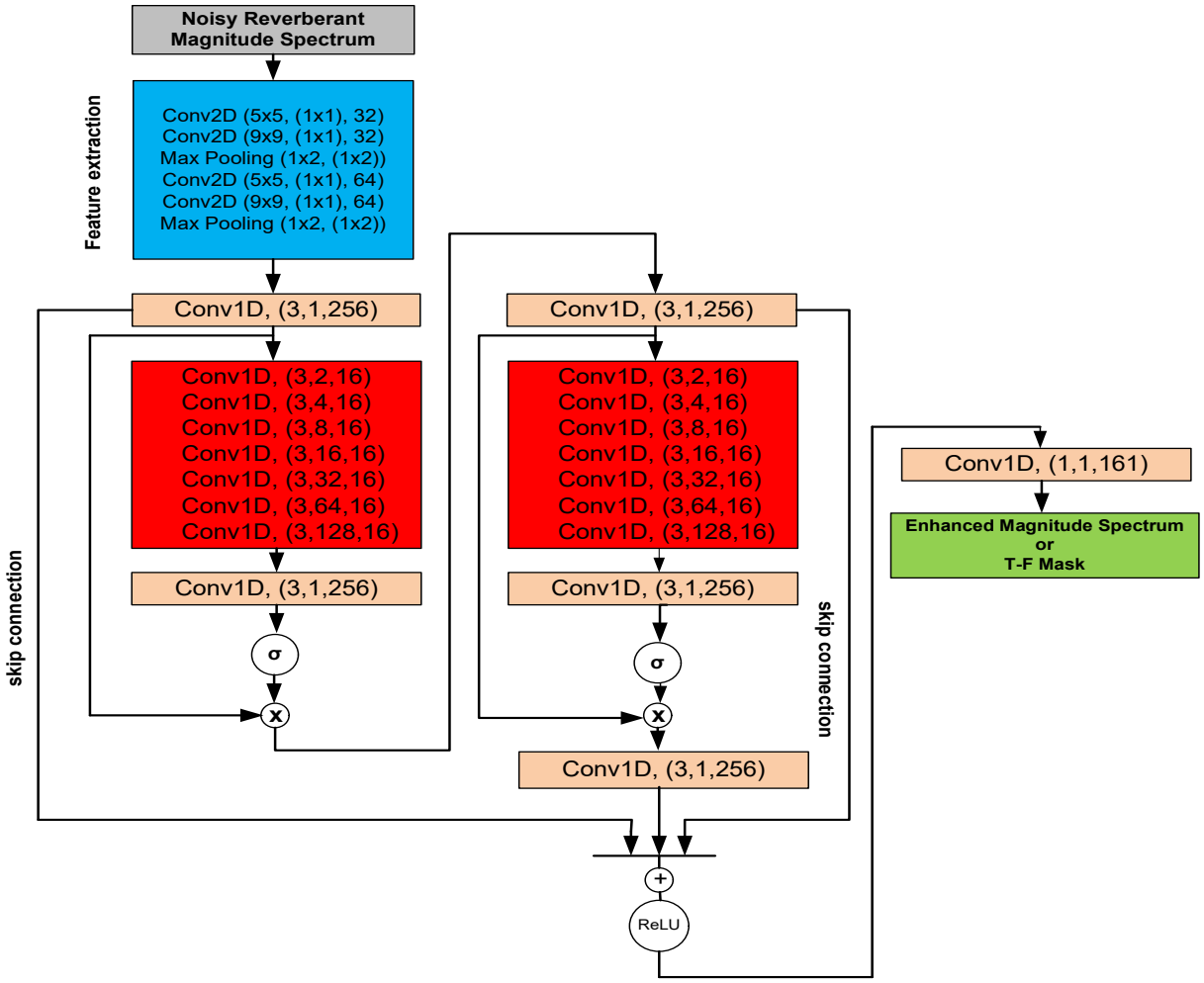


Figure 6.1: Proposed dilated convolutional network (DCN) architecture.

rates [151].

6.4.4 Training targets

In this work, two different training targets of clean speech are examined: ideal ratio mask (IRM) [3] and target magnitude spectrum (TMS) [158]. IRM has been widely used in supervised speech separation, which can interpreted as a smooth version of IBM [2] but with improved performance [3]:

$$\mathrm{IRM}(\omega,\tau) = \sqrt{\frac{s^2(\omega,\tau)}{s^2(\omega,\tau)+n^2(\omega,\tau)}} \tag{6.12}$$

Table 6.1: Architecture of our proposed dilated convolutional network.

Block name	Layer name	hyperparameters
Conv2d+Maxpooling	conv2d_1	$5\times5,(1\times1),32$, BN, ELU
	conv2d_2	$9\times9,(1\times1),32$, BN, ELU
	maxpool_1	$1\times2,(1\times2)$
	conv2d_3	$5\times5,(1\times1),64$, BN, ELU
	conv2d_4	$9\times9,(1\times1),64$, BN, ELU
	maxpool_2	$1\times2,(1\times2)$
Conv1d	conv1d_1	$3,1,256$, BN, ELU
Dilated block1	conv1d_d1	$3,2,16$, ELU
	conv1d_d2	$3,4,16$, ELU
	conv1d_d3	$3,8,16$, ELU
	conv1d_d4	$3,16,16$, ELU
	conv1d_d5	$3,32,16$, ELU
	conv1d_d6	$3,64,16$, ELU
	conv1d_d7	$3,128,16$, ELU
Conv1d	conv1d_2	$3,1,256$, ELU
Conv1d	conv1d_3	$3,1,256$, BN, ELU
Dilated block2	conv1d_d8	$3,2,16$, ELU
	conv1d_d9	$3,4,16$, ELU
	conv1d_d10	$3,8,16$, ELU
	conv1d_d11	$3,16,16$, ELU
	conv1d_d12	$3,32,16$, ELU
	conv1d_d13	$3,64,16$, ELU
	conv1d_d14	$3,128,16$, ELU
Conv1d	conv1d_4	$3,1,256$, ELU
Conv1d	conv1d_5	$3,1,256$, ELU
Conv1d	conv1d_6	$1,1,161$, ReLU/Sigmoid

where $n^2(\omega,\tau)$ and $s^2(\omega,\tau)$ are the energy of noise and speech in each T-F unit, respectively. In masking-based speech separation, the estimated IRM is multiplied with the noisy mixture magnitude spectrum to calculate the magnitude of separated speech. The result is then combined with the phase of the noisy mixture to reconstruct an audio waveform of the processed speech. In mapping-based speech separation, TMS is the training target where the estimated magnitude spectrum is combined with noisy phase to construct the processed speech waveform.

6.4.5 Advantages of the proposed dilated convolution model

In comparison to DNN, our model is able to leverage contextual information using dilated convolution, while DNN can not because of its feedforward connections. Moreover, convolutions can be done in parallel; however, recurrent networks are sequential and for the prediction of next input,

the previous timesteps need to be calculated [159]. In addition, the receptive field in CNN can be increased easily by increasing the depth of the CNN layers, increasing dilation factor, or kernel size. CNNs do not have the exploding or vanishing gradient, while RNNs suffer from this issue. Recurrent networks such as GRU and LSTM have large memory requirements to store the information of their gates, especially if the input sequence is long [159]. Therefore, the required memory for training CNNs are much smaller than recurrent networks. Finally, the number of training parameters of DCN is smaller than LSTM, CRN and GRN, which makes it computationally efficient and the training time is shorter, which is desirable when working with large training datasets. The number of training parameters for each of the different models are shown on Table. 6.2.

Table 6.2: The total number of trainable parameters for each model.

Network type	LSTM	CRN	GRN	**DCN**
Number of parameters (millions)	25.51	17.22	3.11	**2.92**

6.4.6 Experimental setup

6.4.6.1 Dataset

Speech material from the TIMIT corpus [140] were used for training and testing. A total number of 1700 utterances spoken by 462 speakers (male and female) were used for training along with 300 different background noises extracted from the FreeSFX [160] and Freesound [161] corpora. Furthermore, the Babble and Cafeteria noises drawn from the NOISEX [135] and DEMAND [141] repositories were used as unseen noises during testing.

The noisy-reverberant mixtures (Total=34000, training=30600, development=3400) used for training the neural networks in this study were generated first by convolving the clean speech signal with a randomly selected room impulse response (IR). Each single channel IR simulated reverberant room conditions according to the image method technique [136] and was followed by

adding a random selection of the different background noises. The IRs with different reverberation times (T_{60}) in the range of $\{0.3-0.9\}$ s and an anechoic condition with $T_{60}=0.0$ s were created by the room impulse response generator [137]. Room dimension was fixed as $(10\times9\times8)$ m and the microphone was located at $(3,4,1.5)$ m. The target speaker was located at a random position at different distances in the range of 0.5–3 m from the microphone. A random SNR in the range [-5,5] dB was used for the training set and was based on the reverberant target speech instead of anechoic speech.

Two test sets were generated to investigate generalization of speakers and noise in each of the models. Briefly, one set used 6 trained speakers (3 males, 3 females) and the other test set used 6 untrained speakers (3 males, 3 females). The noisy-reverberant test set mixtures were generated by convolving the clean speech signal with real IRs recorded at the University of Surrey [138]. The Surrey database includes IRs recorded from four reverberant rooms with T_{60} of 0.32 s, 0.47 s, 0.68 s and 0.89 s, respectively and the distance between the sound source and the microphone is 1.5 m. These IRs have 2 channels; therefore, when the signal is coming from the left side, the left channel of IR is selected to convolve with speech, while when the signal is coming from the right side, the right channel of IR is selected to convolve with speech. It should be noted that IRs are randomly selected based on their T_{60} and directions for data generation.

- Test set 1: 1000 mixtures were created from 6 trained speakers convolved with random Surrey IRs and mixed with random segment of background noise of Cafeteria and Babble.
- Test set 2: 1000 mixtures were created from 6 untrained speakers convolved with random Surrey IRs and mixed with a random segment of background noise of Cafeteria and Babble.

The SNR for both test sets was set to three levels of -5, 0 and 5 dB. No similar noise segment, speech utterance or IR are included in both training and testing data sets. All noises, speech utterances, and IRs were resampled to 16 KHz. The neural network cost function for all models was chosen as the mean squared error (MSE) and the Adam algorithm [117] was selected for backpropagation optimization. The convolutional models were trained with the mini-batch size of

16 while the batch size in LSTM was 256. The number of training epochs was 50 and all features were normalized to have zero mean and unit variance on each frequency channel before feeding to the networks. Batch normalization was applied on hidden layers for faster training [123].

6.4.6.2 Evaluation criteria

The short-time objective intelligibility (STOI) [142] and perceptual evaluation of speech quality (PESQ) [143] were used to evaluate speech enhancement model performance. Both criteria are objective and do not represent listener performance. STOI is given by the correlation between the temporal envelopes of clean and processed signals, resulting in a score ranging from 0 to 1 with higher scores indicating greater objective speech intelligibility [142]. PESQ measures objective speech quality between the processed speech and clean speech, and results in a score between [-0.5,4.5] where higher values are reflective of greater speech quality [143].

6.4.7 Evaluation results

6.4.7.1 Trained speaker condition

Table 6.3 and Table 6.4 show the STOI and PESQ scores for unprocessed and processed signals for Test set 1 (where the same speakers were used in training and testing) using mapping based (TMS) and masking based (IRM) speech enhancement models. Five different neural network structures are compared in terms of STOI and PESQ in three different SNR levels for two different unseen noise types of Babble (Bab) and Cafeteria (Caf). The proposed DCN model consistently outperforms DNN, LSTM, CRN and GRN in both criteria. The average improvements of STOI for DCN for TMS over unprocessed, DNN, LSTM, CRN and GRN are 14.48%, 9.64%, 8.33%, 7.18% and 1.90%, respectively. Furthermore, DCN has higher PESQ score compared to other four networks. The average score of PESQ is improved by 2%, 2.93%, 3.66%, 6% and 7.02% using DNN, LSTM, CRN, GRN and DCN, respectively. For both types of noise and different SNRs, the proposed dilated network has better performance compared to other networks. DCN model

improves STOI and PESQ significantly in more challenging conditions (zero and negative SNR) for both mapping and masking based models.

In masking-based speech enhancement, DCN still outperforms other network structures in terms of STOI with improvement of 10.82%, 6.37%, 5.83%, 4.92% and 2.36% compared to unprocessed, DNN ,LSTM, CRN and GRN, respectively. In terms of PESQ, the dilated network is better than other network structures. The improvements of PESQ using DNN, LSTM, CRN, GRN and our model in masking based method are 3.13%, 4.33%, 6.13% and 7.60%, respectively. There is also a slightly better performance of masking based models compared to mapping based models in speech quality.

Overall, the mapping-based methods (TMS) have higher STOI scores compared to masking based approaches (IRM), although the IRM based networks have slightly better PESQ compared to TMS methods. The amount of improvements for different network types, SNRs, noise types and training targets are shown in Fig. 6.2. The spectrograms of enhanced speech processed with each network structure are shown on Fig. 6.3. For both negative and zero SNRs, the STOI improvement is higher than positive SNRs in mapping and masking based enhancement. Moreover, the highest improvement of PESQ is found for 0 SNR in IRM and TMS based models. Both dilated neural networks of GRN and DCN have higher performance compared to DNN, LSTM and CRN because these two networks have larger receptive field to capture information about previous and future time frames in final prediction of masks or clean spectra. However, GRN uses 62 layers of convolution on 10 s information versus 26 layers to capture the same amount of time frame information in DCN.

6.4.7.2 Untrained speaker condition

The analysis is repeated for untrained speakers, which means that the speakers used in training and testing are different. Our goal is to evaluate whether the proposed network generalizes well to untrained speakers and noise. Table 6.5 and Table 6.6 compare the performance of DNN, LSTM, CRN, GRN and the DCN model using TMS and IRM for untrained speakers in terms of STOI and

Table 6.3: Model and Training target comparison in terms of STOI on trained speakers.

Metric	**STOI (%)**								
SNR	**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	51.11	55.08	53.09	58.29	60.74	59.51	64.83	67.50	66.17
DNN+TMS	54.90	59.51	57.20	64.16	65.85	65.00	70.39	71.77	71.08
LSTM+TMS	53.95	60.69	57.32	64.98	68.29	66.63	72.48	74.05	73.26
CRN+TMS	57.65	62.20	59.92	66.21	68.49	67.35	72.71	74.12	73.41
GRN+TMS	62.19	68.39	65.29	71.25	74.34	72.75	77.77	79.20	78.48
DCN+TMS	**64.96**	**70.58**	**67.77**	**73.36**	**75.90**	**74.63**	**79.28**	**80.36**	**79.82**
DNN+IRM	54.49	59.14	56.81	63.60	65.55	64.57	70.02	71.49	70.75
LSTM+IRM	54.06	59.83	56.94	63.95	66.72	65.33	70.44	72.49	71.46
CRN+IRM	56.23	60.83	58.53	64.97	67.32	66.14	71.38	72.20	71.79
GRN+IRM	58.51	64.26	61.38	67.35	69.85	68.60	73.42	74.95	74.18
DCN+IRM	**62.35**	**67.23**	**64.79**	**70.14**	**71.60**	**70.87**	**75.28**	**75.88**	**75.58**

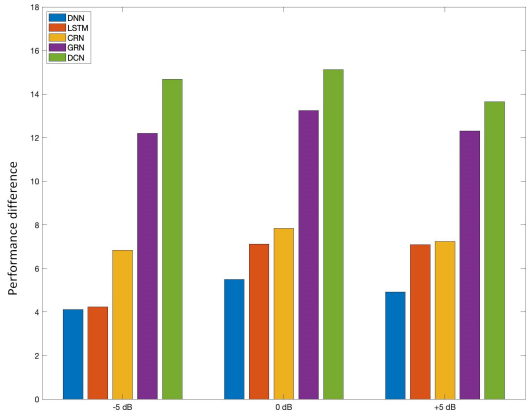
PESQ. The rate of increase in STOI using DCN with TMS is 14.83% compared to unprocessed, while the rate of increase in STOI using DNN, LSTM, CRN and GRN are 5.32%, 7.15%, 8.26% and 12.12% compared to the unprocessed noisy mixture. The average PESQ score for our model is 1.95, while GRN, CRN, LSTM, DNN and unprocessed mixture scores are 1.87, 1.82, 1.79 ,1.74 and 1.62 for mapping based enhancement.

The lower section of Table 6.5 and Table 6.6 show the STOI and PESQ scores of masking based models that predict the IRMs for speech enhancement. DCN with IRM estimation increases STOI by 11.27%, while DNN, LSTM, CRN and GRN have improvements of 4.51%, 5.79%, 6.44% and 9.75%, respectively. For example, in one of the most challenging conditions (SNR=-5 dB and Babble noise) for untrained speakers, the proposed model increases STOI by 15.50% compared to the unprocessed mixture, while LSTM, CRN and GRN have 5.95%, 8.2%, 11.43% improvement.

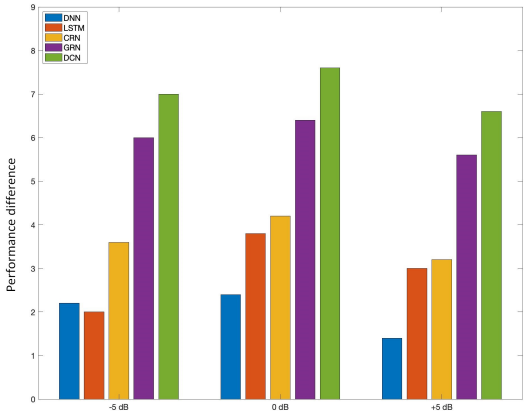
Table 6.4: Model and Training target comparison in terms of PESQ on trained speakers.

Metric	**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	1.31	1.48	1.39	1.57	1.77	1.67	1.87	2.02	1.94
DNN+TMS	1.41	1.59	1.50	1.72	1.87	1.79	1.96	2.07	2.01
LSTM+TMS	1.37	1.62	1.49	1.77	1.96	1.86	2.04	2.15	2.09
CRN+TMS	1.48	1.67	1.57	1.81	1.96	1.88	2.05	2.16	2.10
GRN+TMS	1.58	1.81	1.69	1.92	2.07	1.99	2.17	2.27	2.22
DCN+TMS	**1.64**	**1.85**	**1.74**	**1.98**	**2.13**	**2.05**	**2.22**	**2.33**	**2.27**
DNN+IRM	1.41	1.63	1.52	1.75	1.93	1.84	2.05	2.17	2.11
LSTM+IRM	1.45	1.67	1.56	1.84	2.01	1.92	2.10	2.25	2.17
CRN+IRM	1.49	1.71	1.60	1.85	2.02	1.93	2.12	2.24	2.18
GRN+IRM	1.54	1.77	1.65	1.93	2.10	2.01	2.20	2.33	2.26
DCN+IRM	**1.62**	**1.87**	**1.74**	**2.00**	**2.16**	**2.08**	**2.26**	**2.39**	**2.32**

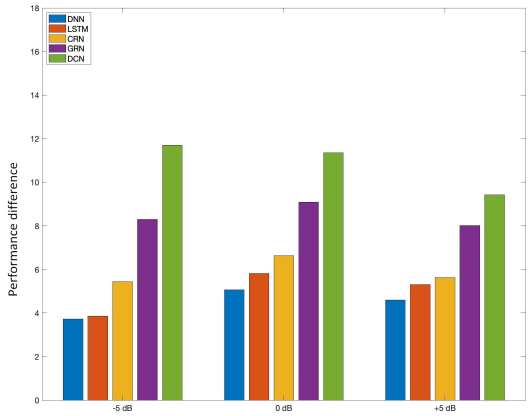
It should be noted that the PESQ scores in masking based speech enhancement are slightly higher than the mapping based models for untrained speaker. The improvements of STOI and PESQ scores per model type, noise type, SNR level and training target are visualized in detail on Fig. 6.4. In conclusion, the performance of the DCN is better than the other four models for untrained speakers in both mapping and masking based speech enhancement. Furthermore, the speech intelligibility and speech quality of processed signals using this model is similar to the ones tested on trained speakers. Therefore, we can conclude that DCN generalizes well to unseen speakers, which is desirable in designing supervised speech separation approaches with applications in ASR, hearing aid devices and telecommunications. One advantage of the proposed model is that it leverages contexts in both frequency and time similar to CRN and GRN which leads to modeling more complex temporal dependency. Moreover, the dilated blocks deepen the network more efficiently



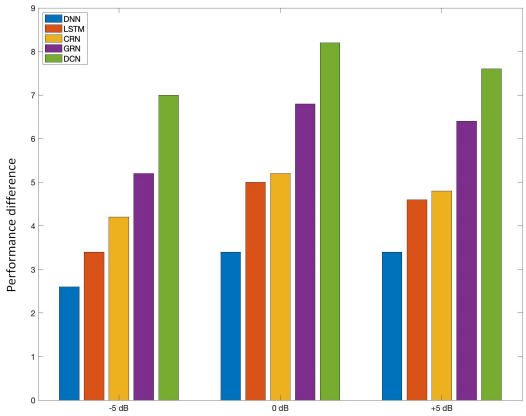
(a) ΔSTOI (%) for mapping based



(b) ΔPESQ (%) for mapping based



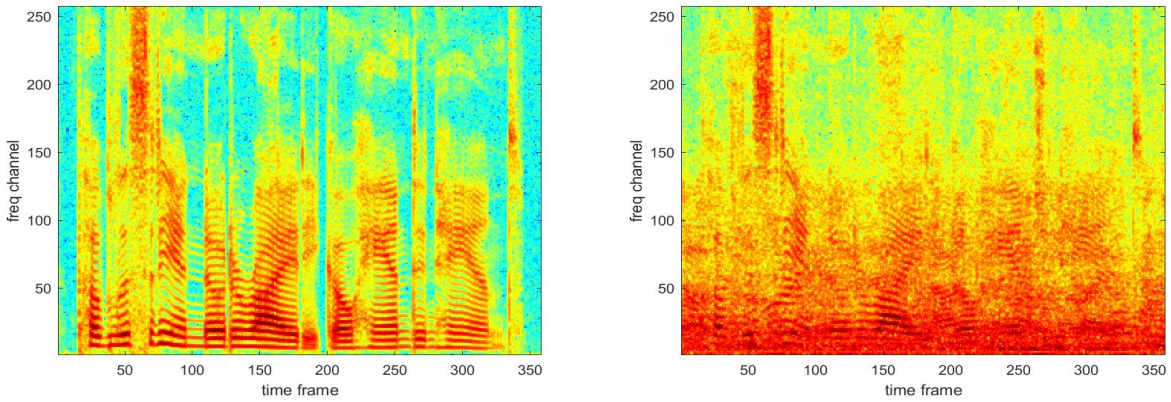
(c) ΔSTOI (%) for masking based



(d) ΔPESQ (%) for masking based

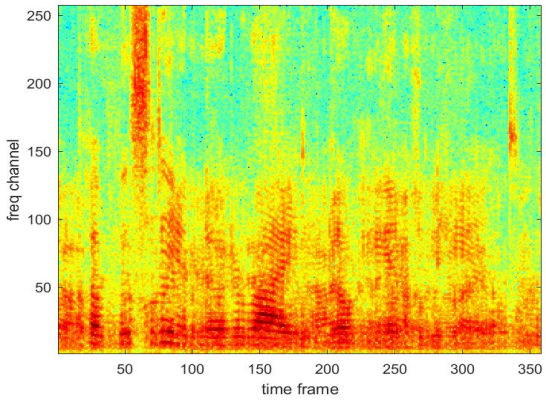
Figure 6.2: Improvements of STOI (%), PESQ (%) for different neural network architectures for mapping based speech enhancement (a,b) and masking based speech enhancement (c,d) on trained speakers.

than increasing the length of the filter and these blocks work as a soft mask on the learned features of previous layers. Another advantage of this model is its higher computational efficiency. The total number of trainable parameters for each model are reported on Table 6.2.

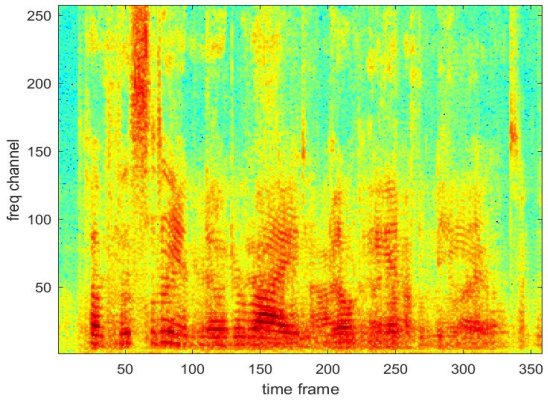


(a) Clean speech

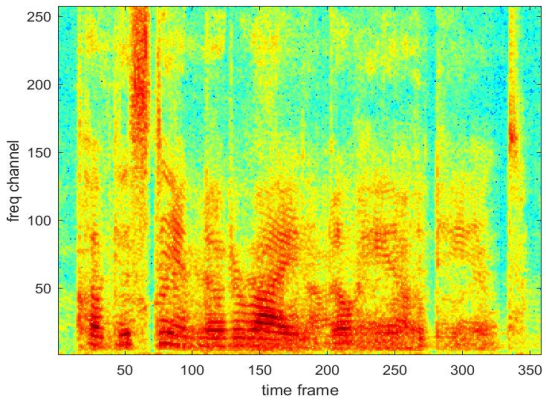
(b) Noisy reverberant mixture



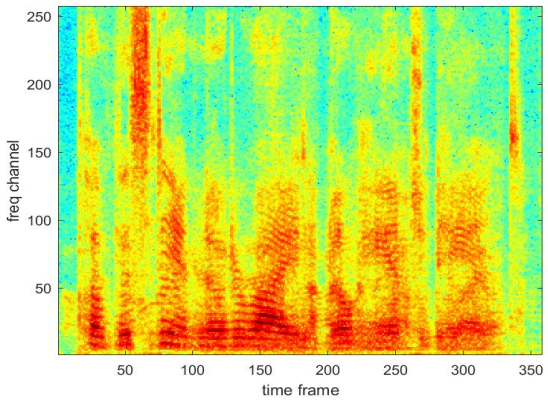
(c) DNN



(d) CRN



(e) GRN



(f) Proposed DCN

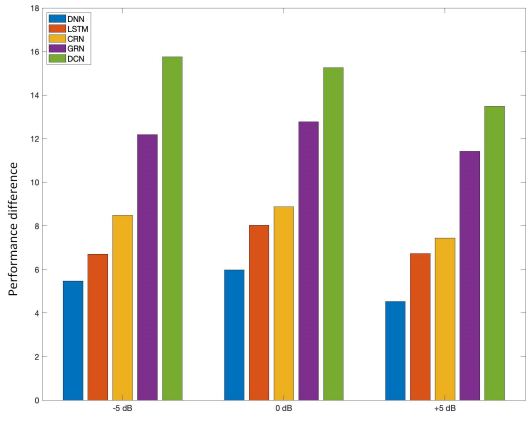
Figure 6.3: Visualization of processed signals with different masking based speech enhancement models, where Babble is the background noise (SNR = 0) and $T_{60} = 0.68s$.

Table 6.5: Model and Training target comparison in terms of STOI on untrained speakers.

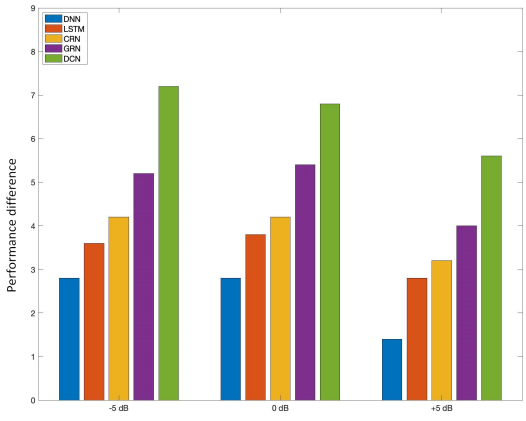
Metric	**STOI (%)**								
SNR	**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	50.31	54.31	52.31	57.10	61.12	59.11	63.89	65.69	64.79
DNN+TMS	55.73	59.82	57.77	63.33	66.85	65.09	69.14	69.51	69.32
LSTM+TMS	56.26	61.76	59.01	65.04	69.24	67.14	71.07	71.96	71.51
CRN+TMS	58.51	63.10	60.80	65.97	69.99	67.98	71.89	72.57	72.23
GRN+TMS	61.74	67.26	64.50	69.79	73.97	71.88	75.67	76.75	76.21
DCN+TMS	**65.81**	**70.33**	**68.07**	**72.71**	**76.04**	**74.37**	**78.07**	**78.49**	**78.28**
DNN+IRM	54.95	58.93	56.94	62.48	66.16	64.32	68.25	68.71	68.48
LSTM+IRM	55.71	60.49	58.10	63.91	67.38	65.64	69.77	69.91	69.84
CRN+IRM	57.78	61.45	59.61	64.33	67.81	66.07	69.80	69.94	69.87
GRN+IRM	61.31	64.99	63.15	67.90	70.89	69.39	73.02	72.83	72.92
DCN+IRM	**63.83**	**66.97**	**65.40**	**69.53**	**71.97**	**70.75**	**74.22**	**73.54**	**73.88**

Table 6.6: Model and Training target comparison in terms of PESQ on untrained speakers.

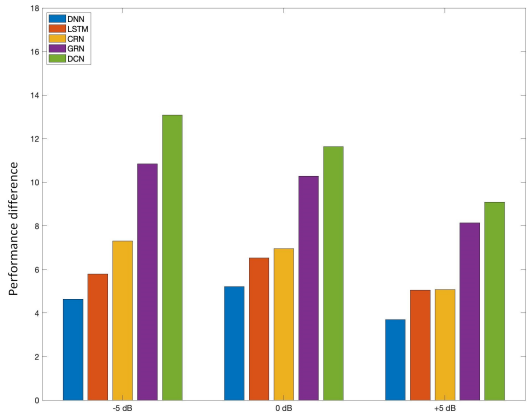
Metric	**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	1.25	1.44	1.34	1.53	1.75	1.64	1.84	1.97	1.90
DNN+TMS	1.38	1.59	1.48	1.68	1.88	1.78	1.95	2.00	1.97
LSTM+TMS	1.39	1.66	1.52	1.73	1.93	1.83	2.00	2.08	2.04
CRN+TMS	1.44	1.67	1.55	1.75	1.95	1.85	2.02	2.10	2.06
GRN+TMS	1.49	1.72	1.60	1.81	2.01	1.91	2.06	2.14	2.10
DCN+TMS	**1.59**	**1.81**	**1.70**	**1.90**	**2.07**	**1.98**	**2.13**	**2.23**	**2.18**
DNN+IRM	1.37	1.61	1.49	1.71	1.93	1.82	2.03	2.12	2.07
LSTM+IRM	1.41	1.69	1.55	1.80	1.99	1.89	2.10	2.17	2.13
CRN+IRM	1.45	1.70	1.57	1.79	2.00	1.89	2.09	2.17	2.13
GRN+IRM	1.56	1.79	1.67	1.91	2.10	2.00	2.21	2.29	2.25
DCN+IRM	**1.59**	**1.84**	**1.71**	**1.95**	**2.14**	**2.04**	**2.25**	**2.32**	**2.28**



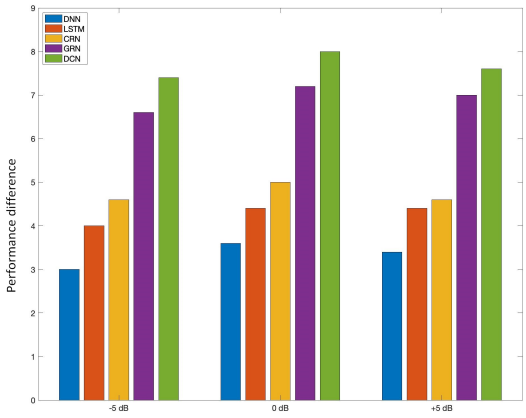
(a) ΔSTOI (%) for mapping based



(b) ΔPESQ (%) for mapping based



(c) ΔSTOI (%) for masking based



(d) ΔPESQ (%) for masking based

Figure 6.4: Improvements of STOI (%), PESQ (%) for different neural network architectures for mapping based speech enhancement on untrained speakers.

6.4.7.3 Speaker and noise dependent model

Furthermore, DCN is evaluated as a speaker and noise dependent model for mapping based speech enhancement. The same procedure was used to generate the training and testing dataset. Same speakers and same noise types (Babble and Cafeteria) are used to generate both training and testing dataset. The goal of this experiment is to compare the performance of DCN as speaker-noise dependent model versus the DCN as speaker-noise independent model to show the generalization ability of DCN. As can be seen from Table 6.7, the DCN speaker-noise independent model performs similarly to the DCN speaker-noise dependent model. Therefore, DCN can perform well in unseen conditions such as unseen noise, speakers and room simulations. DCN is a very efficient model that has fewer number of parameters with faster training and less memory requirements compared to other models.

Table 6.7: STOI and PESQ scores for speaker-noise dependent model versus the speaker-noise independent model for mapping based DCN speech enhancement model.

Metrics	**STOI (%)**									**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**			**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	55.09	50.90	52.99	61.91	56.82	59.36	68.10	64.83	66.46	1.47	1.31	1.39	1.77	1.56	1.66	2.03	1.86	1.94
DCN (Independent model)	64.96	70.58	67.77	73.36	75.90	74.63	79.28	80.36	79.82	1.64	1.85	1.74	1.98	2.13	2.05	2.22	2.33	2.27
DCN (Dependent model)	65.76	70.28	68.02	74.76	76.41	75.58	80.61	80.74	80.67	1.99	2.03	2.01	2.01	1.99	2.00	2.01	2.05	2.03

6.5 Real-time monaural speech enhancement

6.5.1 Dilated convolutional recurrent neural network

In many speech processing applications such as hearing aid devices and cochlear implants, real-time processing and low latency is necessary; therefore, many supervised speech separation algorithms need to operate in real-time in order to improve the performance of these devices. Inspired

by our dilated convolution neural network described in previous section, a new architecture is proposed to enable the dilated network to work in real-time, which means that the neural network only uses the information about previous time frames to estimate the time frequency mask or spectrum of clean speech for current time frame. In this section, a new dilated convolutional recurrent neural network (DCRN) is proposed that combines two network types of dilated causal convolutions [162] with LSTM to make the network suitable for real-time applications where no future time frames are needed in training [163]. The proposed dilated convolutional recurrent network (DCRN) includes 1 block of 2D-convolutions with four convolution layers and 2 max pooling layers (similar to DCN proposed in section 6.4.2 but uses only the frames from the current and past time frames); 2 blocks of dilated causal convolutions with dilation rates of 2, 4, 8, 16, 32, 64, 128 for soft mask estimation on features, one layer of LSTM with 64 units stacked between the 2 blocks of dilated causal convolution and skip connections. This model is very similar to DCN with the only difference in using causal 1D-convolutions and LSTM layer included in the middle of the dilated blocks. Moreover, the kernel size in 2D-convolutional layers are reduced on the time axis to only include current and previous time frames in output prediction. Therefore, DCRN is a completely causal model and it is suitable for real-time processing. The LSTM layer is used to leverage long-term context that the convolutional layers can not provide and it is concatenated with the summation of soft masked higher level features and soft masked lower level features [163].

DCRN utilizes both convolutional layers and recurrent layer (LSTM) to predict the time frequency mask or the spectrum of clean speech. Both types of network use only previous time frames to predict the current time frame. The output of LSTM layer is directly concatenated with the summation of skip connections to equally include both short-term and long-term contexts in our model. Furthermore, DCRN has slightly fewer number of training parameters (2.66 million) compared to DCN with 2.92 million and it can also operate in real-time processing. The DCRN network structure is visualized on Fig. 6.5 and the detailed description of it is given in Table 6.9. The hyperparameters for 2D-convolution layers are shown as (kernel size, (strides), output channel), while the hyperparameters for 1D-convolution are given as (kernel size, dilation rate, output

channels) in Table 6.9. Note that the zero-padding is applied to both 2D-convolution and 1D-convolution layers. DCRN is compared with four different baselines of DNN, LSTM, CRN and GRN(R) explained in the next section. The proposed dilated network uses a large past information to predict the spectrum of clean speech or the time frequency mask for speech enhancement. The total number of trainable parameters for different network structures are given in Table. 6.8.

Table 6.8: The total number of trainable parameters for each neural network model.

Network type	LSTM	CRN	GRN(R)	**DCRN**
Number of parameters (millions)	25.51	17.22	3.08	**2.66**

Table 6.9: Architecture of our proposed dilated convolutional recurrent network (DCRN).

Block name	**Layer name**	**hyperparameters**
Causal Conv2d+Maxpooling	conv2d_1	$2\times5,(1\times1),32$, BN, ELU
	conv2d_2	$2\times9,(1\times1),32$, BN, ELU
	maxpool_1	$1\times2,(1\times2)$
	conv2d_3	$2\times5,(1\times1),64$, BN, ELU
	conv2d_4	$2\times9,(1\times1),64$, BN, ELU
	maxpool_2	$1\times2,(1\times2)$
Causal Conv1d	conv1d_1	$3,1,256$, BN, ELU
Dilated causal convolution block1	conv1d_d1	$3,2,16$, ELU
	conv1d_d2	$3,4,16$, ELU
	conv1d_d3	$3,8,16$, ELU
	conv1d_d4	$3,16,16$, ELU
	conv1d_d5	$3,32,16$, ELU
	conv1d_d6	$3,64,16$, ELU
	conv1d_d7	$3,128,16$, ELU
Causal Conv1d	conv1d_2	$3,1,256$, ELU
LSTM	lstm1	64, BN
Causal Conv1d	conv1d_3	$3,1,256$, BN, ELU
Dilated causal convolution block2	conv1d_d8	$3,2,16$, ELU
	conv1d_d9	$3,4,16$, ELU
	conv1d_d10	$3,8,16$, ELU
	conv1d_d11	$3,16,16$, ELU
	conv1d_d12	$3,32,16$, ELU
	conv1d_d13	$3,64,16$, ELU
	conv1d_d14	$3,128,16$, ELU
Causal Conv1d	conv1d_4	$3,1,256$, ELU
Causal Conv1d	conv1d_5	$3,1,256$, ELU
Causal Conv1d	conv1d_6	$1,1,161$, ReLU/Sigmoid

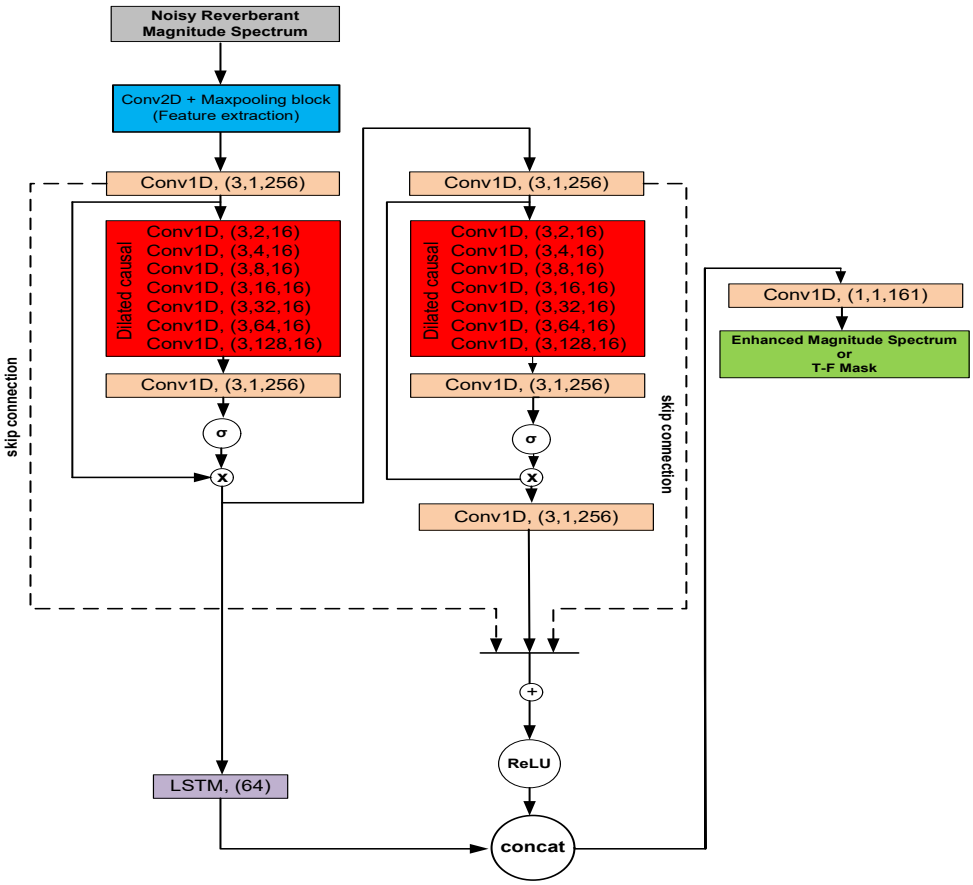


Figure 6.5: Proposed dilated convolutional recurrent network (DCRN) architecture.

6.5.2 Comparison models

DNN: A simple feedforward neural network is used with 4 hidden layers of 1024 units. The activation layer of each hidden unit is ELU and droput rate of each hidden layer is 0.2. Moreover, batch normalization is applied on each layer before passing the output through the activation functions.

LSTM: The LSTM network includes four hidden layers with the size of 1024 units while the input and output layers have 161 units. The time step for these layers is 128 and the dropout rate of 0.2 is used on each layer. Each hidden layer is also batch normalized. The output layer of this model is a fully-connected layer (Dense layer) with the ReLU or sigmoid activation function for the mapping based speech enhancement and the masking based speech enhancement model, respectively.

CRN: The convolutional recurrent network (CRN) [152] has a total of ten 2D-convolution layers

and two LSTM layers. CRN includes an encoder and a decoder with five convolutional layers and five deconvolutional layers, respectively. Moreover, the convolutional layers are causal and have a stride of 2 along frequency. In addition, 2 layers of LSTM are stacked between encoder and decoder to model temporal dynamics of speech and skip connections are used to concatenate the output of each encoder layer to the input of each decoder layer [152]. This model is operating in real-time using causal convolutions in the encoder and decoder structure.

GRN(R): Gated residual network (GRN) is a 62 convolutional layer network, constructed with a stack of frequency-dilated convolutional layers and multiple gated residual blocks with different dilation rates [151]. The original GRN does not have causal layers, therefore, the real-time version of it (GRN(R)) is re-designed for comparison purpose with our DCRN model.

6.5.3 Experimental setup

6.5.3.1 Dataset

The dataset is generated similarly as section 6.4.6.1 from TIMIT corpus [140] using utterances from 462 speakers with 300 different background noise from the FreeSFX [160] and Freesound [161] corpora and simulated IRs generated by the room impulse response generator [137]. The Babble and Cafeteria noises were used as unseen noises during testing. A SNR for training set is selected randomly from the range of [-5,5] dB, while the SNR levels in testing set are -5, 0 and 5 dB.

The noisy-reverberant mixtures (Total=33000, training=29700, development=3300) are generated from 1650 utterances of 462 speakers (female and male). Two test sets were generated as explained in 6.4.6.1 to investigate the generalization of the proposed model to different speakers and noise. One set includes 1000 noisy-reverberant mixtures of 6 trained speakers (3 males, 3 females) and the other test set used 6 untrained speakers (3 males, 3 females) to generate 1000 noisy-reverberant mixture. The noisy-reverberant mixtures were generated by convolving the clean speech signal with real IRs recorded at the University of Surrey [138] as discussed in section. 6.4.6.1. The inputs of the neural networks are the magnitude spectra with the dimension of 161,

calculated from the audio mixtures that were segmented into time frames with a 20 ms hamming window and 10 ms overlap (50%). A 320-point STFT was used to calculate the magnitude spectra.

MSE was employed as the loss function for all models with Adam algorithm [117] for back-propagation optimization. The batch size of 16 was used for convolutional models, while LSTM used 256 batch size. The number of training epochs was 50 and all features were normalized to have zero mean and unit variance on each frequency channel before feeding to the networks. Batch normalization was applied on hidden layers for faster training [123].

6.5.4 Evaluation results

Two different evaluations are reported for trained speaker and untrained speaker conditions. In the trained speaker case, the utterances from the same speakers were used to generate the training and test dataset, while in untrained speaker case, speakers were different in testing and training. The STOI and PESQ scores are reported for evaluation of different models, and training targets for various SNR levels and different unseen noises on Table. 6.10 and Table. 6.11.

6.5.4.1 Trained speaker condition

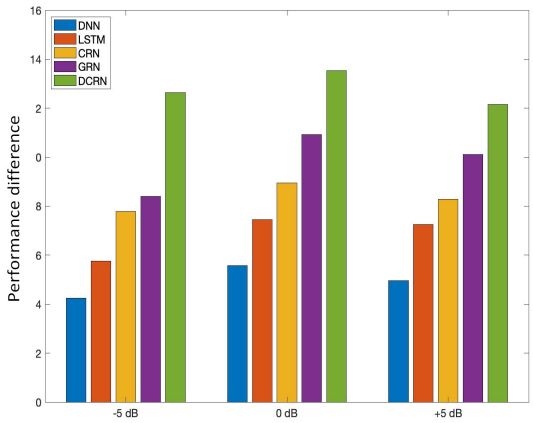
The evaluation results for trained speaker condition are reported on Table 6.10. It should be noted that all the neural network models present in this table can operate in real-time. On the upper section of Table 6.10, the STOI and PESQ scores of mapping based speech enhancement models are shown. It can be seen that DCRN outperforms all the other neural network models for trained speakers in terms of STOI. The average amount of increase in STOI for DCRN, GRN(R), CRN, LSTM and DNN models are 12.77%, 9.81%, 8.34%, 6.81% and 4.93% compared to unprocessed mixtures, respectively. The improvements of STOI and PESQ compared to unprocessed mixtures are drawn on Fig 6.7 in (%). The highest amount of improvement in STOI score for all the trained models is for 0 dB SNR. Furthermore, the same improvement pattern can be seen in the PESQ score. On the other hand, GRN(R) has slightly higher PESQ score for both noises in mapping based case for +5 dB SNR. The average improvement in PESQ score for DCRN, GRN(R), CRN,

LSTM and DNN mapping-based models are 4.8%, 4.53%, 3.86%, 2.4% and 2.13%, respectively.

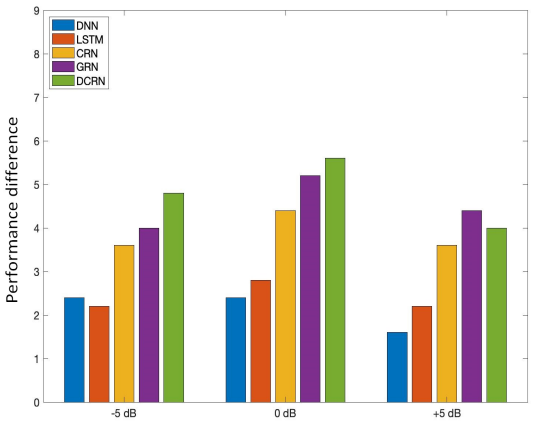
In the lower section of Table. 6.10, the STOI and PESQ scores of masking based models are reported. The average amount of increase in STOI for DCRN, GRN(R), CRN, LSTM and DNN masking based models are 9.51%, 6.86%, 6.43%, 5.54% and 4.43% compared to unprocessed mixtures, respectively. Furthermore, the PESQ scores are improved by 7.33%, 5%, 4.73%, 4.40% and 2.86% using DCRN, GRN(C), CRN, LSTM and DNN, respectively. It should be noted that while the STOI scores of mapping based models are higher than the masking based ones, the PESQ of masking based models are greater than the mapping based models. The amount of improvement for PESQ scores in masking based models are much higher than mapping based models, which is similar to the non-real time models examined in previous section. In conclusion, the proposed causal DCRN provides significant improvement in objective speech intelligibility and speech quality in comparison to other network, while it has fewer trainable parameters which makes it computationally more efficient and suitable for devices with low power supply. The performance of DCN is slightly better than DCRN, while DCRN is a causal model.

Table 6.10: Model and Training target comparison in terms of STOI and PESQ on trained speakers.

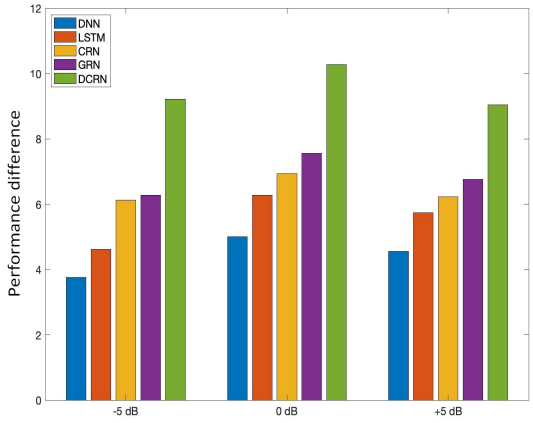
Metrics	**STOI (%)**									**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**			**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	51.11	55.08	53.09	58.29	60.74	59.51	64.83	67.50	66.16	1.31	1.48	1.39	1.57	1.77	1.67	1.87	2.02	1.94
DNN+TMS	54.98	59.71	57.34	64.10	66.08	65.09	70.43	71.81	71.12	1.42	1.60	1.51	1.72	1.87	1.79	1.97	2.07	2.02
LSTM+TMS	56.09	61.60	58.84	65.56	68.37	66.96	72.64	74.19	73.41	1.40	1.61	1.50	1.74	1.91	1.82	2.00	2.11	2.05
CRN+TMS	58.17	63.59	60.88	67.04	69.89	68.46	73.63	75.28	74.45	1.45	1.70	1.57	1.80	1.98	1.89	2.06	2.18	2.12
GRN(R)+TMS	58.71	64.28	61.49	69.07	71.79	70.43	75.62	76.93	76.27	1.48	1.70	1.59	1.85	2.01	1.93	**2.11**	**2.22**	**2.16**
DCRN+TMS	**63.02**	**68.43**	**65.72**	**71.71**	**74.37**	**73.04**	**77.76**	**78.86**	**78.31**	**1.53**	**1.74**	**1.63**	**1.89**	**2.02**	**1.95**	2.10	2.19	2.14
DNN+IRM	54.53	59.17	56.85	63.52	65.50	64.51	69.95	71.48	70.71	1.40	1.62	1.51	1.74	1.93	1.83	2.03	2.16	2.09
LSTM+IRM	55.07	60.35	57.71	64.56	67.00	65.78	71.27	72.53	71.90	1.46	1.68	1.57	1.83	2.00	1.91	2.11	2.25	2.18
CRN+IRM	56.80	61.65	59.22	65.37	67.52	66.44	71.78	73.00	72.39	1.48	1.73	1.60	1.85	2.01	1.93	2.12	2.24	2.18
GRN(R)+IRM	56.24	62.49	59.36	65.65	68.49	67.07	72.11	73.74	72.92	1.47	1.72	1.59	1.86	2.04	1.95	2.15	2.27	2.21
DCRN+IRM	**59.34**	**65.29**	**62.31**	**68.64**	**70.93**	**69.78**	**74.76**	**75.64**	**75.20**	**1.61**	**1.83**	**1.72**	**1.99**	**2.14**	**2.06**	**2.27**	**2.37**	**2.32**



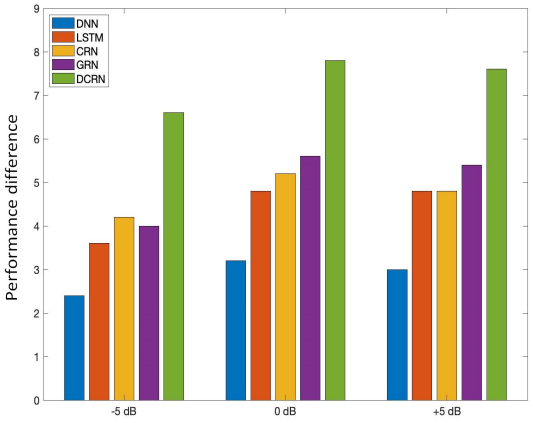
(a) ΔSTOI (%) for mapping based



(b) ΔPESQ (%) for mapping based



(c) ΔSTOI (%) for masking based



(d) ΔPESQ (%) for masking based

Figure 6.6: Improvements of STOI (%), PESQ (%) for different neural network architectures for mapping based and masking based real-time speech enhancement on trained speakers.

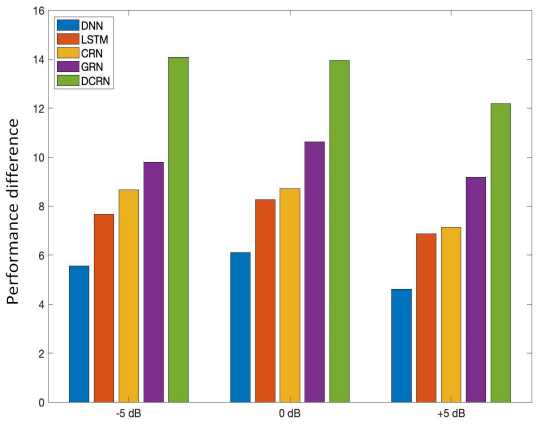
6.5.4.2 Untrained speaker condition

The same analysis was performed on Test data 2 for untrained speakers. The STOI and PESQ scores for different models, unseen noises and SNRs are reported on Table. 6.11. For mapping based models, the proposed DCRN improves STOI by 13.40%, while DNN, LSTM, CRN and GRN(R) improve STOI by 5.42%, 7.60%, 8.17% and 9.86%, respectively. In -5 and 0 SNR, the GRN(R) model has slightly better PESQ score for Cafeteria noise compared to DCRN. The average increase of PESQ score for DCRN is 3.86% while GRN(R) improves PESQ by 3.73%.

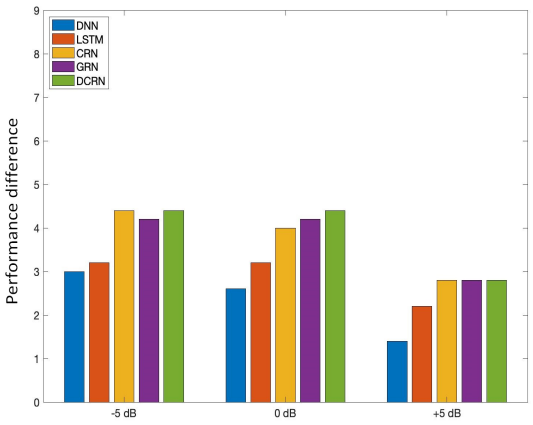
In masking based models, DCRN significantly outperforms other causal models in term of STOI and PESQ. DCRN increases STOI on average by 9.82% and PESQ by 7.20%. The detailed description of the increase of STOI and PESQ for different models and different SNR levels are shown on Fig. 6.7. To conclude, the proposed DCRN outperforms all other causal models in improving objective speech intelligibility and speech quality for both unseen noises and unseen speakers. This model generalizes well to unseen conditions, while having fewer trainable parameters, which leads to shorter training time.

Table 6.11: Model and Training target comparison in terms of STOI and PESQ on untrained speakers.

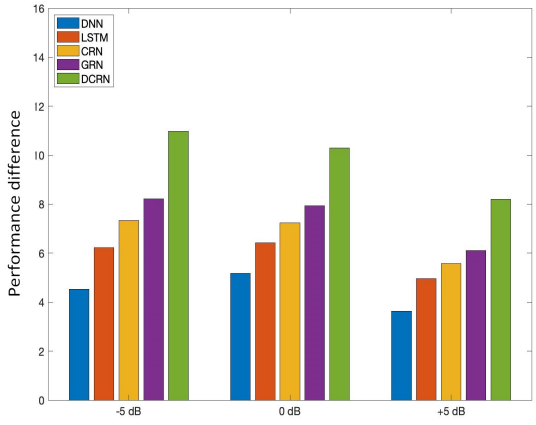
Metrics	**STOI (%)**									**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**			**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	50.31	54.31	52.31	57.10	61.12	59.11	63.89	65.69	64.79	1.25	1.44	1.34	1.53	1.75	1.64	1.84	1.97	1.90
DNN+TMS	55.70	60.04	57.87	63.38	67.06	65.22	69.16	69.64	69.40	1.39	1.59	1.49	1.68	1.87	1.77	1.94	2.00	1.97
LSTM+TMS	57.37	62.59	59.98	65.47	69.30	67.38	71.23	72.10	71.66	1.36	1.64	1.50	1.70	1.90	1.80	1.98	2.05	2.01
CRN+TMS	59.09	62.88	60.98	66.03	69.63	67.83	71.59	72.25	71.92	1.46	1.67	1.56	1.75	1.94	1.84	2.00	**2.09**	2.04
GRN(R)+TMS	59.25	64.95	62.10	67.67	71.80	69.73	73.43	74.52	73.97	1.41	**1.69**	1.55	1.75	**1.95**	1.85	2.00	2.08	2.04
DCRN+TMS	**64.04**	**68.75**	**66.39**	**71.23**	**74.87**	**73.05**	**76.88**	**77.08**	**76.98**	**1.46**	1.67	**1.56**	**1.78**	1.94	**1.86**	2.00	2.08	2.04
DNN+IRM	54.73	58.94	56.83	62.41	66.16	64.28	68.15	68.70	68.42	1.36	1.61	1.48	1.70	1.93	1.81	2.01	2.11	2.06
LSTM+IRM	56.78	60.31	58.54	63.86	67.21	65.53	69.68	69.83	69.75	1.42	1.66	1.54	1.76	1.97	1.86	2.08	2.17	2.12
CRN+IRM	57.82	61.47	59.64	64.54	68.14	66.34	70.38	70.34	70.36	1.45	1.71	1.58	1.81	2.02	1.91	2.10	2.19	2.14
GRN(R)+IRM	58.37	62.67	60.52	65.26	68.82	67.04	70.74	71.06	70.90	1.47	1.73	1.60	1.84	2.02	1.93	2.12	2.20	2.16
DCRN+IRM	**61.42**	**65.17**	**63.29**	**67.97**	**70.83**	**69.40**	**73.14**	**72.84**	**72.99**	**1.56**	**1.80**	**1.68**	**1.93**	**2.12**	**2.02**	**2.23**	**2.30**	**2.26**



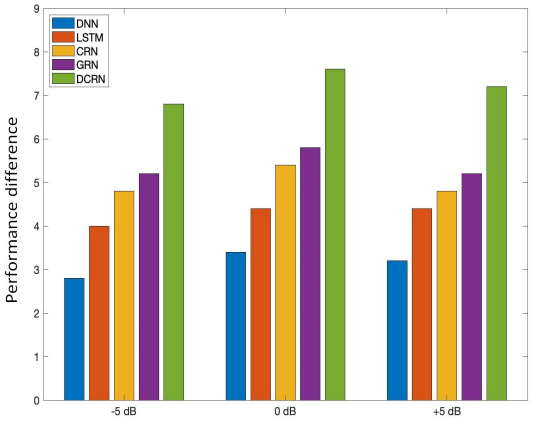
(a) ΔSTOI (%) for mapping based



(b) ΔPESQ (%) for mapping based



(c) ΔSTOI (%) for masking based



(d) ΔPESQ (%) for masking based

Figure 6.7: Improvements of STOI (%), PESQ (%) for different neural network architectures for mapping based and masking based real-time speech enhancement on untrained speakers.

6.5.4.3 Speaker and noise dependent model

In addition, DCRN is evaluated as a speaker-noise dependent model for mapping based speech enhancement. The same procedure was done to generate the training and testing dataset. Same speakers and same noise types (Babble and Cafeteria) are used to generate both training and testing dataset. The goal of this experiment is to compare the performance of DCRN as speaker-noise dependent model versus the DCRN as speaker-noise independent model to show the generalization ability of DCRN. As can be seen from Table 6.12, the DCRN speaker-noise independent model performs similarly to the DCRN speaker-noise dependent (the independent model has slightly lower STOI scores but higher PESQ scores). Therefore, DCRN performance is robust to unseen conditions such as unseen noises, unseen speakers and unseen room simulations. Our DCRN model generalizes well to different unseen conditions which make it a very strong model for supervised speech separation that can deal with generalization issue.

Table 6.12: STOI and PESQ scores for speaker-noise dependent model versus the speaker-noise independent model for mapping based DCRN speech enhancement model.

Metrics	**STOI (%)**									**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**			**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	55.09	50.90	52.99	61.91	56.82	59.36	68.10	64.83	66.46	1.47	1.31	1.39	1.77	1.56	1.66	2.03	1.86	1.94
DCRN (Independent model)	63.02	68.43	65.72	71.71	74.37	73.04	77.76	78.86	78.31	1.53	1.74	1.63	1.89	2.02	1.95	2.10	2.19	2.14
DCRN (Dependent model)	64.43	69.16	66.79	73.60	75.30	74.45	79.49	79.72	79.60	1.61	1.62	1.61	1.91	1.90	1.90	2.08	2.10	2.09

6.6 Discussions

In this chapter, a novel dilated convolutional neural network (DCN) architecture is presented for noise and speaker independent speech enhancement, which exploits dilated convolution to aggregate multi-scale contextual information, while keeping the network parameters small. DCN is an efficient network structure that stacks multiple 1D dilated convolutions to generate soft mask on different features learned from lower levels. The evaluation of speech enhancement performance indicates that DCN outperforms comparable LSTM, CRN and GRN in terms of STOI and PESQ for both trained and untrained speakers. The proposed model has significant improvement in monaural speech enhancement and it has better generalization to different types of challenging unseen noises (such as babble and cafeteria) and untrained speakers compared to previous models.

In addition, a new model (DCRN) capable of real-time processing is presented that combines the recurrent neural network with dilated causal convolutions for monaural speech enhancement that only needs the information about the current and previous time frames for time frequency mask or clean spectrum estimation of the current time frame. DCRN generalizes well to different unseen noises and untrained speakers which means that it is a causal speaker-noise independent model, and is desirable for many real-world applications such as ASR, hearing aid devices, mobile phones and headsets.

Chapter 7

Conclusions and Future Work

7.1 Contribution

Speech separation is one method for solving the cocktail party problem. Over the years, many speech separation algorithms were proposed to solve this issue; however, it is still a challenge to produce high quality and intelligible speech in adverse environment with low SNR level and high reverberation. In this dissertation, different speech separation algorithms are proposed to separate the target speech from the interference in these highly reverberant and noisy environment.

Inspired by the human auditory system in separating target speech from interference using spatial cues, two novel algorithms were proposed to focus on target speakers in reverberant environments in chapter 3. The IM algorithm increases the discrimination of different speakers by magnifying the interaural level difference of interfering sounds using its HRTF. The IM algorithm improves the objective speech quality significantly in both anechoic and reverberant rooms. The second proposed speech separation algorithm (DUET-CP) employs binary time frequency masks based on spatial cues of ILD and ITD to separate target speaker from an interfering speaker in reverberant room. This algorithm first dereverberates the binaural mixtures, then separates them using classical DUET algorithm and finally preserves the location of sound sources after separation. DUET-CP algorithm reduces reverberation, while separating target male speaker from interfering female speaker. Furthermore, the subjective localization test on native English speakers illustrates the effectiveness of this algorithm for cue preservation after speech dereverberation and speech separation.

In chapter 4, one of the most important parts of DNN based supervised speech separation meth-

ods called acoustic features is investigated. Different common spectral features and gammatone features are investigated for masking based speech separation for different room simulations with reverberation(recorded IRs and simulated IRs), matched and unmatched noises and co-channel (2 speakers) condition. The gammatone based features have better performance in increasing speech intelligibility and speech quality for masking based speech separation in matched noise conditions compared to spectral features. Inspired by MRCG, a new gammatone based feature called DMRCG is proposed to further improve the performance of DNN for speech separation in noisy-reverberant environments. The proposed monaural feature includes different levels of contextual information and frequency dynamics of speech spectrum using Laplacian of Gaussian filtering to capture the target speech in noisy-reverberant environment. DMRCG increases objective speech intelligibility and speech quality significantly in different simulated rooms and seen and unseen non-stationary noises. Moreover, this feature is robust to different neural network types and by changing the network type from simple MLP to LSTM, DMRCG still outperforms other features in improving speech quality and intelligibility. In addition, gammatone-domain features and DMRCG has better generalization to unseen conditions which indicates that the features extracted based on human auditory perception are better representation of speech to the neural network.

In chapter 5, a complementary binaural feature set is proposed based on the energy difference and partial interaural coherence between two channel to separate target speech in a very challenging environment with babble noise as background noise and high reverberation. While the reliability of common binaural features of ILD and IPD for source separation decreases in highly reverberant conditions, the proposed feature set provides complementary information about the target speech to the neural network. Furthermore, all binaural features are passed through a 64 channel mel-filtebank for dimension reduction and better frequency selectivity. Different binaural features of ILD, IPD, IC, IC_p, energy difference and normalized energy difference are investigated as an input to DNN based masking speech separation. Comparing the results of Chapter 4 and Chapter 5 indicates the huge improvement of the supervised speech separation algorithms using binaural (spatial) features over monaural for different matched and unmatched testing condition.

In Chapter 6, two different neural network architecture called DCN and DCRN are proposed to solve the generalization issue of supervised speech enhancement methods to unseen noises and unseen speakers in noisy reverberant environment. Most of the current supervised speech enhancement models only deal with background noise; however, the proposed models work in noisy-reverberant environments. The first model (DCN) includes multiple dilated convolution blocks with skip connections to efficiently increase the receptive field of convolutional layers while keeping the number of trainable parameters small. The goal of the dilated convolution blocks with exponentially increasing dilation rate is to generate soft masks on lower level features using multi-level contextual information of previous and future time frames. DCN can be trained to generate either the ideal ratio mask or the target magnitude spectrum. DCN improves the objective speech intelligibility and speech quality significantly in adverse noisy and reverberant environment while generalizing well to unseen noises, unseen speakers and unseen rooms. Therefore, DCN is a noise-speaker independent speech enhancement model with few trainable parameters, which is also computationally more efficient than current supervised speech enhancement models.

Furthermore, the dilated convolutional recurrent network (DCRN) is proposed to adapt DCN for real-time processing applications. DCRN is a causal model that only needs the information about current and previous input time frames to estimate the output. DCRN incorporates both dilated causal convolution blocks and LSTM to exploit short-term context and long-term context for prediction of the output. DCRN is a causal noise-speaker independent model for monaural speech enhancement that improves speech intelligibility and speech quality significantly in noisy-reverberant environment. This model is suitable for real-time processing applications such as ASR and hearing aid devices because of its causality and low number of model parameters.

In this dissertation, different supervised speech separation algorithms (learning model and acoustic features) using deep learning techniques are proposed that operate when a limited number of microphones are available (monaural and binaural case). These models can separate target speech from non-stationary background noise and reverberation with no assumptions on the statistics of the signals. Moreover, the proposed models are independent of the number of micro-

phones (they can operate in under-determined conditions where number of microphones are less than number of sound sources in mixtures). The proposed methods generalize well to different unseen conditions such as untrained speakers, unmatched noises and unseen room simulations.

7.2 Future work

In this dissertation, different approaches were proposed for supervised speech separation in noisy and reverberant environment using deep learning techniques. The majority contribution of this dissertation is the development of different acoustic features for monaural and binaural speech separation and different neural network architectures to improve the generalization of neural network models to unseen environment, unseen speakers and noises. For further improvement of supervised speech separation and real world applications, some areas can be explored as listed below.

- *Multiple speakers and noise separation*. In this dissertation, most of the proposed approaches and acoustic features are used for the case of speech denoising and speech enhancement, where only speech from one speaker is contaminated with background noise. However, in many real acoustic environments, multiple talkers may interfere with the target speaker and the background noise and reverberation are also present. Therefore, extending supervised speech separation models' ability to separate both the interfering speakers, background noise and room reverberation is desirable. For these challenging conditions, multiple stage models can be used to for example separate background noise in one stage and then separate the interfering speakers in following stages.

- *Further evaluation of supervised speech separation models on real recording data*. In this dissertation, all the dataset used for training and testing are simulated. Real-recording dataset can only be used for testing because the supervised speech separation algorithms need the clean signals as training target and the real-recording data is not clean. As a next step, the trained models should be tested with different recorded dataset to show the generalization

of the models to a more realistic condition. Furthermore, it is not possible to evaluate the objective speech quality and objective speech intelligibility of the recorded audio, because the clean recorded audio signal is not available. Therefore, for real-recording evaluation, human subjects should be tested to rate the speech quality and speech intelligibility of the processed signals.

- *Model compression.* The computational cost and power consumption of these supervised speech separation algorithms are very important factors to consider for implementation in portable devices such as phones, headsets and hearing aids. These devices have limited power supply (batteries) and limited memory; therefore, model compression should be used on neural network structure used for supervised speech separation to reduce the number of parameters (such as weights). Therefore, studying techniques for low-computation and low-memory supervised speech separation is essential and next steps for this work.

- *Phase enhancement.* So far, the proposed supervised speech separation algorithms in this dissertation enhanced only the magnitude of noisy-reverberant speech, while using the phase of noisy speech to resynthesize the processed waveform. However, in negative SNRs, the phase of the noisy speech includes more the phase of the background noise than the speech. Therefore, as a next step, the neural networks should be trained to learn both the magnitude and phase of audio signals in highly noisy environment. This can be done by either working on the complex value STFT representation of noisy-reverberant speech or train the network to work directly on the time-domain signals.

References

[1] A. S. Bregman. *Auditory Scene Analysis: The perceptual organization of sound*. Cambridge, MA : MIT Press, 1990.

[2] D. L. Wang. On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi Kluwer, editor, *Speech Separation by Humans and Machines*, pages 181–197. Academic, Norwell, MA, 2005.

[3] Y. Wang, A. Narayanan, and D Wang. On training targets for supervised speech separation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(12):1849–1858, 2014.

[4] J. Chen and D. L. Wang. Long short-term memory for speaker generalization in supervised speech separation. In *International Speech Communication Association (INTERSPEECH)*, pages 3314–3318, 2016.

[5] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions Signal Processing*, 52(7):1830–1847, 2004.

[6] S. Rickard. The duet blind source separation algorithm. In T. W.L̃ee S.M̃akino and H. Sawada, editors, *Blind Speech Separation*, pages 217–241. Dordrecht, The Netherlands, Springer, 2007.

[7] E. C. Cherry. Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America*, 25:975–979, 1953.

[8] J. B. Boldt. *Binary Masking and Speech Intelligibility*. PhD thesis, Aalborg Universitet, 2011.

[9] K. Kokkinakis and P. Loizou. *Advances in Modern Blind Signal Separation Algorithms: Theory and Applications*. San Rafael, CA, USA: Morgan and Claypool, 2010.

[10] S. Pirhosseinloo and F. Almasganj. Discriminative speaker adaptation in persian continuous speech recognition systems. In *Fourth International Conference of Cognitive Science, Procedia Social and Behavioural Sciences, ELSEVIER*, volume 32, pages 296–301, 2012.

[11] S. Pirhosseinloo and S. Javadi. A combination of maximum likelihood bayesian framework and discriminative linear transforms for speaker adaptation. *International journal of information and electronics engineering (IJIEE)*, 2(4):552–555, 2012.

[12] D. B. Ward, R. A. Kennedy, and R. C. Williamson. Constant directivity beamforming. In Michael S. Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 3–17. Springer, 2001.

[13] J. Capon. High resolution frequency-wave number spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

[14] S. Darlington. Linear least-squares smoothing and prediction with applications. *Bell System Technical Journal*, 37:1121–1194, 1952.

[15] M. Er and A. Cantoni. Derivative constraints for broad-band element space antenna array processors. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, (6):1378–1393, 1983.

[16] O. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

[17] O. Frost. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982.

[18] B. R. Breed and J. Strauss. A short proof of the equivalence of lcmv and gsc beamforming. *IEEE Signal Processing Letter*, 9(6):168–160, 2002.

[19] Y. Kaneda and J. Ohga. Adaptive microphone-array system for noise reduction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, (6):1391–1400, 1986.

[20] S. Affes and Y. Grenier. A source subspace tracking array of microphones for double talk situations. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 269–272, 4 1997.

[21] S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions Signal Processing*, 49(8): 1618–1626, 2001.

[22] C. Jutten and J. Hérault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *IEEE Transactions Signal Processing*, 1991.

[23] P. Comon. Independent component analysis a new concept? *IEEE Transactions Signal Processing*, 36(3):287–314, 1994.

[24] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput*, 7(6):1129–1159, 1995.

[25] M Pedersen, J Larsen, U. Kjems, and L. Parra. Convolutive blind source separation methods. In *Handbook of Speech Processing*, pages 1065–1094. Springer, 2007.

[26] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.

[27] D. L. Wang and G. J. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Hoboken, NJ: Wiley/IEEE Press, 2006.

[28] E. De Boer. Synthetic whole-nerve action potentials for the cat. *Journal of the Acoustical Society of America*, 58:1030–1045, 1975.

[29] J Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice. Implementating a gammatone filterbank. Technical report, APU Technical Note. MRC Applied Psychology Unit, Cambridge, 1988.

[30] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. In *Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, pages 269–272, 12 1987.

[31] B. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[32] E. N. Roman, D. L. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114:2236–2252, 2003.

[33] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 2985–2988, 2000.

[34] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Eurospeech*, pages 1009–1012, 2003.

[35] D. L. Wang. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification*, 12(6):332–353, 2008.

[36] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. New York: Wiley, 2001.

[37] G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 1994.

[38] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15:1135–1150, 2004.

[39] G. Hu and D. L. Wang. An auditory scene analysis approach to monaural speech segregation. In E. Hansler & G. Schmidt, editor, *Topics in acoustic echo and noise control*, pages 485–515. Heidelberg, Germany, Springer, 2006.

[40] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, and K. et al. Hiekata. Two-stage blind source separation based on ica and binary masking for real-time robot audition system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS*, pages 2303–2308, 2005.

[41] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of a dominant source signal from mixtures of many sources. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP) , Philadelphia, PA.*, pages 61–64, 2005.

[42] H. Sawada, Araki. S., R. Mukai, and S. Makino. Blind extraction of dominant target sources using ica and time-frequency masking. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 14:2165–2173, 2006.

[43] N. Roman and D. L. Wang. Binaural sound separation for multi source reverberant environments. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP) , Montreal Quebec, Canada.*, pages 373–376, 2004.

[44] J. B Boldt, U. Kjems, M. S. Pedersen, T. Lunner, and D. L. Wang. Estimation of the ideal binary mask using directional systems. In *11th International Workshop on Acoustic Echo and Noise, Seattle, WA.*, 2008.

[45] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Underdetermined blind separation for speech in speech in real environments with sparseness and ica. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP) , Montreal Quebec, Canada.*, pages 881–884, 2004.

[46] D. Kolossa and R. Orglmeister. Nonlinear postprocessing for blind speech separation. In C. G. Puntonet & A. Prieto, editor, *Lecture notes in computer science: 3195. Independent*

component analysis and blind signal separation, pages 832–839. Proceedings of the Fifth International Congress, ICA, Berlin: Springer, 2004.

[47] O. Cappe. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2:345–349, 1994.

[48] H. Erdogan, J. Hershey, S. Watanabe, and J. Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 708–712, 2015.

[49] D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[50] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Adv. in Neural Info. Process. Systems*, 13:556–562, 2001.

[51] P. Smaragdis. Non negative matrix factor deconvolution: extraction of multiple sound sources from monophonic inputs. In *ICA*, pages 494–499, 2004.

[52] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1–12, 2007.

[53] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and spareness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1066–1074, 2007.

[54] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh. Non-negative matrix factorization based compensation of music for automatic speech recognition. In *International Speech Communication Association (INTERSPEECH)*, pages 717–720, 2010.

[55] Y. Wang and D. Wang. A structure-preserving training target for supervised speech separation. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 6148–6152, 2014.

[56] Y. Wang and D. Wang. A deep neural network for time-domain signal reconstruction. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 4390–4394, 2015.

[57] G. J. Mysore and P. Smaragdis. A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 17–20, 2011.

[58] G. J. Mysore, P. Smaragdis, and B. Raj. Non-negative hidden markov modeling of audio with application to source separation. In *LVA/ICA*, 2010.

[59] S. Tamura. An analysis of a noise reduction neural network. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 2001–2004, 1989.

[60] S. Tamura and A. Waibel. Noise reduction using connectionist models. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 553–556, 1989.

[61] F. Xie and D. Van Compernolle. A family of mlp based nonlinear spectral estimators for noise reduction. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 53–56, 1994.

[62] M. L. Seltzer, B. Raj, and R. M. Stern. A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech communication*, 43:379–393, 2004.

[63] G. Kim, Y. Lu, Y. Hu, and P. Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *Journal of the Acoustical Society of America*, 126: 1486–1494, 2009.

[64] K. Han and D. Wang. A classification based approach to speech segregation. *Journal of the Acoustical Society of America*, 132:3475–3483, 2012.

[65] K. Han and D. Wang. Towards generalizing classification based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):168–177, 2013.

[66] N. Roman and D. Wang. Speech segregation based on sound localization. In *International Joint Conference on Neural Networks*, pages 2861–2866, 2001.

[67] N. Roman, D. Wang, and G. J. Brown. A classification-based cocktail-party processor. In *Advances in Neural Information Processing Systems*, 2003.

[68] Y. Wang and D. Wang. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1381–1390, 2013.

[69] Y. Wang and D. Wang. Cocktail party processing via structured prediction. In *Advances in Neural Information Processing Systems*, pages 224–232, 2012.

[70] D. L. Wang and J. Chen. Supervised speech separation based on deep learning: an overview. *arXiv preprint arXiv.1708.07524*, 2017.

[71] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *Journal of the Acoustical Society of America*, 134:3029–3038, 2013.

[72] A. Narayanan and D. Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 7092–7096, 2013.

[73] Y. Jiang, D. Wang, R. Liu, and Z. Feng. Binaural classification for reverberant speech segregation using deep neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(12):2112–2121, 2014.

[74] J. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, London, 1988.

[75] J. Blauert. *Spatial Hearing – The psychophysics of human sound localization*. Cambridge, MA : MIT Press, 1997.

[76] J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Ann. Rev. Psych*, 42:135–159, 1991.

[77] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[78] N. Kayser, S. D. Ewert, J. Anemuller, T. Rohdenburg, V. Hohhmann, and B. Kollmeier. Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Applied Signal Processing*, pages 1–10, 2009.

[79] R. Y. Litovsky, S. H. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *Journal of the Acoustical Society of America*, 106(4):1633–1654, 1999.

[80] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson. Structural composition and decomposition of hrtfs. 2001.

[81] J. S. Bradley, H. Sato, and M. Picard. On the importance of early reflections for speech in rooms. *Journal of the Acoustical Society of America*, 113(6):3233–3244, 2003.

[82] A. J. Watkins and N. J. Holt. Effects of a complex reflection on vowel identification. *Acta Acustica united with Acustica*, pages 532–542, 2000.

[83] J. Lochner and J. Burger. The influence of reflections on auditorium acoustics. *Journal of Sound and Vibration*, 1(4):426–454, 1964.

[84] H. Glyde, J. Buchholz, H. Dillon, S. Cameron, and H. Hickson. The importance of interaural time differences and level differences in spatial release from masking. *Journal of the Acoustical Society of America*, 134:EL 147,EL 152, 2013.

[85] N. I. Durlach and X. D. Pang. Interaural magnification. *Journal of the Acoustical Society of America*, 80:1849–1850, 1986.

[86] G. F. Kuhn. Acoustics and measurements pertaining to directional hearing. In W. A. Yost and G. Gourevitch, editors, *Directional Hearing*, page 3D25. Berlin, Germany: Springer-Verlag, 1987.

[87] N. Kollmeier and J. Peissig. Speech intelligibility enhancement by interaural magnification. *Acta Otolaryngol. Suppl*, 469:215–223, 1990.

[88] N. I. Durlach, B. G. Shinn-Cunningham, and R. M. Held. Supernormal auditory localization. i. general background. *Presence*, 2:89–103, 1993.

[89] B. G. Shinn-Cunningham, N. I. Durlach, and R. M. Held. Adapting to supernormal auditory localization cues i. bias and resolution. *Journal of the Acoustical Society of America*, 103 (6):3656–3666, 1998.

[90] S. Pirhosseinloo and K. Kokkinakis. An interaural magnification algorithm for enhancement of naturally-occurring level differences. In *International Speech Communication Association (INTERSPEECH)*, pages 2558–2561, 2016.

[91] K. Kokkinakis and S. Pirhosseinloo. An algorithm for enhancement of naturally-occurring level differences in bilateral cochlear implants. In *Poster session at 2017 conference on implantable auditory prostheses*, 2017.

[92] IEEE Subcommittee. IEEE recommended practice speech quality measurements. *IEEE Transaction Audio Electroacoustic*, 17:225–246, 1969.

[93] ITU-T P.862. Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders. *ITU-T*, 2001.

[94] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

[95] A. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acoustica*, 86(1):117–128, 2000.

[96] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions Neural Network*, 15(5):1135–1150, 2004.

[97] S. Pirhosseinloo and K. Kokkinakis. Time-frequency masking for blind source separation with preserved spatial cues. In *International Speech Communication Association (INTERSPEECH)*, pages 1188–1192, 2017.

[98] J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *Journal of the Acoustical Society of America*, 62(4):912–915, 1977.

[99] C. Faller and J. Merimma. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America*, 116(5):3075–3089, 2004.

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

[101] T. H. Falk, C. Zheng, and W. Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2006.

[102] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets. Signal-based performance evaluation of dereverberation algorithms. *Journal of Electrical and Computer Engineering*, 2010(7):1–5, 2010.

[103] A. Tsilfidis and J. Mourjopoulos. Blind single-channel suppression of late reverberation based on perceptual reverberation modeling. *Journal of the Acoustical Society of America*, 129(3):1439–1451, 2011.

[104] A. Westermann, J. M. Buchholz, and T. Dau. Binaural dereverberation based on interaural coherence histograms. *Journal of the Acoustical Society of America*, 133(5):2767–2777, 2013.

[105] B. Rakerd and W. M. Hartmann. Localization of sound in rooms. iii: Onset and duration effects. *Journal of the Acoustical Society of America*, 80(6):1695–1706, 1985.

[106] F. Rosenblatt. *Principles of neural dynamics*. New York: Spartan, 1962.

[107] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel distributed processing*, pages 318–362. Cambridge, MA: MIT Press, 1986.

[108] Y. LeCun and et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(8):541–551, 1989.

[109] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Network*, 61: 85–117, 2015.

[110] J. P. J. Werbos. Backpropagation through time: What it does and how to do it. *Proc. IEEE*, 78:1550–1560, 1990.

[111] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pages 1310–1318, 2013.

[112] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comp*, 9:1735–1780, 1997.

[113] K. Cho, B. V. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder decoder for statistical machine translation. *arXivpreprint arXiv.1406.1078*, 2014.

[114] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[115] T. Tieleman and G. Hinton. Lecture 6.5 - rmsprop ,coursera: Neural networks for machine learning. technical report. Technical report, 2012.

[116] M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[117] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[118] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[119] D. A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations (ICLR)*, 2016.

[120] M. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, volume 30, 2013.

[121] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *ArXiv e-prints*, 2017.

[122] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[123] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[124] K. Paliwal, K. Wójcicki, and B Shannon. The importance of phase in speech enhancement. *Speech Communication*, 53(12):465–494, 2010.

[125] D. S. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 24(3):483–492, 2016.

[126] Y Xu, J. Du, L. R. Dai, and C. H. Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letter*, 21:65–68, 2014.

[127] Y Xu, J. Du, L. R. Dai, and C. H. Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 23:7–19, 2015.

[128] H. Hermansky. Perceptual linear predictive (plp) analysis for speech. *The Journal of The Acoustical Society of America*, 87:1738–1752, 1990.

[129] T. H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2:578–589, 1994.

[130] Y. Shao, S. Srinivasan, and D. L. Wang. Robust speaker identification using auditory features and computational auditory scene analysis. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 1589–1592, 2008.

[131] J. Chen, Y. Wang, and D. L. Wang. A feature study for classification based speech separation at low signal-to-noise ratios. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(12):1993–2002, 2014.

[132] M. Delfarah and D. L. Wang. Features for masking-based monaural speech separation in reverberant conditions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25 (5):1085–1094, 2017.

[133] S. Pirhosseinloo and J. S. Brumberg. A new feature set for masking-based monaural speech separation. In *52nd Asilomar Conference on Signals, Systems, and Computers*, pages 828–832, 2018.

[134] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[135] A. Varga and H. J. Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12:247–251, 1993.

[136] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small room acoustics. *The Journal of The Acoustical Society of America*, 65:943–950, 1979.

[137] E. Habets. Room impulse response generator. http://home.tiscali.nl/ehabets/rir generator.html, 2010.

[138] C. Hummersone, R. Mason, and Brookes. T. Dynamic precedence effect modeling for source separation in reverberant environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1867–1871, 2010.

[139] Institute of Sound Recording (IoSR). Simulated room impulse responses software. https://www.surrey.ac.uk/department-music-media/research/institute-sound-recording-iosr/iosrsoftware-and-digital-resources, 2012.

[140] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and

V. Zue. Timit acoustic-phonetic continuous speech corpus. Technical report, Natl. Inst. Standards Technol. Tech. Rep. NISTIR 4930, Gaithersburg, MD, USA, 1993.

[141] J. Thiemann, N. Ito, and E. Vincent. The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of The Acoustical Society of America*, 133:3591–3591, 2013.

[142] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[143] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, volume 2, pages 749–752, 2001.

[144] J. Ma, Y. Hu, and P. C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of The Acoustical Society of America*, pages 3387–3405, 2009.

[145] X. Zhang and D. L. Wang. Deep learning based binaural speech separation in reverberant environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 25(5): 1075–1084, 2017.

[146] T. May. Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 26(2):406–414, 2018.

[147] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Audio, Speech, and Language Processing*, (4):357–366, 1980.

[148] H. Wierstorf, M. Geier, A. Raake, and S. Spors. A free database of head-related impulse response measurements in the horizontal plane with multiple distances. In *130th Convention Audio Engineering Society*, 2011.

[149] M Jeub, M. Schafer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *16th International Conference Digital Signal Processing*, pages 1–5, 2009.

[150] J. Chen and D. L. Wang. Long short-term memory for speaker generalization in supervised speech separation. *The Journal of The Acoustical Society of America*, 141(6):4705–4714, 2017.

[151] K. Tan, J. Chen, and D. L. Wang. Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 27:189–198, 2019.

[152] K. Tan and D. L. Wang. A convolutional recurrent neural network for real-time speech enhancement. In *International Speech Communication Association (INTERSPEECH)*, pages 3229–3233, 2018.

[153] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

[154] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[155] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[156] S. Pirhosseinloo and J. S. Brumberg. Monaural speech enhancement with dilated convolutions. In *International Speech Communication Association (INTERSPEECH)*, pages 3143–1192, 2019.

[157] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *https://arxiv.org/abs/1512.03385*, 2015.

[158] K. Han, Y. Wang, and D. L. Wang. Learning spectral mapping for speech dereverberation. In *IEEE International Conference Acoustic Speech Signal Processing (ICASSP)*, pages 4628–4632, 2014.

[159] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[160] FreeSFX. Freesfx. http://www.freesfx.co.uk/, 2017.

[161] Freesound. Freesound. http://freesound.org/, 2015.

[162] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.

[163] S. Pirhosseinloo and J. S. Brumberg. Dilated convolutional recurrent neural network for monaural speech enhancement. In *53rd Asilomar Conference on Signals, Systems, and Computers*, 2019.