IMPERIAL

Binaural Enhancement of Speech for Hearing Assistive Devices

Vikas Tokala

A thesis submitted in fullment of requirements for the degree of Doctor of Philosophy of Imperial College London

Supervisors - Patrick A. Naylor, Mike Brookes
Speech and Audio Processing Laboratory
Communications and Signal Processing Research Group
Department of Electrical and Electronic Engineering
Imperial College London
6th August 2025

Copyright Declaration

Statement of Originality

I declare that this thesis and the research it contains are the product of my own work under the guidance of my thesis supervisor, Prof. Patrick A. Naylor and Mike Brookes. Any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. The material of this thesis has not been submitted for any degree at any other academic or professional institution.

Abstract

HUMANS naturally hear with two ears, known as binaural hearing, which enhances spatial awareness and speech intelligibility, particularly in noisy environments. While monaural speech enhancement has advanced, it often fails to preserve interaural cues, limiting sound localization and listener comfort. Binaural speech enhancement improves intelligibility while maintaining spatial perception, making it vital for hearing aids, communication systems, and AR/VR devices.

This thesis investigates binaural speech enhancement techniques that preserve spatial cues while improving speech intelligibility in noise. A method extends monaural mask-assisted enhancement to binaural speech using STOI-optimal masks and a feed-forward DNN to improve weighted-STOI scores. To avoid artefacts from applying TF masks directly in the STFT domain, Optimally Modified-LSA is proposed, enhancing perceptual quality. Recent neural network approaches have shown promising results for speech enhancement in both time and time-frequency (TF) domains. Most TF-domain methods enhance only magnitude, using noisy phase for reconstruction. Complex-valued networks that jointly estimate magnitude and phase have outperformed real-valued approaches for monaural speech enhancement.

This thesis introduces an end-to-end binaural enhancement framework based on complex convolutional neural networks with an encoder-decoder structure. Two models are proposed: one incorporating a complex self-attention transformer block, the other using a complex recurrent block. Both estimate complex ratio masks for each ear channel. A custom loss function balances spatial preservation, intelligibility, and noise reduction, outperforming existing methods like binaural TasNet and STOI-optimal masking in isotropic noise. Further, two multichannel binaural enhancement methods are explored: a beamformer-guided complex neural network and an end-to-end multichannel encoder-decoder network. Both leverage multiple microphones in modern binaural devices, pro-

viding significant intelligibility and noise reduction gains while preserving spatial cues, validated through simulations, real-world recordings, and MUSHRA listening tests. An end-to-end model for sound source localization is also developed to assess spatial cue preservation, showing strong performance in noisy conditions compared to conventional localization methods.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Prof. Patrick Naylor and Mike Brookes for their exceptional guidance and for giving me the incredible opportunity to pursue my PhD at the **Speech and Audio Processing (SAP) Lab** at **Imperial College London**. Their unwavering support and encouragement throughout this journey have been instrumental, in making these PhD years the most enriching and rewarding of my career. I am profoundly grateful to the **European Unions Horizon 2020 research and innovation programme** for funding the **SOUNDS project**, of which this work is a part. This project has provided me with invaluable resources, opportunities, and collaborations that have shaped the course of my research. I extend my heartfelt thanks to my co-supervisors, Dr. Jesper Jensen from Oticon A/S and Aalborg University, and Prof. Simon Doclo from Carl von Ossietzky Universität Oldenburg, for their guidance, expertise, and support. The secondments at their institutions were pivotal in advancing my knowledge and skills, and I am deeply indebted to them for their mentorship. My sincere gratitude also goes to Prof. Toon Van Waterschoot, Aldona Niemiro-Sznajder, and Randall Ali from KU Leuven, for their unwavering support and encouragement to all researchers within the SOUNDS project. Their advice and dedication have left an indelible impact on my work. I am deeply humbled and grateful to Prof. Stefan Goetze and Prof. Dan Goodman for dedicating their time to reviewing and evaluating my work. Their insightful feedback has significantly enriched my technical understanding and perspective, both directly and indirectly.

The SOUNDS project has been an incredible journey, not only in terms of research but also in fostering lifelong friendships. I am grateful to all the researchers and supervisors in the project for their collaboration and camaraderie. In particular, I would like to thank Anslem Lohmann, James Brooks-Park, Pia Porysek, Kaspar Müller, Jesper Brunnström, Paul Didier, José Miguel Cadavid Tobón, Mohammad Bokaei, Ximei Yang, Vasudha

Sathyapriyan, Bilgesu Cakmak, and Matthias Blochberger for their support and for the unforgettable moments during our workshops. A special thanks go to Eric Grinstein and Francesco Nespoli, my fellow ESRs from Imperial College. Their friendship, collaboration, and unwavering support have made this journey even more remarkable, and I cannot imagine it without them.

During my time at Imperial College London, I have been fortunate to work alongside many brilliant friends and colleagues in the SAP Lab and the Communications and Signal Processing (CSP) Group. I am deeply grateful to Emilie dOlne and Vincent Neo for being wonderful office mates, always lending a listening ear to my complaints and helping me overcome challenges. I would also like to express my gratitude to my colleagues from the SAP Lab: Giulia Sanguedolce, Aidan Hogg, Alastair Moore, Pierre Guiraud, Yuxi Zhang, Daniel Jones, and Simon McKnight. Additionally, I extend my thanks to Xinze Lyu, Marek Hilton, and Ziang Liu from the CSP Group for their support and insightful discussions.

Beyond academia, the Imperial College Volleyball Club (ICVC) has been an integral part of my time at Imperial, and I cannot imagine my journey without it. Being part of the team has given me memories of a lifetime, and I am incredibly grateful to all my teammates for making this experience unforgettable. A special mention goes to Dylan Tan and Augusto Del Zotto for giving me the opportunity to be part of the team. Being part of the team has given me memories of a lifetime.

I must also express my deepest gratitude for the unwavering love and support I receive from back home in India. I am forever grateful to my parents, Dhanapal and Swarupa, and my brother Sachin for their constant love and encouragement. Additionally, I am lucky to have incredible childhood friendsKaushik, Venu, Anirudh, Pruthvi, Akhilesh and Roopakwho have been my support system throughout this PhD journey.

Finally, I am deeply appreciative of everyone who contributed, directly or indirectly, to this PhD journey. Your support, encouragement, and belief in me have been my greatest source of inspiration. Thank you.

Contents

List of acronyms

A

AI articulation index

ANN artificial neural network

ASA auditory scene analysis

ASR automatic speech recognition

ATF acoustic transfer function

B

BCCRN Binaural Complex Convolutional Recurrent Network

BCCTN Binaural Complex Convolutional Transformer Network

BiTasNet binaural TasNet

BN batch normalization

BRIR binaural room impulse response

BSOBM binaural STOI-optimal masking

BTE behind the ear

C

CASA computational auditory scene analysis

CBP complex-valued back-propagation

CCC cross correlation coefficient

CED convolutional encoder-decoder

CIRM complex ideal ratio mask

CNN convolutional neural network

CQS continuous quality scale

CRM complex ratio mask

CRN convolutional recurrrent network
CVNN complex-valued neural network
D

DBSTOI deterministic binaural STOI
DFT discrete Fourier transform
DNN deep neural network
DOA direction of arrival
DSP digital signal processing
E

EC equalisation cancellation
ELU exponential linear unit
ERB equivalent rectangular bandwidth
ERM empirical risk minimization
ESTOI extended STOI
ETF elementary transcendental function
F

FCIM fixed cylindrical isotropic MVDR
fwSegSNR frequency-weighted segmental SNR
G

GCC Generalized Cross-Correlation
GCC-PHAT Generalized Cross-Correlation with Phase Transform
GFCIM guided FCIM
GMM Gaussian Mixture Model
GMVDR guided MVDR
GOMVDR guided OMVDR
GRU gated recurrent unit
H

HATS head and torso simulator
HRIR head related impulse response
HRTF head-related transfer function

HSWOBM high-resolution stochastic WSTOI-optimal binary mask
I

IACC interaural cross-correlation coefficient
IBM ideal binary mask
IC interaural coherence
IDFT Inverse Discrete Fourier Transform
ILD interaural level difference
IPD interaural phase differences
IRM ideal ratio mask
iSNR input SNR
ISPD Interaural Signal Phase Difference
ISTFT inverse STFT
ITD interaural time difference
ITF interaural transfer function
L

LC local criterion
LMS least mean square
LReLU leaky rectified linear unit
LSA log spectral amplitude
LSTM long short-term memory
M

MBCCTN multichannel binaural complex convolutional transformer network
MBSTOI modified binaural STOI
MISO multiple inputs single output
MLP multi-layer perceptrons
MUSHRA MUltiple Stimuli with Hidden Reference and Anchor
MVDR minimum variance distortionless response
MWF multichannel Wiener filter
N

NCM Noise Covariance Matrix
NLP natural language processing

O

OM-LSA optimally-modified log spectral amplitude
OMVDR Oracle MVDR
P

PESQ perceptual evaluation of speech quality
POLQA perceptual objective listening quality assessment
PReLU parametric rectified linear unit
R

ReLU rectified linear unit
RMS root mean square
RMSProp root mean square propagation
RNN recurrent neural network
RReLU randomized rectified linear unit
RTF relative transfer function
RVNN real-valued neural network
S

SAA sample average approximation
SELU scaled exponential linear unit
SGD stochastic gradient descent
SI-SNR scale invariant SNR
SII speech intelligibility index
SNR signal-to-noise ratio
SOBM STOI-optimal binary mask
SPP speech presence probability
SRM spatial release from masking
SRP steered response power
SRP-PHAT SRP with phase transform
SRT speech reception threshold
SSN speech shaped noise
STFT short time Fourier transform
STI speech transmission index

STOI short-time objective intelligibility measure

T

TCN temporal convolutional network

TF time-frequency

TNN transformer neural networks

V

VSSNR voiced-speech-plus-noise to noise ratio

W

WGN white Gaussian noise

WSTOI weighted STOI

List of Symbols

y	noisy speech in time domain
x	clean speech in time domain
v	noise in time domain
n	discrete time index
Y	noisy speech in time-frequency domain
X	clean speech in time-frequency domain
V	noise in time-frequency domain
k	frequency index
ℓ	time index
$\mathcal{M}_{IBM}$	ideal binary mask (IBM)
ν	local criterion (LC)
$\hat{S}$	enhanced speech in time-frequency domain
$\mathcal{M}_{IRM}$	ideal ratio mask (IRM)
η	tunable parameter for IRM
$\mathcal{M}_{IRM}$	complex ideal ratio mask (CIRM)
r	real part
i	imaginary part
J	total number of octave bands
K_j	lowest short time Fourier transform (STFT) frequency bin within the frequency band j
$\mathbf{x}$	clean speech modulation vector
$\mathbf{y}$	clean speech modulation vector
$\|\cdot\|$	Euclidean norm
V^p	communication model production noise

W	communication model output
R	communication model production noise
N_e	communication model production noise
I	mutual information
a	Kneser-Ney free parameter
H_0	speech absence hypothesis
H_1	speech presence hypothesis
G	LSA gain function
p	speech presence probability (SPP)
q	a priori probability of speech absence
υ	a posteriori SNR
ξ	a priori SNR
α	smoothing parameter
Ψ	number of features for binaural STOI-optimal masking (BSOBM)
w	triangular window
μ	feature subset for BSOBM
f_s	sampling frequency
$\digamma$	equivalent rectangular bandwidth (ERB) scale triangular window function
Φ	ERB transformation
Λ	triangular function
B	target high-resolution stochastic WSTOI-optimal binary mask (HSWOBM)
$\mathcal{J}$	loss function for BSOBM
H_{n+1}	output of the transformer for $n+1^{th}$ hidden state
H_n	output of the encoder for n^{th} hidden state
M	estimated complex ratio mask (CRM)
$\mathcal{L}$	loss function of the convolutional neural networks (CNNs)
$\mathcal{L}_{SNR}$	Signal-to-noise ratio (SNR) loss

$\mathcal{L}_{STOI}$	Short-time objective intelligibility measure (STOI) loss
$\mathcal{L}_{ILD}$	Interaural level difference (ILD) loss
$\mathcal{L}_{IPD}$	Interaural phase differences (IPD) loss
E	energy of the speech signal
$\mathcal{T}$	threshold of energy for mask estimation
N_{fft}	FFT length
$\mathbf{R_V}$	Noise Covariance Matrix (NCM)
$\mathbf{d}$	relative transfer function (RTF) vector of the speech source
$h_e(n)$	impulse response of the internal ear noise

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

THE objectives of this work are to research digital signal processing and machine learning methods applied to audio signal processing for binaural speech enhancement. Humans have a natural ability to hear with two ears, which is known as Binaural Hearing. Binaural speech enhancement has been established in recent years as the state-of-the-art approach for enhancement in hearing aid [3], communication, and augmented/virtual reality devices [4,5]. Binaural speech improves the natural listening experience by reproducing speech as it would be heard in real-life environments by preserving those spatial aspects of sound that are perceptually significant. In recent times teleconferencing and videoconferencing for communication is used extensively. In such situations where the speaker is not physically present with the listener, binaural speech can provide a more natural and realistic representation of the speaker's voice which can make it easier to follow and reduce fatigue over long periods. This is particularly useful for hearing aid users who often report physical and mental fatigue after being in situations that require prolonged listening in noisy acoustic environments [6]. Binaural speech helps listeners to accurately perceive the direction of the source and localize the source in space [7]. Accurate localization of sound sources can be helpful in various applications such as automobile navigation, operation of industrial machinery, and emergency response situations to name a few. The location of the ears makes them receive different signals from the same sound source. Acoustic cues and neural mechanisms help humans localize sound sources in a 3-dimensional space and

also help in reducing background noise from our auditory perception [8]. Binaural speech helps to bridge the gap between recorded audio and the way humans naturally perceive sound in real life. Binaural speech can offer a more immersive and realistic listening experience across various applications, enhancing communication and entertainment by capturing and reproducing spatial audio cues. Section 1.3 provides the technical review of binaural speech and spatial cues, and Sec. 1.4 introduces binaural speech enhancement.

1.2 Research Goals

To develop, evaluate, and advance binaural speech enhancement techniques, ensuring improved speech intelligibility, noise suppression, and spatial information preservation.

1.2.1 Original Contributions

To the best of the authors knowledge, the novel contributions of this thesis are:

1. **Extension of STOI-Optimal Masking to Binaural Speech Enhancement:**
   - Development of a novel application of STOI-optimal masking for binaural speech scenarios, leveraging high-resolution stochastic WSTOI-optimal binary mask (HSWOBM) and optimally-modified log spectral amplitude (OM-LSA) techniques and validated through simulation experiments.
2. **Binaural speech enhancement using CVNNs:**
   - Proposal of an end-to-end binaural speech enhancement framework using complex convolutional neural networks with convolutional encoder-decoder (CED) architectures.
   - Introduction of two CED architectures for binaural speech enhancement:
     (a) A complex valued self-attention-based transformer integrated in the CED architecture.
     (b) A complex valued long short-term memory (LSTM) based Recurrent neural network (RNN) integrated in the CED architecture.

- Development of a novel loss function emphasizing the preservation of spatial information, alongside improving speech intelligibility and noise suppression.
- Training and evaluation of the models to estimate complex ratio mask (CRM) for binaural channels in the time-frequency domain, achieving superior performance compared to existing methods.
- Evaluation of the proposed methods using a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test.

3. **Development of Multichannel Binaural Speech Enhancement Methods:**
   - Introduction of two methods for multichannel binaural speech enhancement:
     (a) A beamformer-guided complex convolutional neural network.
     (b) An end-to-end CED based network for multichannel binaural speech enhancement.
   - Introduction of the size-reduced versions of the proposed models with minimal reduction in performance.
   - Demonstration of the methods capability to enhance binaural speech intelligibility, reduce noise, and preserve spatial cues across isotropic noise and real-world multichannel group conversation recordings.
   - Evaluation of the proposed methods using a MUSHRA test.
4. **End-to-End Binaural Speech Source Localization:**
   - Development of a lightweight convolutional recurrrent network (CRN) for accurate binaural localization of speech sources in the frontal azimuthal plane under noisy and reverberant conditions.
   - Execution of listening tests to align algorithmic performance with human auditory localization and intelligibility in noisy conditions, bridging the gap between machine-based and human-centric evaluations.
   - Establishment of the model as a potential tool for assessing interaural cue preservation in binaural speech enhancement methods.

1.2.2 Publications

During the course of this work, the following have been published or are being considered for publication:

- V. Tokala, M. Brookes, and P. A. Naylor, Binaural Speech Enhancement Using STOI-optimal Masks, in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Germany, Sep. 2022, pp. 15.
- V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, Binaural Speech Enhancement using Deep Complex Convolutional Recurrent Networks, in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, USA, 2023.
- V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, Binaural Speech Enhancement using Deep Complex Convolutional Transformer Networks, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Seoul, South Korea, 2024.
- V. Tokala, E. dOlne, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, Multichannel binaural speech enhancement using deep complex convolutional transformer networks, *submitted to IEEE/ACM Trans. Audio, Speech and Language Proc. (TASLP)*, 2025
- V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, Binaural Localization of Speech in Noise, in *Accepted to Forum Acusticum*, Spain, 2025.

1.3 Binaural Speech

It was first proposed by Lord Rayleigh [9] in his 'Duplex Theory' that the ability to localize sound was mainly due to two acoustic cues. The difference in level between the two ears is known as the interaural level difference (ILD) and the difference in arrival time between the two ears is known as interaural time difference (ITD). ILDs and ITDs are the primary interaural cues helpful in localizing and dereverberating sounds, boosting the perceived loudness, and improving intelligibility [7]. The primary interaural cues operate in complementary frequency ranges of sound. Lord Rayleigh suggested that ITDs were

used for low-frequency sound localization and ILDs was used for high-frequency sound localization. When a high-frequency sound source is located on one particular side of the listener, the intensity of sound reaching the ear present on the other side will be reduced. This reduction in intensity at the ear distant from the sound source is known as Head Shadow [10]. Low-frequency sounds that have wavelengths greater than the size of the head do not produce a head shadow. However, as the two ears are separated in space, there is a delay between the times the sound arrives at one ear closer to the source compared to the one that is distant, and this causes the time difference in arrival known as the ITD. Due to ITDs, there will be a phase difference between the sound wave that reaches each ear, which is referred to as interaural phase differences (IPD). Considering the size of an average adult human head the maximum ITD that can be generated is around 700 μs [11] and therefore ITDs are generally useful at frequencies below about 1.5 kHz. The ILDs are most pronounced at frequencies above 1.5 kHz because the head is large enough compared to the wavelength of the incident sound wave causing substantial attenuation causing the aforementioned head shadow. The actual measured values of the interaural cues have been found to empirically depend on the angle of arrival of the sound source, frequency, and distance from the source [12]. Another binaural measure, interaural cross-correlation coefficient (IACC) is used to predict source lateralization [13–15] and is based on the cross correlation coefficient (CCC) that is calculated for the signals received at each ear and measures the similarity between the two signals. The offset time period used in the calculation of IACC is varied over a range that would ensure that the maximum ITD caused by the physical separation of human ears is encompassed, which is typically around ± 1 ms [16]. Interaural coherence (IC) is calculated as the maximum of the IACC for values of lag within the ITD range. Studies have shown that IC is an important factor in spatial perception [17] and models have been developed to estimate it using other binaural cues [18, 19]. The human auditory system is capable of using these binaural cues to suppress interfering sounds that come from different directions other than the target sound. This phenomenon is referred to as spatial release from masking (SRM) [20]. SRM enables us to have 10 dB improvement in speech reception threshold (SRT) by performing innate speech enhancement through spatial filtering over speech and noise sources [21, 22].

The fissures of the outer ears, or pinnae, cause further spectral colouration of the signals which helps the listener better perceive the elevation of the target source and mitigate

binaural front-back confusion [23–26]. In [27, 28], the authors followed the instrumentation techniques developed in [29] and used probe microphones in the ear to measure the transfer function from the sound source to the ear in anechoic environments, which is now commonly referred to as the head-related transfer function (HRTF). HRTF in the time domain is referred head related impulse response (HRIR).

Binaural hearing helps listeners achieve better localization in enclosed and reverberant spaces due to a phenomenon known as the *precedence effect*. Despite the fact that there are multiple reflections that are caused by the walls of an enclosed room, listeners can determine the location of the source quite accurately. Human hearing has a perceptual weighting system for the directional cues arriving at the ears where the cues from the direct sound are given a higher weighting than those from the reflected sound [30].

1.4 Binaural speech enhancement

Given the importance of the spatial cues in binaural speech, enhancement algorithms should be designed to preserve these acoustic cues which hold the information critical to spatial auditory perception. Early enhancement methods for hearing devices used beamformers that were applied bilaterally. Using the multiple microphones present on the device, two multiple inputs single output (MISO) beamformers were independently used to produce a dual channel output from the left and right channel microphone arrays [3]. This method corresponds to applying two single-channel or monoaural beamformers at each ear separately. Studies indicated that this technique did not preserve the interaural cues and hearing aid users did not have an improvement in localization of sound sources [3, 31, 32]. The same effect can be observed if monoaural speech enhancement methods are used bilaterally [32]. For designing binaural algorithms for hearing aids, a communication channel had to be established for the left and right channel arrays to communicate [3]. Multiple studies have been carried out to extend monoaural beamformers and enhancement algorithms to work with binaural audio [3, 22, 33–39] using interaural transfer function (ITF) estimation. In most cases, the beamformers collapse the entire sound scene in the direction of the target source towards which it was steered thereby using all spatial information. Listening tests with hearing aid users showed that the

listeners perceived both speech and noise in the same direction [31, 32]. Due to limited processing power in hearing aids and latency limitations, complex enhancement algorithms could not be deployed historically. Binaural enhancement solutions using multichannel Wiener filter (MWF) which spatially separate the target source and residual noise were proposed [33, 34], however, they could not preserve the correct speech and noise cues simultaneously [22, 32]. Some algorithms that were proposed for binaural enhancement introduced a trade-off parameter between noise reduction and cue preservation [35–38].

Binaural audio with multiple speech and noise sources can make binaural speech enhancement a challenging task while preserving the spatial cues of all the sources in the soundscape. The number of hearing aid users is steadily increasing [40] and the use of virtual/augmented reality devices is gaining popularity [41, 42] which makes binaural speech enhancement an important challenge to address. This project explores digital signal processing (DSP) and machine learning-based models to perform binaural speech enhancement.

1.5 Overview of the thesis structure

This work is structured with Ch. 1 giving the motivation and supported by the literature review in Ch. 2. Chapter 3 presents the use of short-time objective intelligibility measure (STOI) optimal masking for binaural speech enhancement. Chapter 4 presents the use of deep complex convolutional networks for binaural speech enhancement. Two complex convolutional networks with transformer and recurrent neural network blocks are studied. In Ch. 5, binaural speech enhancement using multichannel microphone data will be presented. Chapter 6 will introduce a binaural source localization method that is used as a metric to analyze the preservation of binaural cues by speech enhancement algorithms.

Chapter 2

Literature Review

2.1 Introduction

This chapter provides an overview of existing speech enhancement techniques followed by a review of the speech quality and intelligibility metrics. The fundamentals of deep neural network (DNN)-based speech enhancement are discussed along with the introduction of complex neural networks. Individual layers of the DNNs used in this work to build the enhancement networks are discussed.

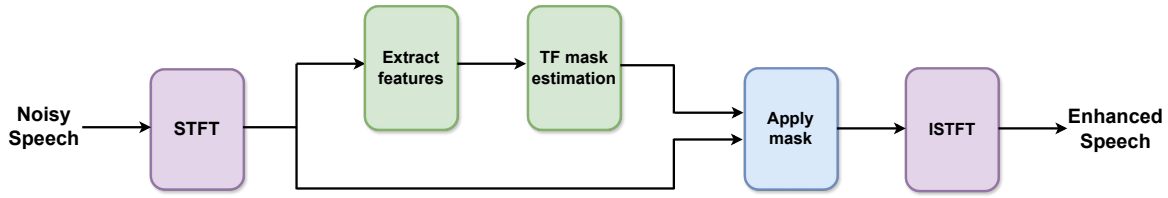


Figure 2.1: Block diagram of a typical Time-frequency (TF) mask-based enhancer.

2.2 Mask-based speech enhancement methods

2.2.1 Overview

Studies have shown that the intelligibility of noisy speech can be improved by applying a two-dimensional multiplier or 'mask' in the TF-domain [43–47]. Figure 2.1 shows the block diagram of a typical mask-based enhancer and multiple mask-based enhancement techniques have been proposed that have a similar structure. The masks described in Sec.2.2.3 and Sec.2.2.4 are estimated using oracle information, i.e., knowledge of the clean speech or noise signal. In a typical speech enhancement system, the mask is estimated solely from the noisy signal without oracle information. Masks estimated with oracle information are commonly used as training targets. However, different techniques vary in how the TF mask is estimated and the features utilized. The estimated mask is then applied to the noisy speech in the TF-domain, and the enhanced spectrogram is transformed back into the time domain.

2.2.2 Signal model

The noisy time domain signal is given by

$$y[n] = x[n] + v[n], \tag{2.1}$$

where y is the noisy speech, x is the clean speech, v is the noise and n is the discrete time-index. The complex short time Fourier transform (STFT) coefficients of these signals are given by $Y(k,\ell)$, $X(k,\ell)$, and $V(k,\ell)$ with (k,ℓ) the frequency and time frame indices respectively.

2.2.3 Ideal binary mask (IBM)

The ideal binary mask (IBM) is a binary-valued mask that was introduced in [48] for computational auditory scene analysis (CASA) which aims to separate a target signal from a signal that contains a mixture of multiple interfering sources. Studies have shown that

IBM can improve speech intelligibility [44, 46, 49] by suppressing by the use of a masking process, the TF bins which are dominated by noise and directing the listener's hearing attention to bins which are dominated by speech. This is based on the principle of auditory scene analysis (ASA) proposed in [50] which explains the human auditory capability to segregate acoustic signals to multiple sound source streams. The IBM computation technique proposed in [51, 52] assigns a mask value of 1 to each TF bin for which the target energy exceeds the set threshold of interference or noise energy, and 0 otherwise. The IBM, $\mathcal{M}_{IBM}$, defined as

$$\mathcal{M}_{IBM}(k,\ell) = \begin{cases} 1 & \text{if } |X(k,\ell)|^2 > \nu |Y(k,\ell)|^2 \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where ν is the local criterion (LC), which determines the signal-to-noise ratio (SNR) threshold above which the mask value is assigned to 1. This mask is applied to the noisy speech magnitude, $|Y(k,\ell)|$, in the TF domain to obtain the magnitude of enhanced speech $\hat{S}$ given by

$$|\hat{S}(k,\ell)| = |Y(k,\ell)| \mathcal{M}_{IBM}(k,\ell). \tag{2.3}$$

Figure 2.2(a) shows the magnitude spectrogram of the clean speech signal, Fig 2.2(b) shows the magnitude spectrogram of the noisy speech signal, and Fig 2.3(a) shows the IBM generated to be applied to the enhanced signal. The noisy signal was generated by adding white Gaussian noise (WGN) at a SNR of 5 dB to the clean speech signal. The IBM was generated using a ν (LC) of -5. Figure 2.3(b) shows the spectrogram of the enhanced signal after applying the IBM to the noisy speech spectrogram multiplying the respective TF bins of the spectrograms.

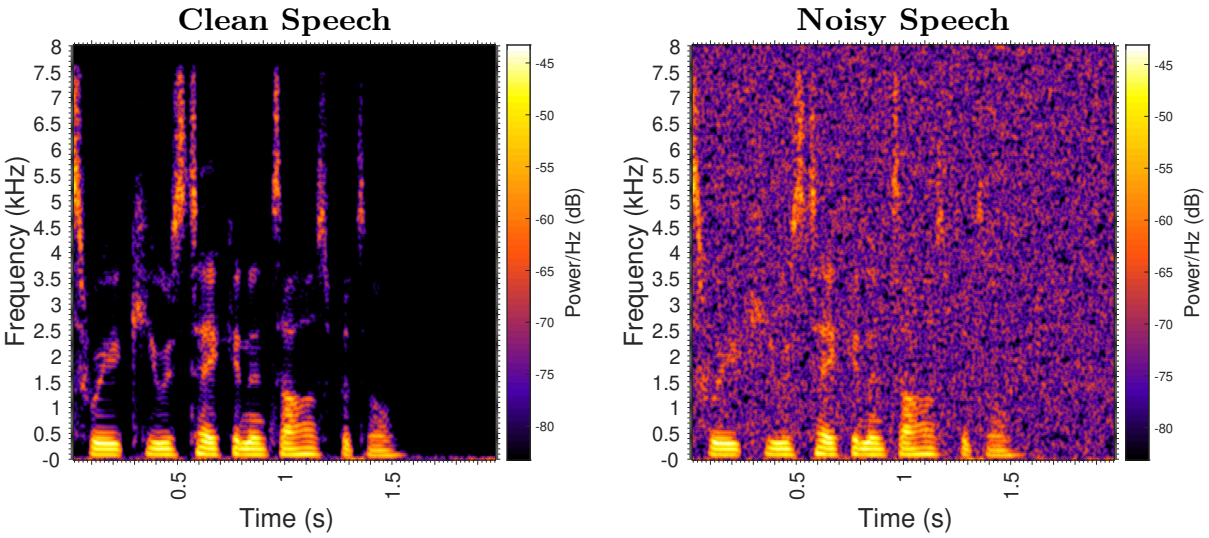


Figure 2.2: (a) Magnitude spectrograms of the clean speech signal and (b) the noisy speech signal.

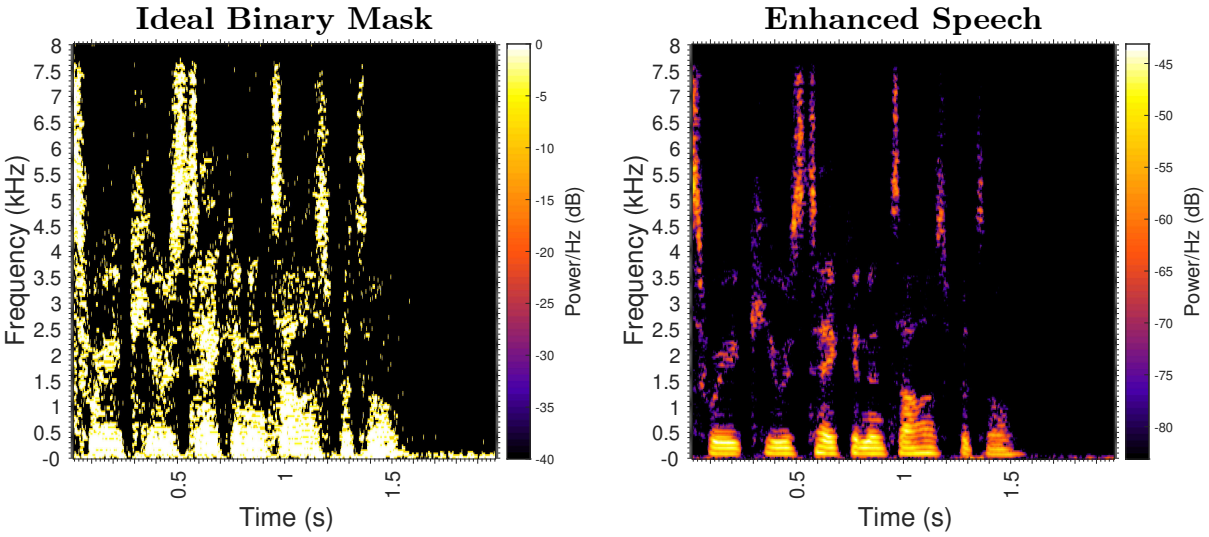


Figure 2.3: (a) Spectrogram of the estimated ideal binary mask and (b) magnitude spectrogram of the enhanced speech signal.

2.2.4 Ideal ratio mask

A ratio mask or a soft mask has continuous values corresponding to the probability of TF bins being speech or noise-dominated and the values typically range from 0 to 1

[53–55]. Ratio masks have previously been used in automatic speech recognition (ASR) applications [56, 57]. The ideal ratio mask (IRM), $\mathcal{M}_{IRM}$ is given by

$$\mathcal{M}_{IRM}(k,\ell) = \left(\frac{|X(k,\ell)|^2}{|X(k,\ell)|^2 + |V(k,\ell)|^2}\right)^{\eta}, \tag{2.4}$$

where η is a tunable parameter that can be used to scale the mask [54]. When η is set to 1, and the speech and noise are assumed to be uncorrelated stationary stochastic processes [53], (2.4) gives the gain of the Wiener filter. In [54], the authors state that η set to 0.5, (2.4) resembles the square root Wiener filter and provides the gains for the optimal unbiased estimation of the power spectrum [58] and found that $\eta = 0.5$ was the best choice for IRM estimation. Equation (2.4) can be rewritten using the SNR in the TF-domain as

$$\mathcal{M}_{IRM}(k,\ell) = \left(\frac{SNR(k,\ell)}{SNR(k,\ell)+1}\right)^{\eta}. \tag{2.5}$$

Figure 2.4(a) shows the spectrogram of the IRM computed for the noisy signal shown in Fig 2.2(b) using $\eta = 0.5$. Figure 2.4(b) shows the enhanced signal obtained by applying the IRM to the noisy spectrogram. Several studies [59–61] have used IRMs as training targets for training DNNs. Commonly, both IBM and IRM are applied to enhance the magnitude of the unprocessed noisy speech and use the noisy phase to reconstruct the signal. When the phase of noisy speech with negative SNRs is used for signal reconstruction, the phase of noisy speech is influenced more by the phase of background noise rather than the phase of the target speech [62, 63] leading to lower speech intelligibility and musical artefacts in the enhanced signal lowering the quality of the speech.

The complex ideal ratio mask (CIRM) was introduced in [63] where complex-valued spectrograms were used to derive the IRM. The CIRM, $\mathcal{M}_{IRM}$ is given by

$$\mathcal{M}_{CIRM}(k,\ell) = \frac{|X(k,\ell)|}{|Y(k,\ell)|}\cos(\theta) + i\frac{|X(k,\ell)|}{|Y(k,\ell)|}\sin(\theta) \tag{2.6}$$

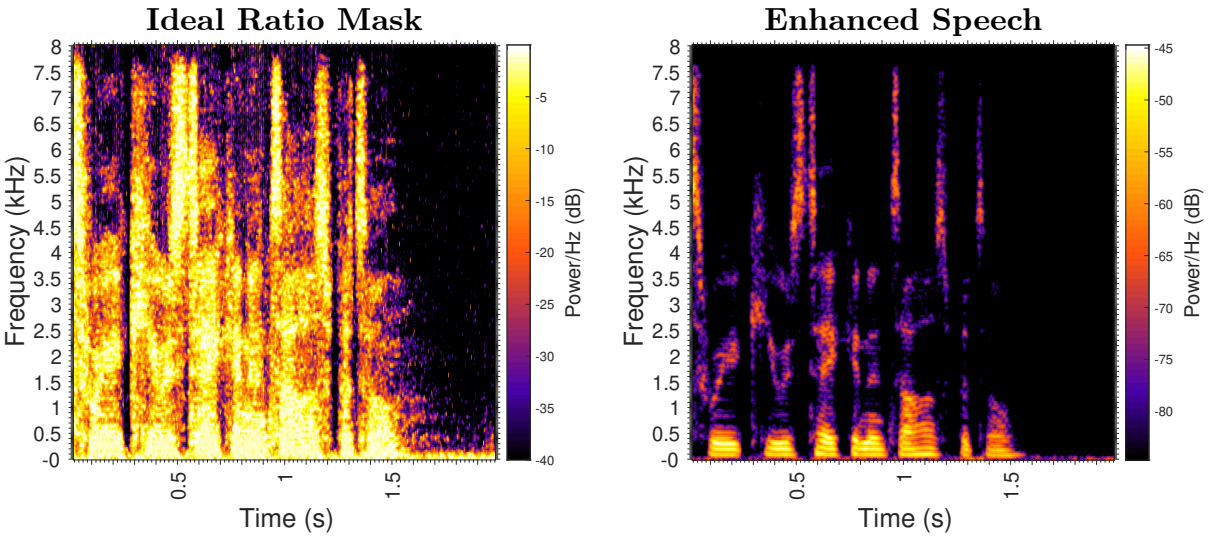


Figure 2.4: (a) Spectrogram of the estimated IRM and (b) magnitude spectrogram of the enhanced speech signal using IRMs.

where $\theta = \theta_Y - \theta_X$, $\theta_Y = \angle(Y(k,\ell))$, and $\theta_X = \angle(X(k,\ell))$. The enhanced speech is obtained by applying the estimated complex mask $\mathcal{M}_{CIRM}(k,\ell) = (M_r + jM_i)$ to the complex-valued noisy signal $(Y_r + jY_i)$ in the TF domain (omitting k and ℓ indices),

$$\hat{S}_r + j\hat{S}_i = (M_r + jM_i)\,(Y_r + jY_i)\,, \tag{2.7}$$

where r and i indicate the real and imaginary parts. The resulting definition of the CIRM is

$$M_r + jM_i = \frac{\hat{S}_r + j\hat{S}_i}{Y_r + jY_i} = \frac{Y_r\hat{S}_r + Y_i\hat{S}_i}{Y_r^2 + Y_i^2} + j\frac{Y_r\hat{S}_i - Y_i\hat{S}_r}{Y_r^2 + Y_i^2}. \tag{2.8}$$

The IRM has values in the range [0,1] however, $\hat{S}_r, \hat{S}_i, \hat{Y}_r$ and $\hat{Y}_i \in \mathbb{R}$ therefore M_r and $M_i \in \mathbb{R}$. With this, the CIRM can have large real and imaginary values in the range $[-\infty, \infty]$. These large values make the CIRM estimation difficult and complicate its use as a training target for DNNs and the authors in [63] used hyperbolic tangent function to

compress the values into the range $[0, 1]$.

2.3 Deep neural networks

Artificial neural networks (ANNs) are made up of 'neurons' designed in a structure inspired by the biological functioning of human brain neurons. The neuron is the basic building block of a neural network, sometimes also referred to as the linear unit [64]. Figure 2.5 shows the schematic of a typical artificial neuron. X is the input to the neuron, Y is the output of the neuron, w is the weight, and b is the bias of the model. Normally the bias is independent of the input and is determined by the model architecture. The output of the neuron is given by the equation,

$$Y = wX + b, \tag{2.9}$$

and resembles the slope of a straight line [65]. The fundamental single-layer neural network is called a perceptron whose output is determined by a single activation method followed by the neuron.

Figure 2.6 shows the structure of a perceptron [64, 65]. The inputs to the perceptron are denoted by $[X_1, \ldots, X_n]$, the applied bias is denoted by b, and the weights are denoted by $[w_0, \ldots, w_n]$. The inputs and the bias are multiplied by their respective weights, and a weighted sum of the inputs is obtained. Then, an activation function is applied, and in some cases is also followed by a normalization step. Both activation and normalization

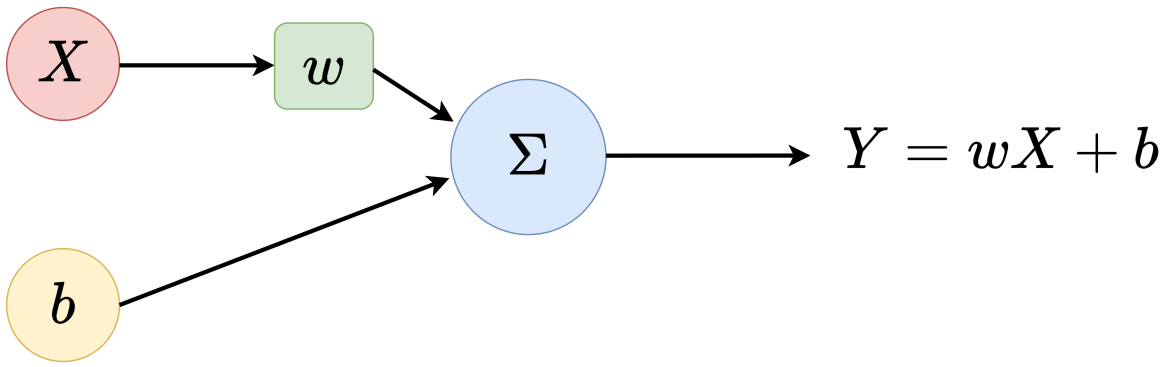


Figure 2.5: Schematic of a typical neuron.

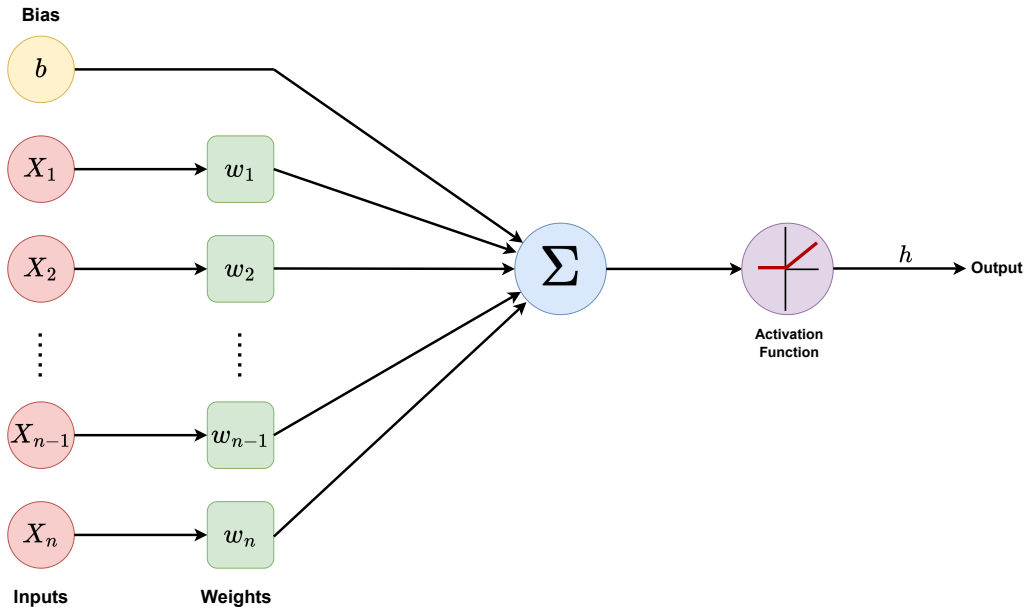


Figure 2.6: Structure of a perceptron layer.

methods are explained in the following sections. A deep learning network is made up of multiple layers of perceptrons and each layer of perceptrons is fed by the previous layer. The topology of a network configures how the layers are connected and the flow of information through the network. This leads to various configurations of DNNs and can be used for different applications like classification, regression, etc.

2.4 Complex-valued neural networks

The earliest use of complex-valued neural networks can be traced to the ADALINE machine which is a single-neuron, single-layer trainable network based on the Rosenblatt perceptron [66]. The optimal weights for ADALINE were derived using least mean square (LMS) and stochastic gradient descent (SGD) and were extended to the complex domain in [67]. The SGD was computed for the real and imaginary parts. Wirtinger calculus [68] was used in [69] to genralize the SGD such that the gradient was computed using complex variables instead of using real-valued components. Complex-valued neural networks (CVNNs) were first comprehensively introduced in [70] and the back-propagation [71] and the network parameters [72] were extended to be complex-valued. In [73, 74], it is shown

that CVNNs are highly compatible with wave-related applications such as speech, radar etc., as both magnitude and phase information can be extracted during training. Analysis operations involving complex numbers have sometimes treated real and imaginary components independently as real numbers and it is shown in [74] that applying the same concept to CVNNs would not be accurate. Assuming the CVNNs to be a two-dimensional real-valued neural network (RVNN) limits the degree of freedom of the CVNNs at the synaptic weighting due to the complex multiplication operation [74, 75]. Studies have also shown that using complex-valued back-propagation (CBP) helps the network avoid singular points during training where the gradients are ambiguous [72,76,77]. CVNNs are said to have better data representation than RVNNs when dealing with a complex-valued problem as RVNNs treat the real and imaginary parts of the data independently which doubles the number of learnable parameters for the same task [78].

In [75], the authors emphasize there are two main reasons for the use of complex-valued neural networks.

1. In applications such as speech and audio processing, wireless communication, radar, etc., the data inherently or by design is complex-valued, where there is a strong correlation between the real and imaginary parts of the signal and the use of CVNNs would be more appropriate. The Fourier transform is a type of linear transformation where there is a direct relationship between multiplying a signal by a constant in the time domain and scaling the magnitude of the signal in the frequency domain. A circular rotation of the signal in the time domain is equivalent to a phase shift in the frequency domain which implies there is a statistical correlation between the real and imaginary parts of the signal. This relationship may be disrupted when RVNNs are used, particularly in the frequency domain.

2. If the relationship between the magnitude and phase of the signal to the learning objective is known *a priori*, the use of complex-valued models is more suited for the purpose as they impose relevant constraints on the complex data than a RVNN.

2.4.1 Types of CVNNs

CVNNs were introduced as split-CVNNs, where the real and imaginary parts of the data were treated as separate inputs and the network was trained using the real and imaginary

parts (or magnitude and phase) independently using RVNNs [74, 77] as shown below.

$$\{z_1, z_2\} = \{(a_1, b_1), (a_2, b_2)\} \tag{2.10}$$

$$\{z_1, z_2\} = \{(r_1, \phi_1), (r_2, \phi_2)\} \tag{2.11}$$

where $z_{1,2} \in \mathbb{C}$, $z = a + ib$, in which a and b are the real and imaginary components, r and ϕ are the magnitude and phase of the complex-valued input respectively.

Split-CVNNs were further divided into two categories: split-CVNNs with both activation function and weights being real-valued and the ones with real-valued activation function and complex-valued weights. The split-CVNN has a higher input dimension and learnable parameters which in turn increases the overall network size. As the network weights are updated by using real-valued gradients inaccuracies in output approximation were observed due to phase distortion [79]. Several studies have shown that using a real-valued activation function and complex weights can overcome the phase distortion problem [80–83].

Alternatively, fully-CVNNs to tackle complex-valued data with complex-valued activation and weights were introduced in [84]. Since then several studies have used fully-CVNNs in various applications such as image processing [85, 86], telecommunication [87, 88], and speech enhancement [63,89,90]. Fully-CVNNs have increased computation complexity and have difficulty in choosing an appropriate activation function that is complex-differentiable or holomorphic [72] for back-propagation.

2.5 Activation layers

In comparison to the neuron-based model in our brains, the activation function decides which neurons have to fire and carry the signal to the adjacent neurons. Similarly in an ANN, the activation function decides which outputs and what level from the previous cells or layers can be supplied as the input to the subsequent layers. Activation functions play a vital role in neural networks by introducing non-linearity that allows ANNs to learn

complex non-linear patterns and relationships in the data. The absence of activation layers in the network architecture would reduce the network to a linear model which would render it useless against complex non-linear problems [91]. Activation functions help restrict the output of the neuron or perception where the input to the activation layer is given by (2.9) which is unbounded and may reach very high levels of magnitude. Furthermore, activation functions facilitate the propagation of gradients through backpropagation, enable efficient optimization for SGDs during training and reduction of network complexity. As ANNs uses gradient descent to train, it is crucial that the designed activation functions are differentiable everywhere [92]. Activation functions are broadly classified into two classes [93].

- **Piecewise linear activation function**: A function which has a limited number of linear segments connected at specific intervals, providing piecewise linear approximations of nonlinear transformations. Example: Rectified linear unit (ReLU) and its variants.
- **Locally quadratic activation function**: A locally quadratic function is any smooth non-linear activation function with a nonzero second derivative. Example: Swish, exponential linear unit (ELU).

Designing a suitable activation function for CVNNs is a challenging task. Liouville's theorem states that any bounded 'entire function' that is analytic everywhere in the complex plane is a constant. The activation functions must capture the nonlinearity in the input data in the complex domain and using real-valued conventional bounded and analytic functions such as *tanh* or *sigmoid* might not be viable. Most of the activation functions for CVNNs are designed by choosing one of the two properties i.e., either analyticity or boundedness. Complex activation functions are mostly modified versions of the real-valued functions which are bounded except for the ReLU function. The complex activation functions are broadly classified into split activation function and fully complex activation function [75, 77]. The split activation function has the real and imaginary parts independently computed while the fully complex activation function treats them as a single entity. The split activation function was proposed in previous studies [70, 71] where

the real-valued function is adopted for the complex domain as

$$f_C(z) = f_R(a) + if_R(b), \tag{2.12}$$

where $z = a + ib$, $f_R(.)$ is the real-valued activation function and f_C is the complex split activation function. Not many fully complex activation functions are available as they are computationally expensive and the constraints imposed by Liouville's theorem [75, 77]. In [94], fully complex activation functions based on several elementary transcendental functions (ETFs) such as $\sin, \sinh, \arccos$ etc., were introduced. The choice of activation function is based on the type of data, network architecture and goal of the network. In the following sections, some commonly used activation functions are reviewed. The real-valued activation functions are first introduced followed by their complex variants.

2.5.1 Rectified linear unit (ReLU) and its variants

ReLU is an activation function that was inspired by neuroscience research. The activation function of neurons in the human brain can be approximated by a rectifier function [91,95]. ReLU was introduced in [95] for restricted Boltzmann machines. In recent years ReLU and its modifications are the most commonly used activation function and have been shown to outperform other functions like sigmoid, hyperbolic tangent and softmax [89, 96–99]. The definition of ReLU for an input signal x is given by,

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}, \tag{2.13}$$

and the gradient of ReLU is given by

$$\text{ReLU}^{'}(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \tag{2.14}$$

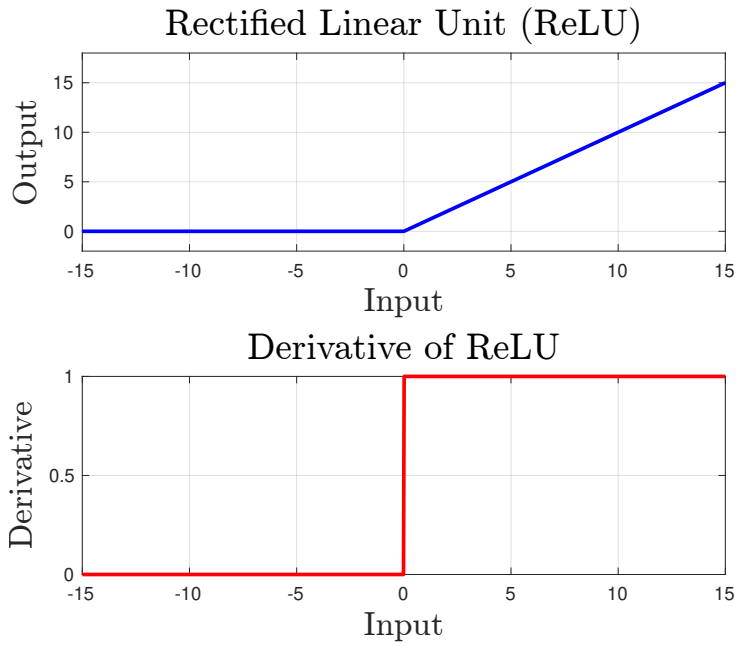


Figure 2.7: ReLU function and it's derivative.

which is a constant for inputs $x > 0$ as ReLU is a non-saturated function. Figure 2.7 shows the plot of the typical ReLU activation function and its derivative. Studies have shown that ReLU has the following advantages over other activation functions [91, 93, 96, 100].

- Computation is cheaper as there are no exponential functions.
- Faster convergence compared to other saturating activation functions like sigmoid or tanh in terms of gradient descent training.
- Higher efficiency in learning as the sparse representation of the data for the network is easily achieved.
- ReLU has a constant gradient which helps the network evade being stuck in a local minima and also resolve the 'vanishing gradient effect' that commonly occurs with sigmoid or tanh.

Given the numerous advantages of ReLU, the major drawback is that all neuron units with negative outputs are set to zero and it is unlikely that these neurons will be activated again. This is a special case of the vanishing gradient problem and is called the 'dying ReLU effect' [101, 102]. ReLU function is not recommended for use in RNNs as the

function is unbounded and is not suitable for long sequences for RNN units which may result in an exploding gradient [93]. To mitigate the above drawbacks several variants of ReLU have been proposed.

Leaky ReLU, Randomized ReLU and Parametric ReLU

The hard right saturation nature of ReLU meant that some neurons of the network would be permanently disabled during training. In [103] the authors proposed the leaky rectified linear unit (LReLU) which is given by

$$\text{LReLU}(x) = \begin{cases} x & if\ x \geq 0 \\ 0.01x & if\ x < 0 \end{cases}, \tag{2.15}$$

and its derivative is

$$\text{LReLU}^{'}(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0.01 & if\ x < 0 \end{cases}. \tag{2.16}$$

Figure 2.8 shows the plot of LReLU and it can be observed that for inputs that are negative a small gradient is allowed instead of completely shutting of the neurons. This allows for the revival of the neurons if necessary and the problem of zero gradients or 'dying ReLU effect' is avoided. Experiments from [103] showed that even though sparsity is lost due to the 'leakage' introduced for the negative inputs, the performance of the neural network in terms of learning ability became significantly robust. The parametric rectified linear unit (PReLU) was proposed in [98] where the constant 0.01 in the LReLU was replaced by a learnable parameter α, and is defined as

$$\text{PReLU}(x) = \begin{cases} x & if\ x \geq 0 \\ \alpha x & if\ x < 0 \end{cases}. \tag{2.17}$$

It is shown in [98] that PReLU significantly outperforms ReLU with lower training error

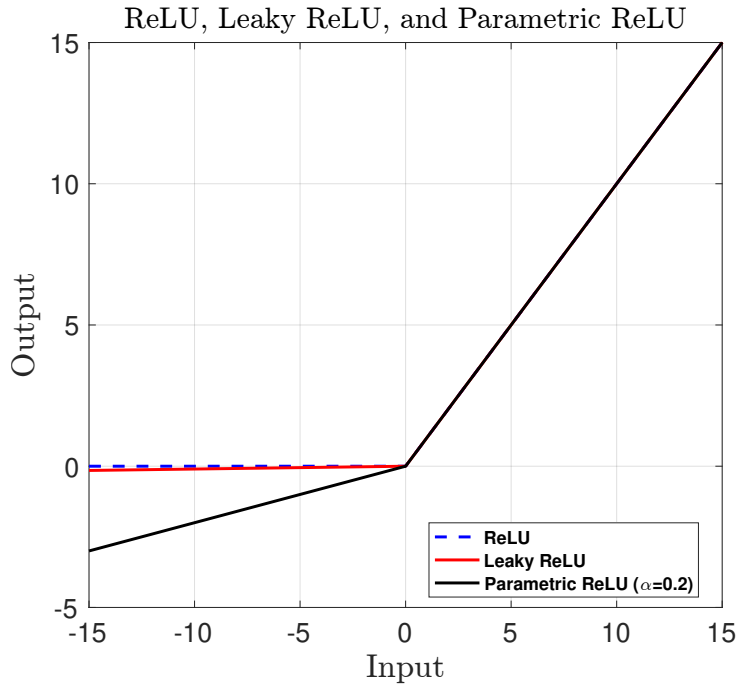


Figure 2.8: Comparison between ReLU, LReLU and parametric rectified linear unit (PReLU) functions.

and faster convergence rate. It is also verified that the introduction of a learnable parameter in the activation function does not result in the model overfitting during training. In [104], randomized rectified linear unit (RReLU), which used a randomized asymmetric weight initialisation, was proposed. The learnable parameter from PReLU is replaced by a randomised parameter chosen from a predefined distribution during training and fixed during testing. In specific conditions and for a given task, RReLU is shown to have better performance compared to its other ReLU counterparts if the right distribution of parameters is chosen [105]. The complex-valued ReLU and its variants are designed based on the split methodology described earlier [106].

$$\mathrm{ReLU}(z) = \mathrm{ReLU}(\Re(z)) + i\mathrm{ReLU}(\Im(z)) \tag{2.18}$$

where z is the complex-valued input. Several modifications to optimise ReLU for use with complex-valued networks have been proposed in the literature based on the application [107, 108]. In [109], the modified ReLU, which is a fully complex activation function, was

proposed. The modified ReLU is given by

$$\text{modReLU(z)} = \text{ReLU}(|z| + \alpha)\frac{z}{|z|}, \tag{2.19}$$

where α is the learnable parameter and *modReLU* does not satisfy the Cauchy-Reimann aligns and is not holomorphic. *ModReLU* has been shown to perform better than the split ReLU activation function when used with unitary RNNs [109, 110].

2.5.2 Sigmoid

Sigmoid or Logistic function is a common activation function which is bounded and translates the input in the range of $(-\infty, +\infty)$ to $(0, 1)$. It is a nonlinear continuous function with a smooth derivative and is given by

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2.20}$$

and its derivative is given by

$$\sigma^{'}(x) = \frac{e^{-x}}{(1 + e^{-x})^2}. \tag{2.21}$$

Figure 2.9 shows the plot of the sigmoid function and its derivative. The sigmoid function is only used in shallow neural networks and due to the bounded nature of the function it is often used in the output layers of deep networks [111, 112]. The complex-valued sigmoid function is designed based on the split complex-valued activation architecture [75,77] and is given by

$$\sigma(z) = \sigma(\Re(z)) + i\sigma(\Im(z)). \tag{2.22}$$

Due to the soft saturation effect of the Sigmoid function, the vanishing gradient problem

is often observed in networks with more than 5 layers [113].

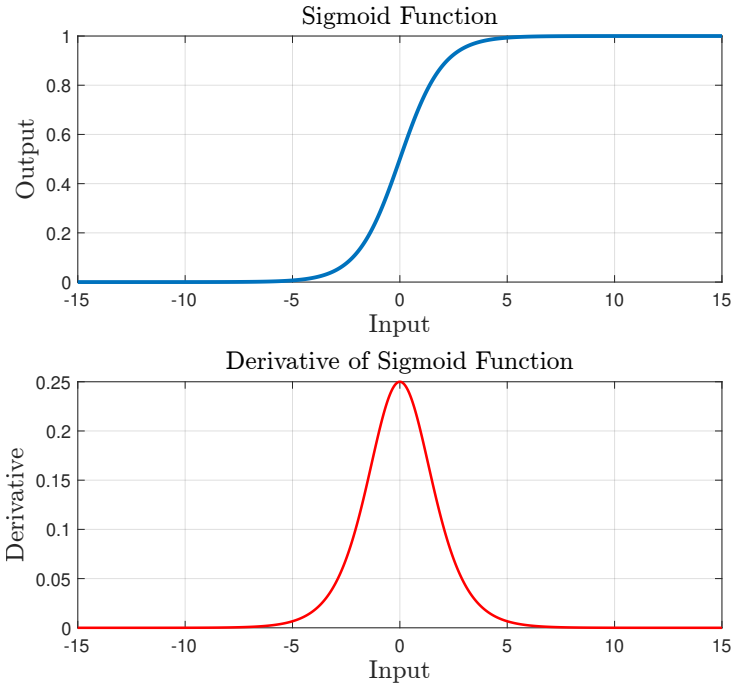


Figure 2.9: Sigmoid function and its derivative.

2.5.3 Hyperbolic tangent (tanh)

The hyperbolic tangent (tanh) has a bounded structure similar to the sigmoid function which limits the output to (-1,1). The tanh function is defined as

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{2.23}$$

$$\tanh x = \frac{2}{1 + e^{-2x}} - 1. \tag{2.24}$$

The tanh function can also be derived from the sigmoid function ($\sigma(x)$) defined in (2.20) by

$$\tanh x = 2\sigma(2x) - 1. \tag{2.25}$$

The derivative of tanh is given by

$$\tanh' x = 2\sigma'(2x) - 1 = \frac{4e^{-2x}}{(1 + e^{-2x})^2}. \tag{2.26}$$

The plots of tanh function and its derivative are shown in Fig. 2.10. The tanh function

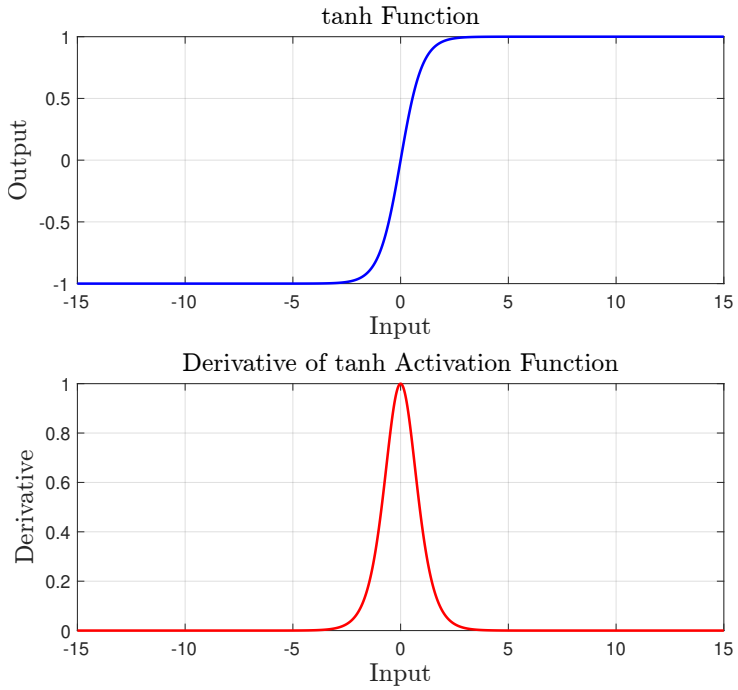


Figure 2.10: Hyperbolic tangent (tanh) function and its derivative.

is a monotonic continuous function that is differentiable everywhere and is symmetric about the origin as shown in Fig. 2.10. The derivative of the tanh is steeper compared to the sigmoid function and is shown to converge faster and have lower classification errors compared to networks that used the sigmoid function [77, 113, 114]. The complex-valued

tanh function is designed by splitting the real and imaginary parts similar to ReLU and sigmoid, given by

$$\tanh(z) = \tanh(\Re(z)) + i \tanh(\Im(z)). \tag{2.27}$$

Similar to the sigmoid function, due to the soft saturation and bounded nature of tanh, the vanishing gradient problem is persistent in deep networks. In [115], the authors linearly scale the tanh function and show that it mitigates the vanishing gradient problem and achieves faster convergence.

2.5.4 Other activation functions

Various activation functions have been proposed depending on the network architecture and the application. Table 2.1 lists some commonly used activation functions and their derivatives. The real-valued activation functions that take x as input. The complex cardioid activation function takes the complex-valued z as the input.

Activation Function	**Equation**	**Derivative**
Exponential linear unit (ELU) [116]	$\max(0,x) + \min(0, \alpha(e^x - 1))$	$\begin{cases} 1 & \text{if } x \geq 0 \\ \alpha e^x & \text{if } x < 0 \end{cases}$
Scaled exponential linear unit (SELU) [117]	$\gamma(\max(0,x) + \min(0, \alpha(e^x - 1)))$	$\gamma \begin{cases} 1 & \text{if } x \geq 0 \\ \alpha e^x & \text{if } x < 0 \end{cases}$
Swish [118]	$\frac{x}{1+e^{-x}}$	$\frac{e^x(x+e^x+1)}{(e^x+1)^2}$
Mish [119]	$x \cdot \tanh(\ln(1+e^x))$	$\frac{e^x \cdot (4e^{2x}+4e^x+e^{3x}+6x^2+4x+6)}{(e^x+2)^2 \cdot (e^{2x}+2e^x+1)^2}$
Complex Cardioid [120]	$z \cdot \frac{1}{2}(1+\cos(\angle z))$	$-\frac{-i}{4}\sin(\angle z)\frac{z}{\lvert z \rvert}$

Table 2.1: Activation functions and their derivatives

2.6 Normalization

Training DNNs becomes complicated by the fluctuating distribution of inputs in each layer throughout the training process, influenced by modifications in previous layer parameters. This phenomenon, known as internal covariate shift [121], hampers training efficiency, necessitating lower learning rates, meticulous parameter initialization, and posing challenges in training models with saturating nonlinearities. In [121], the authors introduced the concept of batch normalization as an integral part of the network architecture to avoid the internal covariate shift. The normalization was performed for 'each training mini-batch' and it allows the use of higher learning rates, random initialization and in some cases elimination of the dropout layer. Model architectures with batch normalization were shown to converge 14 times faster [121]. DNNs converge faster during training if the inputs are whitened or linearly transformed to have zero mean and unit variances [114]. As whitening each layer input is computationally costly, the authors in [121] proposed the batch normalization (BN)-transform which is inserted in the network architecture and can represent the identity transform. The batch normalization is given by

$$\hat{x} = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta, \tag{2.28}$$

where $\hat{x}$ is the batch normalized input, $E[.]$ is the expected value, Var[.] is the variance, γ is the scaling factor and β is the shifting factor. In practice, during training BN-transform is applied to each dimension of the input data with γ and β programmed to be learnable parameters. Complex batch normalization was proposed in [110] by treating it as 2D-vector whitening by scaling the data by the square root of their variances along the principal components. This is done by multiplying the zero-centred data by the inverse square root of the covariance matrix. Computationally this is very costly during training and studies have shown that applying batch normalization by splitting the real and imaginary parts has a marginal effect on the accuracy and convergence of the network [89, 110, 122, 123]. To maximize the effect of batch normalization, additional steps such as increasing the learning rate, removing dropout, shuffling training examples, accelerating the learn decay rate, and reducing L_2 weight regularization can be implemented in the

network architecture and training [121].

2.7 Optimization

During training, neural networks compare the model output with the desired output and the difference is back-propagated back into the model to repeatedly modify the weights of the network until the model converges. Several optimization algorithms have been proposed to accelerate and efficiently compute the weights. Some commonly used optimization algorithms will be briefly introduced below.

1. *Stochastic gradient descent (SGD)*

   Gradient descent was originally proposed by Cauchy in 1847 and the improved SGD was proposed in [124, 125]. In machine learning the SGD is used to solve the empirical risk minimization (ERM) or sample average approximation (SAA) to minimize the loss function [126]. SGD starts with initialization by predetermined or random model parameters which represent the weights and biases of the neural network. At each iteration, the gradient of the loss function is computed with respect to the model parameters. The parameters of the model are updated by taking a small step in the opposite direction of the gradient direction and this step size is controlled by the learning rate parameter. Unlike traditional gradient descent, SGD induces randomness by computing the gradient using a subset or batch of the dataset. This stochasticity helps SGD evade the local minima, introduces noise in the optimization process which helps prevent the model overfitting and improves the generalization of the model [126].

2. *Root mean square propagation*

   Root mean square propagation (RMSProp) is a popular adaptive gradient-based optimization algorithm proposed in [127]. Studies have shown that for certain applications like computer vision RMSProp outperformed SGD and Adagrad [128,129]. Similar to SGD, RMSProp computes the gradients of the loss function using mini-batches of the training dataset. Instead of updating the model based on the current gradient parameters, RMSProp maintains a moving average of squared gradients of every parameter computed using a decay rate parameter. RMSProp has an adaptive

learning rate for each model parameter or weight based on the ratio of the current gradient to the root mean square (RMS) of the past gradients. Parameters with large gradients have their learning rates decreased, while parameters with small gradients have their learning rates increased. The adaptive learning rate of RMSProp helps tackle the problem of slow convergence and model oscillations observed in traditional gradient-based optimization algorithms [127, 128].

3. *Adam*

   An efficient stochastic optimization method named Adam was proposed in [130]. Adam only requires first-order gradients with smaller memory requirements compared to other commonly used optimization methods. The name Adam is said to be derived from 'adaptive moment estimation'. Adam is designed by combining the advantages of AdaGrad [131] and RMSProp [127]. AdaGrad works well with sparse gradients and RMSProp handles online and non-stationary conditions efficiently [130]. Adam combines the advantages of both AdaGrad and RMSProp by maintaining two moving averages of gradients: the first moment (mean) and the second moment (uncentered variance). The two moving average variables are initialized to zero and the gradient is computed for the loss function for each iteration. The parameters are updated based on a combination of the first and second gradient moments. The adaptive learning rate for each parameter is computed as a function of both the mean and variance of the gradients. Bias correction is performed to mitigate the effect of zero initialization in the first few iterations. Adam optimization has been used in various types of neural networks in a range of applications such as image processing, speech and audio processing, classification, computer vision etc. [132–135]

Some other commonly used optimization methods are AdaDelta [136], Nesterov accelerated gradient [137], Strongly-Convex (SC) AdaGrad, SC-RMSProp [128].

2.8 Feed forward dense networks

A feed-forward network is a fundamental architecture of DNNs for classification and regression tasks [138]. Figure 2.11 shows a diagram of a feed-forward network, and the

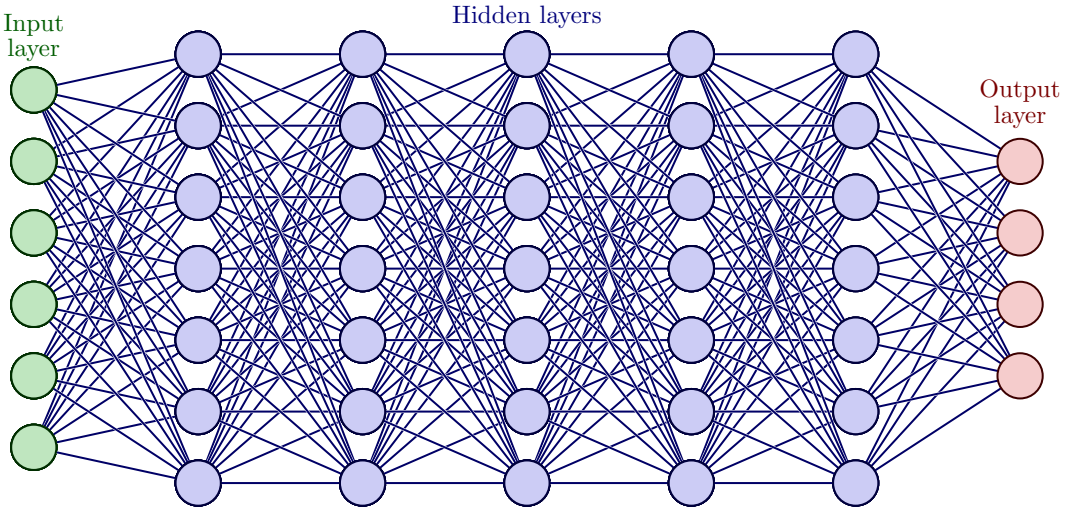


Figure 2.11: Diagram of a feed-forward deep neural network.

basic structure consists of an input layer, one or more hidden layers, and an output layer. Each neuron in a given layer is connected to all the neurons in the subsequent layer. The information flows in one direction from the input layer to the output layer through all the hidden layers. Hence this type of network is known by the term 'feed-forward' and the layers are called 'fully connected dense' layers. Each neuron in the network is defined by (2.9) and the weights and bias values of these neurons are referred to as the network or model parameters which are trained. Modifying the basic neuron equation for the feed-forward network is given by

$$Y_i = w_i h_{i-1} + b_i, \tag{2.29}$$

where Y_i is the output of the i^{th} layer, h_{i-1} is the transformation applied by the hidden layers at the $(i-1)^{th}$ layer and b_i is the bias at the i^{th} layer. In practice, the feed-forward network includes a non-linear activation function after the output layer. The complex-valued feed-forward network can be implemented by using complex input data, weights and biases. The generalization of CVNNs for feed-forward networks was studied in [74] and found that CVNN had a smaller generalization error.

2.9 Recurrent networks

Recurrent neural networks (RNNs) are a class of neural networks which are designed to handle sequential data by introducing connections between individual units across time steps. RNNs are therefore known as auto-associative or feedback networks and they are extensively used in applications involving time series and sequential information. RNNs are designed to have directed cycles allowing them to maintain internal state and capture temporal dependencies [139]. The basic equation for RNN is given by

$$h_n = f(W_{hh} \cdot h_{n-1} + W_{xh} \cdot x_n + b_h), \tag{2.30}$$

where n is the discrete time step, h_n is the hidden state at the time step n, x_n is the input at the time step n, W_{hh} is the weight matrix for the recurrent connections, W_{xh} is the weight matrix for the input connections, b_h is the bias vector for the hidden units and $f(\cdot)$ is the activation function. A major drawback of standard RNNs is the difficulty in capturing long-term dependencies as the signal length increases leading to the vanishing gradient problem [140]. To deal with this problem of long-term dependencies and vanishing gradient, the LSTM cell was introduced [141, 142] by using a 'gates' in the cell. The LSTM cell is designed to have 3 gates i.e., input gate, forget gate and output gate.

Figure 2.12 shows the architecture of a LSTM cell. LSTMs use gated mechanisms to control the flow of information. The aligns that govern these gates are briefly introduced here. The input gate (i_n) determines how much new information is allowed to be added to the current cell state c_n and is given by

$$i_n = \sigma(W_{xi} \cdot x_n + W_{hi} \cdot h_{n-1} + b_i) \tag{2.31}$$

where x_n is the current input, h_{n-1} is the previous hidden state, $\sigma(\cdot)$ is the sigmoid function and W is the weight matrix where the subscripts indicate the associated with the input gate (i), forget gate (f), current input (x), hidden state (h), cell state (c) and output gate (o). The forget gate (f_n) determines how much of the past cell state (c_{n-1})

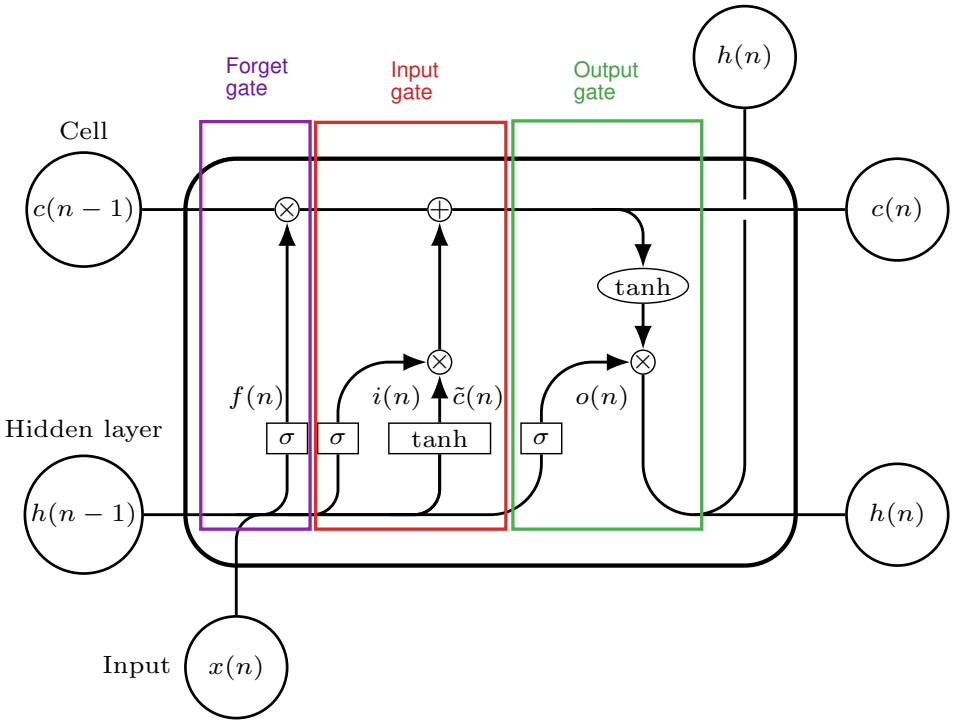


Figure 2.12: Architecture of a LSTM cell.

should be retained and is given by

$$f_n = \sigma(W_{xf} \cdot x_n + W_{hf} \cdot h_{n-1} + b_f). \tag{2.32}$$

$\tilde{c}_n$ is the candidate cell state that represents the new information that can be added to the cell state and is given by

$$\tilde{c}_n = \tanh(W_{xc} \cdot x_n + W_{hc} \cdot h_{n-1} + b_c). \tag{2.33}$$

The update cell state (c_n) combines the new information with the previous cell state and

is defined as

$$c_n = f_n \cdot c_{n-1} + i_n \cdot \tilde{c}_n. \tag{2.34}$$

The information to be updated is decided by combining the forget gate's decision about the information to be discarded and the input gate's decision to add new information. This section represents the 'long-term memory' of the LSTM block. The output gate (o_n) determines the hidden state (h_n) based on the current input and the updated cell state and is given by

$$o_n = \sigma(W_{xo} \cdot x_n + W_{ho} \cdot h_{n-1} + b_o) \tag{2.35}$$

The hidden state (h_n) carries information from previous time steps to the current time step and is given by

$$h_n = o_n \cdot \tanh(c_n), \tag{2.36}$$

which combines the updated cell state with the output gate's decision, and this represents the 'short-term memory' of the LSTM block for the current time step. In summary, the LSTM block uses gates to regulate the flow of information, allowing it to learn long-term dependencies in sequential data by selectively updating, forgetting and remembering information using short-term memory over time [139,141,142]. Complex-valued RNNs have been proven to have enhanced learning stability [143, 144]. However, poor performance was observed when an appropriate activation function was not used [145], but improved performance was observed when combined with complex convolutional layers [89]. In complex-valued LSTMs, matrix multiplication, arithmetic operations, and nonlinearities are performed using complex arithmetic. The standard activation functions, such as the sigmoid and tanh, are replaced with modified ReLU and other complex-valued split activation functions. Furthermore, during backpropagation, Wirtinger calculus is employed instead of the standard real-valued calculus typically used in real-valued LSTMs.

2.10 Convolutional networks

Convolutional neural networks (CNNs) were inspired by the visual cortex operation of animals and have been widely used for recognition, classification and enhancement tasks [146, 147]. Real-valued CNN for handwriting recognition of digits was introduced in [148] and has been improved for applications in computer vision [149], time-series prediction, image classification and recognition [150]. The basic structure of a CNN has a convolutional layer, followed by activation and pooling layers and the final output is usually through a fully-connected layer.

Convolutional layer

This is the fundamental layer of the CNNs and consists of a set of feature maps arranged with the learnable parameters in the form of kernels. The convolution operation involves sliding a small window or kernel across the input data and computing the dot product between the kernel and the local region of the input at each position. The kernel is moved spatially across the input to extract the features by detecting local patterns [147,149,150]. The convolved output is then used as the input to an activation layer to generate a novel feature map or latent representation of the input data [77]. Some CNNs have used the convolutional layers as feature extractors, as each kernel or filter bank used in the convolutional layer is capable of capturing local features of the input data [146, 147]. CNNs with complex-valued weights are shown to be more stable and numerically efficient [77, 150]. Some of the tunable parameters of a typical convolution layer are the kernel size, number of kernels, stride and padding. The kernel size defines the spatial extent of the local region considered for the convolution operation. The kernel size also decides if a 1D or 2D convolution operation is performed on the input data. The number of kernels determines the depth of the output feature maps. The stride decides the step size at which the kernel is moved across the input. A larger stride produces a smaller output feature map. Padding can be applied to the input to ensure that the spatial dimensions of the output feature maps match the input and also preserves the information at the borders of the input data.

Pooling layer

CNN architectures have alternating convolutional layers and pooling layers that help in the reduction of the spatial dimension of the feature maps and the number of parameters. Pooling layers help the CNN to avoid overfitting the model and summarize the presence of specific features in the feature map [77]. Hence, pooling layers are also referred to as the down-sampling layers, where the dimension of the feature map is lowered which produces a lower-resolution version of the input. Some of the common pooling operations include max pooling, average pooling, and stochastic pooling. Similar to convolutional kernels, pooling is also performed by sliding (usually non-overlapping) a pooling window over the feature map after the convolutional operation. In the case of max pooling, the maximum value within each region of the sliding pooling window is the output of the pooling layer. For CVNNs, max pooling can be unreliable as the 'max' operation is undefined for complex numbers. Max pooling for CVNNs can be performed separately on the real and imaginary components, similar to the split-activation functions [77, 151]. In [152], two new approaches to perform pooling on CVNNs called max-by-magnitude and max-by-softmax were introduced. In the first approach, the value of the complex domain is projected onto the real domain. In the second approach, the softmax function is used to perform pooling. Certain CVNNs use fully-connected linear layers as the final output layer.

2.11 Transformer Networks

Transformers are a prominent development in the deep learning model architectures and were originally introduced for natural language processing (NLP) [153]. Since their introduction, transformers have been widely adopted in various applications such as computer vision [154, 155] and audio processing [90, 156–158]. Given the vast advantages of transformers, a few drawbacks were also discovered [155]. The inefficiency of transformer networks in processing long sequences due to computation and memory limitations. Transformers are said to have a flexible architecture and make limited assumptions on the structural bias of the input data which makes it hard to train on small-scale data.

Figure 2.13 shows the architecture of the attention-based transformer model proposed in [153]. The model is a sequence-to-sequence model which consists of an encoder and de-

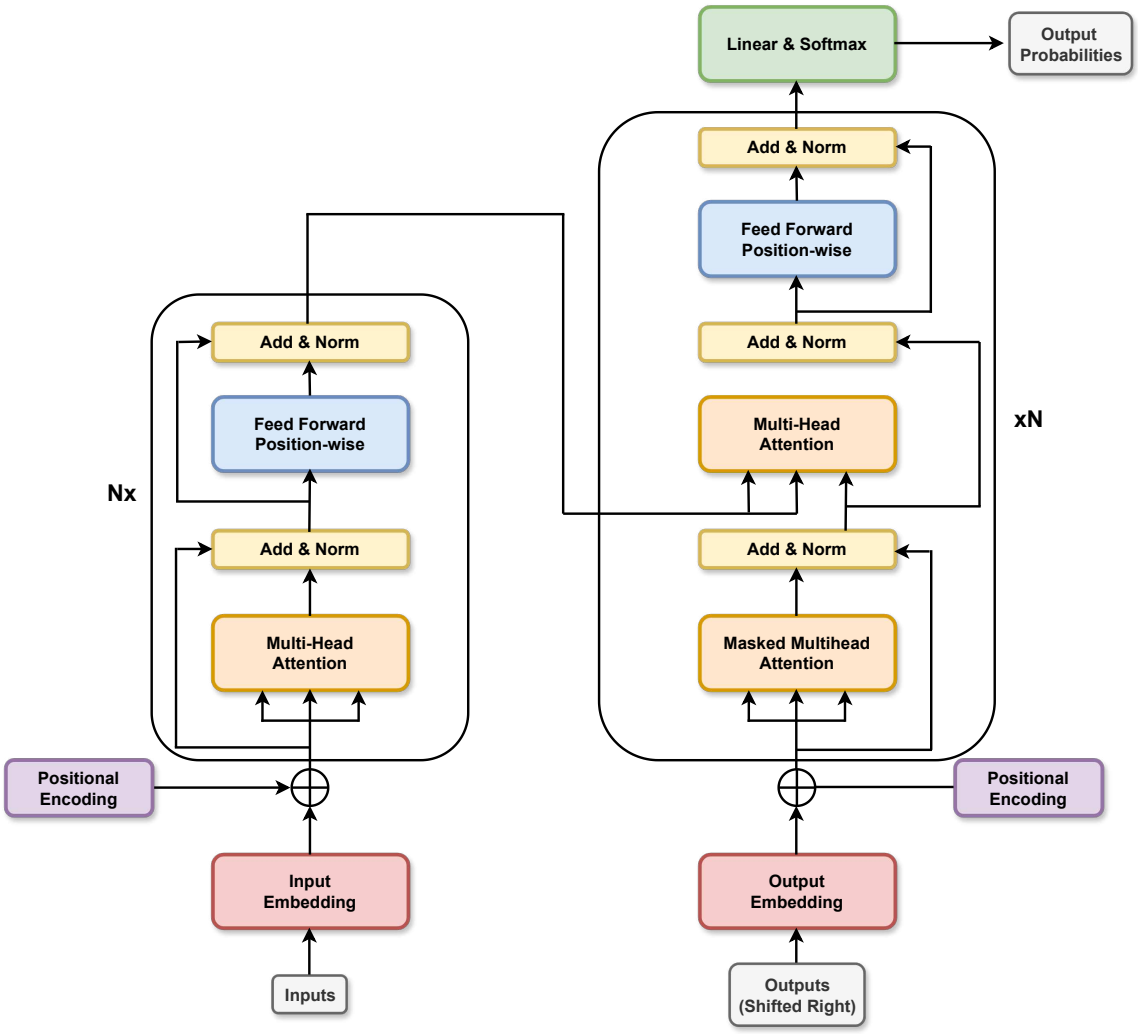
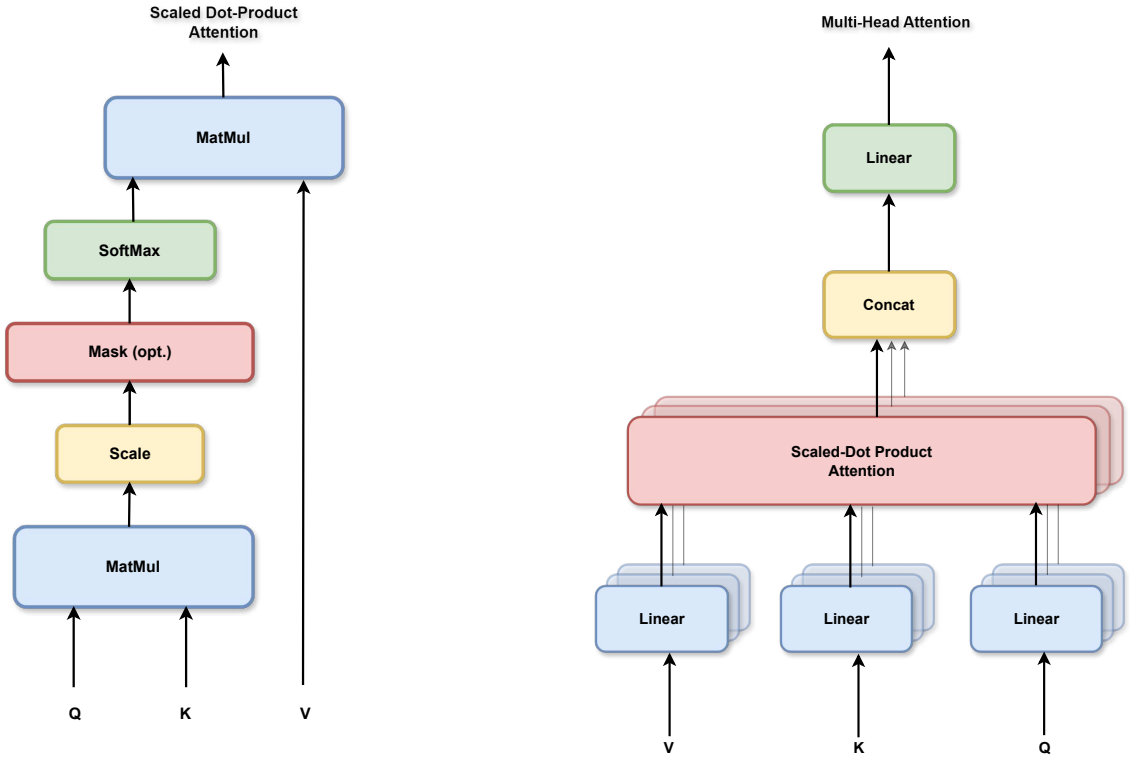


Figure 2.13: Architecture of a transformer.

coder that are a stack of N identical blocks. The encoder consists of multi-head attention blocks and a feed-forward position-wise network. The structure of the multi-head attention layer is shown in Fig. 2.14 along with the scaled dot-product attention mechanism. The Add & Norm block first applies a residual connection, adding the sub-layers input to its output, and then performs layer normalization. The residual connections help mitigate the vanishing gradient problem and allow deeper networks to be trained effectively, while layer normalization ensures that the inputs to each layer maintain a stable distribution, facilitating faster and more robust learning. The point-wise feed-forward network layers



(a) **(b)**

Figure 2.14: (a) Structure of Scaled-dot product attention and (b) structure of Multi-head attention which has several layers running in parallel.

consist of two linear layers with ReLU activation function in between. Positional encoding blocks hold the order sequence information about the relative or absolute positions of tokens in the sequence, which are added to the input embeddings of the encoder and decoder. The authors in [153] recommend the usage of sine and cosine functions of different frequencies that have the same dimensions as the input embedding. The softmax function is used on the learned linear transformation at the decoder output to convert the decoder output to predicted next-token probabilities.

Attention

Attention is defined as the mapping of a query vector and a set of key-value pair vectors to an output vector. The output is the weighted sum of the values with weights decided by a compatibility function of the query with the corresponding key [153]. Transformers use a particular type of attention called the 'scaled dot-product attention' shown in Fig. 2.14(a) [153]. The input consists of queries and keys of dimension d_k, and the values of dimension d_v. The dot products of the query with all the keys are computed and scaled by the value of $\sqrt{d_k}$. The softmax function is applied to obtain the attention weights. For ease of implementation and computational efficiency, matrices are used to compute the attention and are given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.37}$$

where Q, K and V are the matrix representations of query, key, and values, respectively.

Multi-head attention

The schematic of the multi-head attention is shown in Fig. 2.14(b). Several layers of attention are linearly stacked with queries, keys and values projected on the different layers with their dimensions [153, 155]. The attention computed for each layer is then concatenated to obtain the multi-head attention which allows the networks to jointly capture information from different representations simultaneously. The multi-head attention is given by

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O \tag{2.38}$$

where

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{2.39}$$

and W in (2.39), is the respective projection matrices for query, key, value and outputs.

Together, these components contribute significantly to the Transformers performance on a wide range of sequence modeling tasks. The design of the complex-valued transformer is introduced in Chapter 4.

2.12 Evaluation metrics

The most reliable method to assess speech enhancement algorithms is by conducting listening tests with expert listeners to gauge the quality and intelligibility of the enhanced signals. Subjective measures like listening tests involve human perception and judgment. However, listening tests are not practical during the algorithm development phase and evaluation metrics provide insights into the performance of the algorithms. Objective measures are based on mathematical algorithms and do not require human judgment. Objective measures provide standardised and reproducible assessments but may not always correlate perfectly with human perception in certain conditions. In this section, the metrics used in the algorithm development and testing will be briefly introduced. Quality metrics such as perceptual evaluation of speech quality (PESQ) and perceptual objective listening quality assessment (POLQA) evaluate the quality of the enhanced speech signal by comparing it to the clean signal. Intelligibility metrics such as short-time objective intelligibility measure (STOI) and modified binaural STOI (MBSTOI) estimate the intelligibility of the signals by correlating them with the clean speech signals. Noise reduction capabilities of the enhancement algorithms can be assessed by measuring the frequency-weighted segmental SNR (fwSegSNR).

2.12.1 Quality metrics

Frequency-weighted segmental SNR (fwSegSNR)

SNR measures the ratio of signal power to noise power in a speech signal. It is widely used to assess the noise reduction capabilities of an enhancement algorithm. To measure the segmental SNR, the clean reference and noisy signals are split into non-overlapping frames and SNR of each frame is calculated in dB [159]. Before computing the segmental SNR the signals are A-weighted by filtering them. The fwSegSNR for the processed signal

$\hat{s}$ is given by

$$\text{fwSegSNR} = \frac{10\log_{10}}{L}\sum_{i=1}^{L}\frac{\|s_i^2\|}{\|s_i^2 - \hat{s}_i^2\|} \tag{2.40}$$

where s is the clean signal and $\|\cdot\|$ is the L_2 norm, and L is the total number of frames the signal is divided into. The use of fwSegSNR is better suited for speech processing as the SNR calculation is weighted differently over frequency bands and is closer to human perception of the signal [160]. An online implementation of the fwSegSNR is available in [161].

Precisely estimating a listener's subjective assessment of speech quality through testing remains a prominent area of study, leading to the development of several industry standards [162]. Perceptual evaluation of speech quality (PESQ) is a widely used standard monaural speech quality metric introduced in [163, 164]. It operates by simulating the human auditory perception and evaluating the perceptual similarity between the processed and reference speech signals. PESQ analyses various aspects of speech quality, including distortion, background noise and intelligibility and produces a quality score ranging from -0.5 to 4.5, where higher values indicate higher speech quality.

Perceptual objective listening quality assessment (POLQA) is a more advanced objective quality measurement algorithm introduced in [165,166]. It is a standardised objective measure extensively used in telecommunication applications. In contrast to PESQ, POLQA is designed to evaluate the quality of speech signals subjected to various impairments such as codec compression, packet loss, and network delays. POLQA provides a quality score between 0 to 4.5, with higher scores indicating better quality. POLQA is proven to be more accurate and reliable compared to PESQ [162].

2.12.2 Intelligibility metrics

Speech intelligibility metrics are used to evaluate the extent to which the speech is understood or comprehended by listeners. The SNR-based metric for intelligibility, the speech intelligibility index (SII) [167] was designed based on articulation index (AI). SII performs well with stationary additive noise but was unable to predict the intelligibility of

speech signals with fluctuating or modulating noise. In contrast to SNR-based metrics, correlation-based metrics such as short-time objective intelligibility measure (STOI) and modified binaural STOI (MBSTOI) have been shown to have higher accuracy in predicting the intelligibility of speech signals [168].

Short-time objective intelligibility measure (STOI)

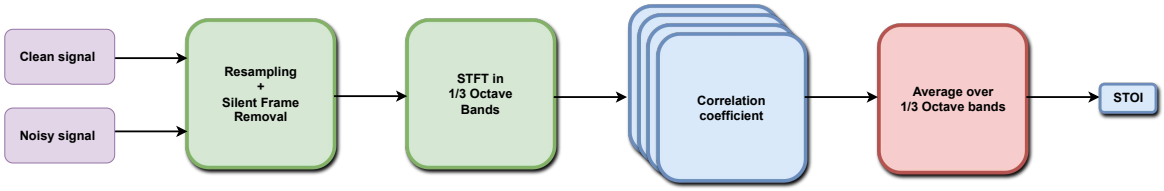


Figure 2.15: Block diagram of STOI computation.

Several studies have shown that STOI scores correlate with subjective intelligibility results in noisy and enhanced speech processed with speech enhancement algorithms [160, 169, 170]. A major drawback of STOI is that it performs poorly with additive noises with strong temporal modulations and does not correlate with human perception of intelligibility affected by modulations [171]. An improved version of STOI has been proposed to adapt STOI for temporal modulations called Extended STOI (ESTOI) [171]. In [172], the authors propose a modified version of STOI, where each TF cell contribution for the final STOI score is weighted by the estimated intelligibility content of the TF cell. Both metrics compare a clean speech signal with a degraded speech signal. Also, for computing the weighted STOI (WSTOI), the correlation comparison is computed on individual STFT frequency bins rather than third-octave bands.

Modified binaural STOI (MBSTOI)

STOI is one of the most popular single-channel intrusive intelligibility metrics, and it was extended into a binaural metric using an equalisation cancellation (EC) stage in the deterministic binaural STOI (DBSTOI) [173]. However, it was realised that DBSTOI overestimates the intelligibility score for spatially distributed interferers with very noisy conditions and low SNRs. The MBSTOI metric was introduced in [174] to rectify this inaccuracy and improve the robustness of binaural intelligibility estimation by optimising

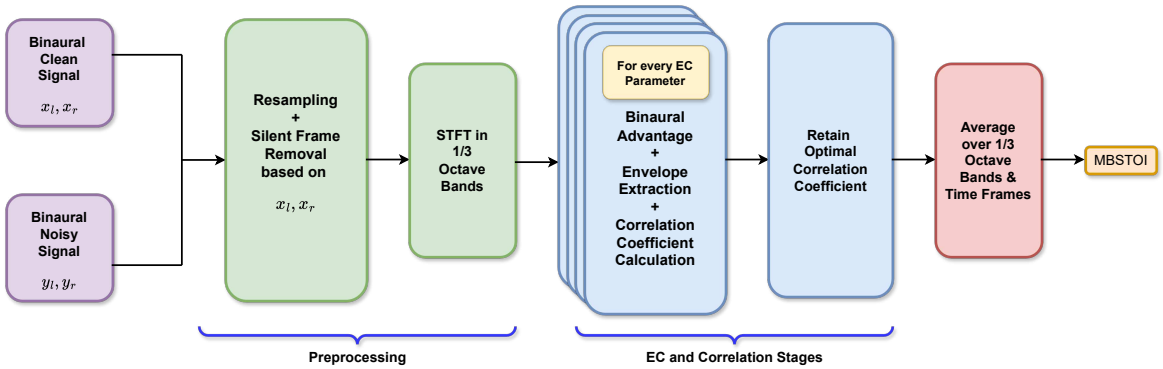


Figure 2.16: Block diagram of MBSTOI computation.

the EC parameter estimation criterion. The block diagram for MBSTOI calculation is shown in Fig. 2.16 [168, 175]. MBSTOI is an intrusive metric and takes both noisy and clean binaural signals as inputs. Similar to the STOI score computation, the first step includes resampling the signals to 10 kHz and the silent frames in the signal are removed using the clean reference signal. The signals are then transformed into the STFT domain and aggregated into $1/3^{rd}$ octave bands. The STFT coefficients of both ear-signals are combined to model the binaural advantage in intelligibility [174]. Human listeners perceive higher intelligibility when there is a spatial separation between the target and the interferer [7]. The EC stage uses parameters that closely represent the ILD and IPD of the dominant interferer. As the direction information about the target and the interferers are not known, an exhaustive grid search is performed for optimal EC parameters with resolutions of 1 dB and 20 μs. Based on the optimal criterion defined in [174], a single pair of values for the EC parameters is selected. The correlation between clean and degraded envelopes is computed similarly to the STOI metric and averaged across frequency bands and time frames to estimate a MBSTOI value between 0 and 1 for the noisy signal. A more detailed description of MBSTOI, experiments and discussion can be found in [168, 173–175].

From the metrics discussed above, in this work, fwSegSNR is used to evaluate the noise reduction capabilities of the enhancement algorithms. For binaural signals the average values of the left and right channels are considered. To estimate the improvement in binaural speech intelligibility, MBSTOI is used in this work.

2.12.3 MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) Test

Listening tests are a common method for evaluating the quality of audio systems, particularly in controlled settings with carefully selected participants and professional audio equipment. In the speech and audio domain, such tests are conducted to assess the improvements that new audio systems or algorithms provide over current state-of-the-art systems. Key guidelines for designing these tests include recommendations such as ITU-R BS.1534 for intermediate-quality audio coding assessments [176] and ITU-R BS.1116 for high-quality audio systems [177]. Despite their value, listening tests are resource-intensive due to the time and effort required to recruit participants, conduct the tests, and analyse the results. In contrast, objective evaluation methods offer the advantage of being faster, typically requiring only seconds or minutes, and their results are reproducible as long as test conditions remain consistent. However, the results of listening tests are not fully reproducible, as they depend on the statistical significance of the participants' ratings. The primary benefit of listening tests, compared to objective methods, is that they accurately capture human perception, which is the target of most audio developments [178].

The ITU introduced the first version of the MUSHRA methodology in 2001 under Recommendation BS.1534-0 for evaluating intermediate audio quality [176]. This was later updated in Recommendation BS.1534-2, which made changes to both test design and result analysis [179]. In MUSHRA tests, listeners are presented with an open reference stimulus and several test stimuli, including a hidden reference stimulus, two anchor stimuli (low-quality and mid-quality), and stimuli processed by the systems under evaluation. A maximum of eleven conditionscomprising eight systems under test, two anchors, and the hidden referencecan be assessed in a single trial. Since the order of conditions is randomised, assessors do not know whether a condition corresponds to a system under test, the hidden reference, or an anchor. All conditions are rated relative to the open reference. The inclusion of the hidden reference implicitly sets a benchmark for the highest rating, reducing the chance that a system under test, even if it introduces minor artefacts, receives the highest score. The hidden reference also serves to verify whether assessors correctly rate it as identical to the open reference. If an assessor consistently rates the hidden reference too low across multiple trials, their results may be excluded during post-screening. BS.1534-2 recommends using two anchor stimulilow-pass filtered versions of

the reference stimuluswith cut-off frequencies of 3.5 kHz for the low-quality anchor and 7 kHz for the mid-quality anchor. While rating, assessors can switch instantly between the conditions and the reference and can loop short, critical sections of the stimulus to detect artefacts. The ratings are made on a continuous quality scale (CQS) ranging from 0 to 100, with labels for each interval: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100).

Although MUSHRA was initially designed to assess audio coding quality, it is now widely applied to evaluate various speech and audio algorithms [178, 180]. A web-based MUSHRA test API was introduced in [178], which allows participants to take the test on a web audio plugin on an internet browser.

Chapter 3

Binaural Speech Enhancement Using STOI-Optimal Masking

This chapter details the use of short-time objective intelligibility measure (STOI)-optimal masking for binaural speech enhancement. The technical background section gives a mathematical formulation of the problem and an overview of weighted STOI (WSTOI) metric, the high-resolution stochastic WSTOI-optimal binary mask (HSWOBM) and the use of optimally-modified log spectral amplitude (OM-LSA) and speech presence probability (SPP) for mask application. The proposed extension of the masking method to binaural speech is then studied. Simulation experiments are presented to evaluate the proposed method which is followed by the conclusion. An earlier version of the work presented in this chapter has been published in [181].

3.1 Introduction

For binaural signals, along with enhancement of signal-to-noise ratio (SNR) and intelligibility, binaural cues must also be preserved. Studies have shown that preserving the binaural cues is helpful for sound localization and speech intelligibility in noisy environments due to *binaural unmasking* [8, 182]. Interaural level differences (ILDs) and interaural time differences (ITDs) are particularly helpful in localizing sound, dereverberation, improving intelligibility and boosting the perceived loudness [7, 183]. Binaural speech enhancement methods using beamformers [37, 184], multichannel Wiener filters [36, 185]

and mask-informed enhancement methods [186, 187] were previously proposed.

Mask-based speech enhancement has been well studied for monoaural signals and has been shown to improve the SNR and intelligibility [46, 188]. Classical enhancement methods apply a time-frequency (TF) gain in by transforming the noisy signal into the TF domain using short time Fourier transform (STFT), applying the gain mask and then transforming the signal back to the time-domain [189–191]. The TF gains applied to the noisy signal are estimated based on an assumed statistical model to minimize the mean-squared log spectral amplitude (LSA) error between the clean and noisy signals. Studies have shown that these methods are successful in improving the SNR of the noisy signal but were not able to improve the intelligibility of the speech [58, 192]. The ideal binary mask (IBM) introduced in [48], uses a binary-valued gain of 0 and 1 (sometimes 0.1 and 1) for noisy speech enhancement and studies have shown that intelligibility can be improved using the binary mask [43, 44, 46]. A detailed review of IBM is discussed in Sec. 2.2.3. The ideal ratio mask (IRM) introduced in [53], uses continuous values for the TF gains depending on the probability of TF bins being speech-dominated or noise-dominated. Section 2.2.4 has a detailed review of IRM. Although these algorithms improve the SNR considerably, enhanced speech signals contain significant audible artefacts which impact the intelligibility and quality of the speech signal [45, 46].

The STOI-optimal binary mask (SOBM), designed to optimize STOI, was introduced in [190] for monoaural signals. A method for extending the monoaural mask-assisted enhancement to binaural speech by using HSWOBM version of the SOBM, which optimizes WSTOI [172], as our training target to estimate a continuous-valued mask is introduced. Directly applying the TF mask as a gain in the STFT domain produces significant artefacts and so in [193], the OM-LSA is proposed as an alternative method of applying the TF mask which improves the perceptual quality of the enhanced speech [194]. The improvement in frequency-weighted segmental SNR (fwSegSNR) is computed to measure the noise reduction, and modified binaural STOI (MBSTOI) is estimated to predict the binaural intelligibility improvement. The root mean square (RMS) error in ILD is calculated to study the distortion of ILD of the target speaker.

3.2 Technical Background

3.2.1 Description of STOI and WSTOI

STOI is an intrusive intelligibility metric based on the correlation between the spectral envelopes of clean and degraded versions of the speech [170]. To compute STOI, these signals are resampled to 10 kHz and converted into the STFT domain using a 50%-overlapping Hanning window of length 25.6 ms and this results in the clean and degraded STFT signals, $X(k,\ell)$ and $Y(k,\ell)$, with (k,ℓ) the frequency and time frame indices. STFT frames that have total energy less than 40 dB or more compared to the frame with the highest energy are considered to be silent frames. These frames are deleted from the clean and noisy signals and are not considered for STOI score computation. $X(k,\ell)$ is combined into 15 third-octave bands (J) by calculating the amplitudes of the TF cells denoted by $X_j(\ell)$ where $j = 1,\ldots,J$ and is given by

$$X_j(\ell) = \sqrt{\sum_{k=K_j}^{K_{j+1}-1} |X(k,\ell)|^2} \quad \text{for} \quad j = 1,\ldots,J \tag{3.1}$$

where K_j is the lowest STFT frequency bin within the frequency band j. The modulation vector for each ℓ is defined as

$$\mathbf{x}_{j,\ell} = [X_j(\ell - L + 1), X_j(\ell - L + 2), .., X_j(\ell)]^T, \tag{3.2}$$

where $L = 30$. Similar processing is applied to $Y(k,\ell)$ to obtain $Y_j(\ell)$ and $\mathbf{y}_{j,\ell}$. The modulation vectors for clean and noisy signals, denoted by $\mathbf{x}_{j,\ell}$ and $\mathbf{y}_{j,\ell}$, are computed by performing a correlation between clean and degraded speech vectors of duration $(25.6 \times 30)\ /2 = 384$ ms. The clipped TF cell amplitudes to limit the impact of frames having low speech energy, denoted by $\widetilde{[\cdot]}$, are determined as

$$\widetilde{Y}_j(\ell) = \min\left(Y_j(\ell), \lambda \frac{\|\mathbf{y}_{j,\ell}\|}{\|\mathbf{x}_{j,\ell}\|} X_j(\ell)\right) \tag{3.3}$$

where $\lambda = 6.623$ and $\|\cdot\|$ denotes the Euclidean norm. The corresponding clipped modulation vector is $\tilde{\mathbf{y}}_{j,\ell}$. The STOI contribution of each TF cell (j, ℓ) is then given by

$$d(\mathbf{x}_{j,\ell}, \tilde{\mathbf{y}}_{j,\ell}) \triangleq \frac{(\mathbf{x}_{j,\ell} - \bar{x}_{j,\ell})^T \tilde{\mathbf{y}}_{j,\ell}}{\|\mathbf{x}_{j,\ell} - \bar{x}_{j,\ell}\| \|\tilde{\mathbf{y}}_{j,\ell} - \bar{\tilde{y}}_{j,\ell}\|}, \tag{3.4}$$

where $\bar{x}_{j,\ell}$ and $\bar{\tilde{y}}_{j,\ell}$ denote the mean of the vectors $\mathbf{x}_{j,\ell}$ and $\tilde{\mathbf{y}}_{j,\ell}$, respectively. The overall STOI metric is computed by averaging the contributions of the TF cells over all bands, j, and frames, ℓ. The STOI metric is given by

$$\text{STOI} = \frac{1}{JL} \sum_{j=1}^{J} \sum_{\ell=1}^{L} d(\mathbf{x}_{j,\ell}, \tilde{\mathbf{y}}_{j,\ell}). \tag{3.5}$$

In [172], a weighted version of STOI was proposed. The weights are computed by estimating the mutual information between the unpredictable component of a hypothetical signal which is assumed the speaker intended to communicate and the version of the signal that the listener has perceived in an imagined scenario where the listener hears the clean speech signal [172]. A simple communication model is considered to model this scenario.

The clean speech, $X_j(\ell)$, is written as

$$X_j(\ell) = S_j(\ell) + V_j^p(\ell), \tag{3.6}$$

where $S_j(\ell)$ is the speaker's intended speech and $V^p{}_j(\ell)$ is the 'production noise' proposed in [195], which models the natural variation in human speech. The residual linear prediction which is the message signal at the input of the communication model is given by

$$R_j(\ell) = \hat{S}_j(\ell) - S_j(\ell), \tag{3.7}$$

where $\hat{S}_j(\ell)$ is the linear prediction of $S_j(\ell)$. The output of the communication model is $W_j(\ell)$ which is the sum of $R_j(\ell)$, $V_j(\ell)$ and the frequency band-dependent hypothetical internal ear-noise that models the absolute threshold of human hearing, denoted by $N_{ej}(\ell)$. The mutual information denoted by $I(\cdot)$ is computed by

$$I(\mathbf{r}_{j,\ell}; \mathbf{w}_{j,\ell}) = 0.5 \times \log_2 \left(1 + \frac{\|\mathbf{r}_{j,\ell}\|^2}{N_{ej}(\ell) + a\|\mathbf{x}_{j,\ell}\|^2} \right), \tag{3.8}$$

where $\mathbf{r}_{j,\ell}$ and $\mathbf{w}_{j,\ell}$ are defined analogous to (3.2), and a is a fixed co-efficient which models the ratio of power of the production noise to that of the speech signal [172, 195]. The free parameter, a, in (3.8) is determined by using the Kneser-Ney model which is a modified interpolated phoneme level trigram language model from [196, 197].

Using (3.8) as weights the WSTOI metric is computed as the weighted average of (3.4) over all bands j and across all frames ℓ and is given by

$$\text{WSTOI} = \frac{1}{\sum_{j,\ell} I_{j,\ell}} \sum_{j,\ell} I_{j,\ell} \quad d(\mathbf{x}_{j,\ell}, \tilde{\mathbf{y}}_{j,\ell}). \tag{3.9}$$

3.2.2 Optimally-modified log spectral amplitude estimator

LSA estimator was introduced in [191] and an optimally-modified log spectral amplitude (OM-LSA) estimator was proposed in [193]. The speech signals are transformed into the TF-domain using overlapping Hamming windows which results in the clean and degraded STFT signals, $X(k, \ell)$ and $Y(k, \ell)$, with (k, ℓ) the frequency and time frame indices. Under hypothesis $H_0(k, \ell)$ it is considered that speech is absent and under hypothesis $H_1(k, \ell)$ it is considered that speech is present in a particular TF-bin, (k, ℓ). The STFT coefficients of noise and speech components are modelled as statistically independent Gaussian random variables. The gain function, $G(k, \ell)$, in frequency bin k of frame ℓ that satisfies

$$G(k,\ell)|Y(k,\ell)| = \exp\{\mathbb{E}[\log|X(k,\ell)|\ |Y(k,\ell)|]\} \tag{3.10}$$

where $\mathbb{E}[\cdot]$ is the expectation operator. Under the constraint that the gain $G(k,\ell)$ is larger than the threshold, G_{min} when speech is absent, it is shown in [191] that

$$G(k,\ell) = \{G_{H_1}(k,\ell)\}^{p(k,\ell)}\ G_{min}^{1-p(k,\ell)} \tag{3.11}$$

where the gain under the hypothesis $H_1(k,\ell)$, $G_{H_1}(k,\ell)$ is given by

$$G_{H_1}(k,\ell) = \frac{\xi(k,\ell)}{1+\xi(k,\ell)} \exp\left(\frac{1}{2}\int_{\upsilon(k,\ell)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{3.12}$$

The conditional SPP $p(k,\ell)$ is defined as

$$p(k,\ell) \triangleq P(H_1(k,\ell)|Y(k,\ell)) \tag{3.13}$$

and computed as

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\exp(-\upsilon(k,\ell))\right\}^{-1} \tag{3.14}$$

where $q(k,\ell) \triangleq P(H_0(k,\ell))$ is the *a priori* probability of speech absence in the hypothesis $H_0(k,\ell)$, $\upsilon(k,\ell)$ is the *a posteriori* SNR given by

$$\upsilon(k,\ell) \triangleq \frac{\gamma(k,\ell)\xi(k,\ell)}{1+\xi(k,\ell)} \tag{3.15}$$

where

$$\gamma(k,\ell) \triangleq \frac{|Y(k,\ell)|^2}{\mathbb{E}[|N(k,\ell)|^2]} \tag{3.16}$$

and $\xi(k,\ell)$ is the *a priori* SNR given by

$$\xi(k,\ell) \triangleq \frac{\mathbb{E}[|X(k,\ell)|^2]H_1(k,\ell)]}{\mathbb{E}[|N(k,\ell)|^2]}. \tag{3.17}$$

Using the modified decision-directed approach from [189] an estimate of $\xi(k,\ell)$, $\hat{\xi}(k,\ell)$ is computed and given by

$$\hat{\xi}(k,\ell) = \alpha G_{H_1}^2(k,\ell-1)\gamma(k,\ell-1) + (1+\alpha)\max\{\gamma(k,\ell)-1,0\} \tag{3.18}$$

where α is a smoothing parameter.

Speech presence probability prior

In [191], $\hat{q}(k,\ell)$, an estimator was used to obtain the speech absence probability, $q(k,\ell)$, from $\hat{\xi}(k,\ell)$. In [193] it was proposed to obtain $q(k,\ell)$ from a binary mask, $d(k,\ell)$. The *a priori* probability parameter $q(k,\ell)$ from [191] is defined as

$$q(k,\ell) = \begin{cases} Q^1 & d(k,\ell) = 1 \\ Q^0 & d(k,\ell) = 0 \end{cases} \tag{3.19}$$

where Q^1 and Q^0 are free parameters [193]. G_{min} is estimated similarly and is given by

$$G_{min} = \begin{cases} G^1 & d(k,\ell) = 1 \\ G^0 & d(k,\ell) = 0 \end{cases} \tag{3.20}$$

where G^1 and G^0 are free parameters [193]. Using binary mask values to control the probability of speech presence and absence the algorithm is said to softly impose the spectro-temporal modulations captured in the mask vital for speech intelligibility [45, 170, 193]. Using the SNR-dependant TF gain $G(k,\ell)$ improves the perceived quality of the speech along with preserving the intelligibility.

3.3 Proposed Method

A block diagram of the proposed method is shown in Fig 3.1. The L and R labels indicate the left and right channels of a binaural signal. The left and right channels are processed in parallel for all stages of the algorithm. The noisy binaural signal is transformed into the STFT domain for TF-based processing. From the STFT domain signals, 3 feature subsets as defined in Sec. 3.3.1 are extracted once per STFT frame and the same features are used for the training and prediction stages of the algorithm. STOI optimal masks correlate to intelligible speech in the signal and SPP is computed from the estimated continuous-valued mask and then supplied to the OM-LSA. The output signal of the OM-LSA is transformed back into the time domain by applying an inverse STFT (ISTFT).

3.3.1 Mask Estimation Features

The feature extraction and mask estimation for noisy binaural speech follow a similar approach from [198] proposed for monaural speech. The first step in feature extraction for mask estimation is to normalize the noisy speech active level to 0 dB using the ITU P.56 objective speech active level [199] to make the algorithm level independent. Each of the three feature subsets has Ψ features, giving 3Ψ features in total.

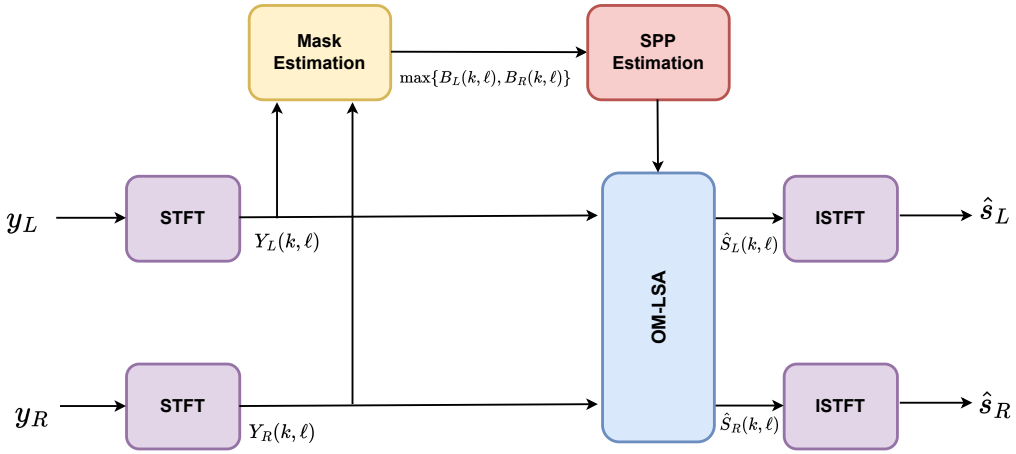


Figure 3.1: Block diagram of the proposed method.

Feature subset 1

The first feature subset is constructed from the TF gains estimated by minimising the mean-squared error in the log-spectral amplitudes using the classical Log-MMSE speech enhancement algorithm [161, 190]. The TF gains are expected to be high when speech is present and low otherwise. As this algorithm uses a noise estimator [200], these features are expected to help generalize the mask estimator to unseen noise types. Recalling the gain function $G(k,\ell)$ from (3.10) [190]

$$G(k,\ell)|Y(k,\ell)| = \exp\{\mathbb{E}[\log|X(k,\ell)|\ |Y(k,\ell)|]\},$$

the first feature subset is given by

$$\boldsymbol{\mu}_\ell^{(1)} = \left[\mu_{1,\ell}^{(1)}, ..., \mu_{\Psi,\ell}^{(1)}\right]^T \tag{3.21}$$

where it is $\Psi \times 1$ vector found by averaging $G(k,\ell)$ in $\Psi = 30$ triangular windows, $w_i(k)$ for $i = 1, .., \Psi$ equally spaced on equivalent rectangular bandwidth (ERB) scale centre frequencies with 50% overlap between windows and then computing the natural logarithm.

For frame ℓ, the feature subset, $\mu_{i,\ell}^{(1)}$ is given by

$$\mu_{i,\ell}^{(1)} = \ln \left\{ \frac{\sum_{k=0}^{K/2} w_i(k) G(k,\ell)}{\sum_{k=0}^{K/2} w_i(k)} \right\} \quad \text{for } i = 1, ..., \Psi \tag{3.22}$$

where K is the discrete Fourier transform (DFT) length. The human auditory filter based ERB scale spaced window function, $w_i(k)$ is given by

$$w_i(k) = \digamma_i \left(\Phi \left(\frac{k f_s}{K} \right) \right) \quad \forall k, \tag{3.23}$$

where f_s is the sampling frequency, $\digamma_i$ is the ERB scale triangular window function, $\Phi(f)$ is the ERB transformation whose inverse derivative $\frac{df}{d\Phi(f)}$, approximates the human auditory filter bandwidths between 0.1 and 6.5 kHz [161, 201] and is given by

$$\Phi(f) = 11.17268 \cdot \ln \left(1 + \frac{46.06538 \cdot f}{f + 14678.49} \right). \tag{3.24}$$

Feature subset 2

The second feature subset, denoted by

$$\boldsymbol{\mu}_{\ell}^{(2)} = \left[\mu_{1,\ell}^{(2)}, ..., \mu_{\Psi,\ell}^{(2)} \right]^T, \tag{3.25}$$

is the estimate of the level-normalized enhanced speech amplitude in each frequency band ℓ and is obtained by multiplying the gains from the first feature subset with the noisy speech. These features are included as they provide a direct correlation to speech presence in each TF bin [202]. The second feature subset is given by

$$\mu_{i,\ell}^{(2)} = \ln\left\{\frac{\sum_{k=0}^{K/2} w_i(k)G(k,\ell)|Y(k,\ell)|}{\sum_{k=0}^{K/2} w_i(k)}\right\} \quad \text{for } i = 1, ..., \Psi. \tag{3.26}$$

Feature subset 3

The third feature subset

$$\boldsymbol{\mu}_{\ell}^{(3)} = \left[\mu_{1,\ell}^{(3)}, ..., \mu_{\Psi,\ell}^{(3)}\right]^T, \tag{3.27}$$

is an estimate of the local voiced-speech-plus-noise to noise ratio (VSSNR) in the TF regions and is obtained using the PEFAC algorithm [203]. This helps to detect the presence of voiced speech energy in local TF regions and helps to provide an independent way to estimate speech presence in very noisy conditions [202]. The fundamental frequency ($f_0(\ell)$) of the speech in each frame (ℓ) is calculated. Then, the VSSNR is estimated across the Ψ frequency bands by comparing the energy at the harmonics of the fundamental frequency with the energy located halfway between consecutive harmonics. A set of triangular gains with peaks at harmonics or multiples of the estimated fundamental frequency is multiplied by the noisy speech power, $|Y(k,\ell)|^2$. These values are then averaged within Ψ overlapping triangular frequency bands equally spaced on the ERB scale and are given by,

$$\mu_{i,\ell}^{(3)} = \ln\left\{\frac{\sum_{k=0}^{K/2} \rho_i(k)h_p(k)|Y(k,\ell)|^2}{\sum_{k=0}^{K/2} \rho_i(k)h_t(k)|Y(k,\ell)|^2}\right\} \quad \text{for } i = 1, ..., \Psi. \tag{3.28}$$

where

$$h_p(k) = \sum_j \Lambda\left(\frac{2(f - f_0(\ell)\cdot j}{f_0(\ell)}\right)\Bigg|_{f=\frac{kf_s}{\ell}}$$

and

$$h_t(k) = \sum_j \Lambda\left(\frac{2(f - f_0(\ell)\cdot (j+0.5)}{f_0(\ell)}\right)\Bigg|_{f=\frac{kf_s}{\ell}}.$$

$\rho_i(k)$ is identical to $w_i(k)$ in (3.26) with a minimum window width imposed on the triangular windows [198] and Λ is the triangular function given by

$$\Lambda(f) \triangleq \begin{cases} 1-|f| & |f|<1 \\ 0 & \text{otherwise} \end{cases} \tag{3.29}$$

The complete feature set for frame ℓ is obtained by concatenating the three subsets to obtain the feature set

$$\boldsymbol{\mu}_\ell = \left[\boldsymbol{\mu}_\ell^{(1)T}, \boldsymbol{\mu}_\ell^{(2)T}, \boldsymbol{\mu}_\ell^{(3)T}\right]^T. \tag{3.30}$$

Each ear channel of the binaural signal is separately used to compute the features, and processed for training the deep neural network (DNN).

3.3.2 Target mask estimation

Several studies have used oracle binary masks estimated from clean speech as targets for DNN training to enhance noisy speech [45, 63, 204]. In [205, 206], it is shown that the intelligibility of speech depends both on the instantaneous spectrum and its temporal modulation which has led to the development of intelligibility metrics such as speech transmission index (STI) [207] and STOI [208]. The oracle binary masks SOBM and

HSWOBM that maximize intelligibility metrics were proposed in [198,202]. The difference between SOBM and HSWOBM is that the latter is designed to optimize WSTOI rather than STOI and is computed over a higher number of frequency bands (129 instead of 15) to increase the resolution of the mask. In the proposed method, HSWOBMs are used as target masks and are computed using the three-stage dynamic programming technique described in [202] which consists of carrying out three passes of estimation with a pruning strategy to compute the target masks to optimize the WSTOI from (3.9). The $J = 129$ ERB frequency bands with centre frequencies between 100 Hz and 5 kHz are used. The target masks are computed for the left and right channels individually based on the speech content in each channel. The target HSWOBM, $B_j(\ell)$ is computed by forming a masked signal given by,

$$Z_j(\ell) = B_j(\ell)Y_j(\ell) \tag{3.31}$$

where $B_j(\ell) \in \{0, 1\}$. The clipped modulation masked vector $\tilde{\mathbf{z}}_{j,\ell}$ is then computed analogously to (3.3) to limit the impact of frames having low speech energy. B_j is numerically optimized using the least squares method separately in each band, j, by computing

$$B_j(\ell) = \underset{\{B_j(\ell):\ell=1,\ldots,N\}}{\operatorname{argmax}} \left(\sum_{\ell=1}^{N} I_{j,\ell} \mathbb{E}[d(\mathbf{x}_{j,\ell}, \tilde{\mathbf{z}}_{j,\ell})] \right) \tag{3.32}$$

where N is the number of frames and

$$\mathbb{E}[d(\mathbf{x}_{j,\ell}, \tilde{\mathbf{z}}_{j,\ell})] \approx \frac{(\mathbf{x}_{j,\ell} - \bar{x}_{j,\ell})^T \, \mathbb{E}[\tilde{\mathbf{z}}_{j,\ell}]}{\|\mathbf{x}_{j,\ell} - \bar{x}_{j,\ell}\| \sqrt{\mathbb{E}[\|\tilde{\mathbf{z}}_{j,\ell} - \bar{\tilde{z}}_{j,\ell}\|^2]}}. \tag{3.33}$$

3.3.3 Mask Estimation network

A feed-forward DNN is trained to estimate a continuous mask from the features extracted from the noisy binaural signals by using the HSWOBM target masks estimated by dynamic programming [198,202]. The DNN consists of 4 layers of 'dense' or 'fully-connected'

layers each with 500 hidden neuron layers. Rectified linear unit (ReLU) is used as the activation function for all the layers except the output layer where the Sigmoid activation function is used to limit the range of output values. The design and structure of feed-forward network architecture, ReLU and Sigmoid activation function are detailed in Sec. 2.3. Dropout layers are placed between the layers with a dropout factor of 20% which prevents overfitting of the neural network model.

Loss function

A weighted mean square error is used as the loss function for the DNN. The loss function, $\mathcal{J}$, is given by

$$\mathcal{J} = \frac{1}{\Delta} \frac{\sum_{p=1}^{\Delta} \phi_p (\mathrm{c}_p - \zeta_p)^2}{\sum_{p=1}^{\Delta} \phi_p} \tag{3.34}$$

where c_p is the output of the DNN, Δ is the number of pairs of feature vectors and corresponding mask value pairs available for training, ζ_p is the training target and ϕ_p is the corresponding value of the weight. The weights are computed from the WSTOI sensitivity (3.9) and band importance weighting obtained from the WSTOI metric [172]. Adam optimization, a gradient descent-based optimization technique for DNN training is applied. The DNN is modelled and trained to estimate a continuous-valued mask by mapping the input features to the target mask values. The network learns to map the values from the noisy speech features with the target binary mask based on the intelligibility metric which is then supplied to the OM-LSA for mask application.

3.3.4 Better-ear Processing

Studies [7,187] have shown that human listeners are capable of better ear listening, which focuses on the speech signal from the ear which has higher SNR and intelligibility. A similar approach is adopted for the selection of mask values or gains in each TF bin. Choosing the maximum gain value from the two estimated masks for the left and right channels is equivalent to selecting the gain associated with lower noise levels and higher

speech presence from the SPP estimation. To preserve the spatial cues of the binaural signal, the same gain must be applied to both channels. Therefore, the gains selected from this method are used to compute a single SPP as input to the OM-LSA enhancer for both channels. The better-ear continuous-valued mask is given by

$$B(k,\ell) = \max\{B_L(k,\ell), B_R(k,\ell)\} \tag{3.35}$$

where B_L and B_R are the estimated continuous valued masks from the DNN for the left and right channels of the binaural signal respectively.

3.3.5 Optimally modified-LSA

Mask application follows the method in [193], which is shown to improve the perceptual quality of masked speech. From the computed HSWOBM, an intermediate continuous-valued mask is estimated and, instead of applying this mask as a TF domain gain, it is used to supply the probability of speech presented to the OM-LSA described in Sec. 3.2.2. aligns (3.19) and (3.38) are modified to work with continuous-valued mask estimated from the DNN and replaced with

$$\hat{q}(k,\ell) = Q^0 + (Q^1 - Q^0)B(k,\ell) \tag{3.36}$$

and

$$G_{min} = G^0 + (G^1 - G^0)B(k,\ell) \tag{3.37}$$

The gain function in (3.10) is modified to suppress musical noise in silent frames [198] and is given by

$$G'(k,\ell) = \begin{cases} \Omega & B(k,\ell) < \Gamma \\ G(k,\ell) & B(k,\ell) \geq \Gamma \end{cases} \tag{3.38}$$

G^1	G^0	Q^1	Q^0	Ω	Γ
-3 dB	-43 dB	0.7	1	0.2	0.25

Table 3.1: Optimal values for OM-LSA algorithm estimated for continuous-valued mask application.

The optimal values for $(G^1, G^0, Q^1, Q^0, \Omega, \Gamma)$ were experimentally determined in [198] using a grid search of a subset of speech utterances from the TIMIT database [209]. The optimal values for $(G^1, G^0, Q^1, Q^0, \Omega, \Gamma)$ are listed in Table 3.1. The computed mask is applied to both the left and right channels of the noisy binaural signal in the TF domain and then transformed into the time domain using ISTFT.

3.4 Experimental Setup

Simulation experiments were performed using speech utterances from TIMIT [209] and noise signals from NOISEX-92 [210] and all signals were resampled to 10 kHz. To simulate binaural signals and directional and isotropic noise, the head related impulse responses (HRIRs) from [211] were used. To generate the training target HSWOBM masks, 4000 speech signals from TIMIT and the anechoic in-ear HRIRs were used. The source azimuth was randomly selected between $-90°$ and $+90°$ with a resolution of $5°$ and with an elevation of $0°$. The distance between the source and the head and torso simulator (HATS) was either 80 cm or 300 cm. The target masks for the left and right channels were computed as described in Sec. 3.3.2. A feed-forward DNN was designed in Python using TensorFlow. Around 3 million trainable parameters were generated from the three extracted feature sets. The DNN was trained to estimate a continuous mask to optimize the WSTOI metric in each channel by minimizing the cost function in (3.34). The generated trainable data was split into 70% for training, 15% for validation and 15% for testing while training the model. For evaluation, 3000 unseen male and female speech utterances were randomly selected from the TIMIT test dataset and VCTK corpus [212]. Five types of noise signals from NOISEX-92 were used and spatialized using HRIRs [211] were used to generate the test data. Isotropic noise was generated using uncorrelated noise sources uniformly spaced every $5°$ in the azimuthal plane [186] using HRIRs from [211] with the source distance of 300 cm. Experiments with different SNRs were performed by adding isotropic noise to normalized binaural speech signal so that $(SNR_L + SNR_R)/2$

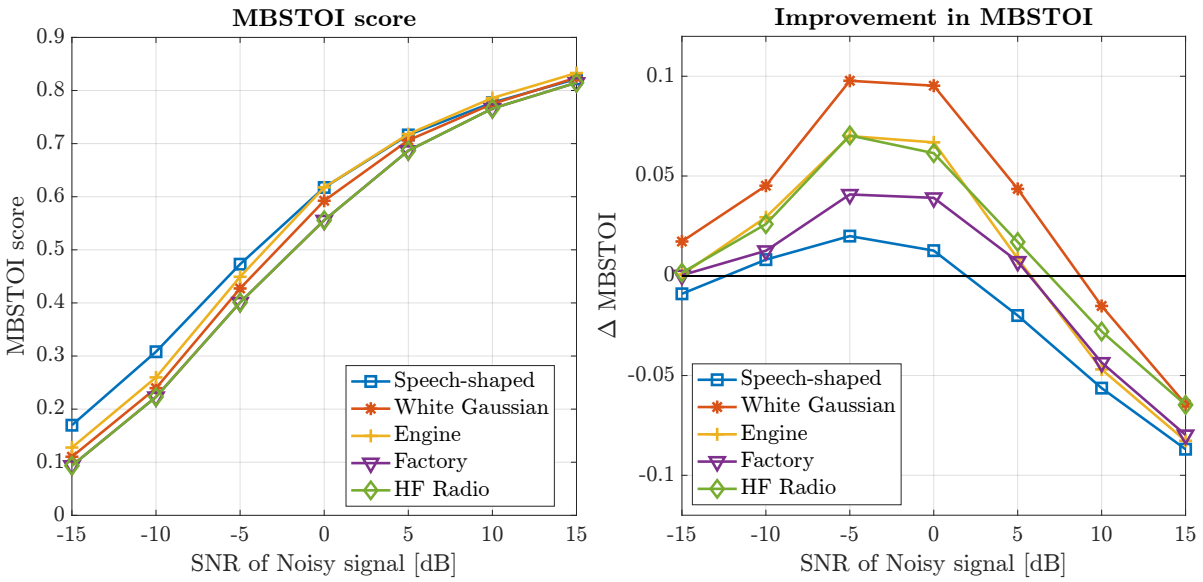
lies between -15 dB to +15 dB in steps of 5 dB. The A-weighted fwSegSNR [161, 213] is measured to show the noise reduction, MBSTOI [174] to show the binaural intelligibility improvement and the ILD error to quantify the preservation of binaural cues. The better ear WSTOI is estimated by choosing the maximum score between the left and right ear channels. The ILD for binaural signals is computed in each TF bin as

$$\mathrm{ILD}_X(k,\ell) = 20\log_{10}\left(\frac{|X_L(k,\ell)|}{|X_R(k,\ell)|}\right), \tag{3.39}$$

where X_L and X_R are the left and right channel STFT domain signals for the clean speech signals. The RMS ILD error is computed by averaging the error over all the frequencies. The RMS ILD error was averaged over frequency because in the experiments, no significant changes in the error were observed over frequencies. The proposed method does not alter the phase of the noisy signal and reuses the phase of the unprocessed signal while processing and therefore ITDs of the enhanced signal was found to be the same.

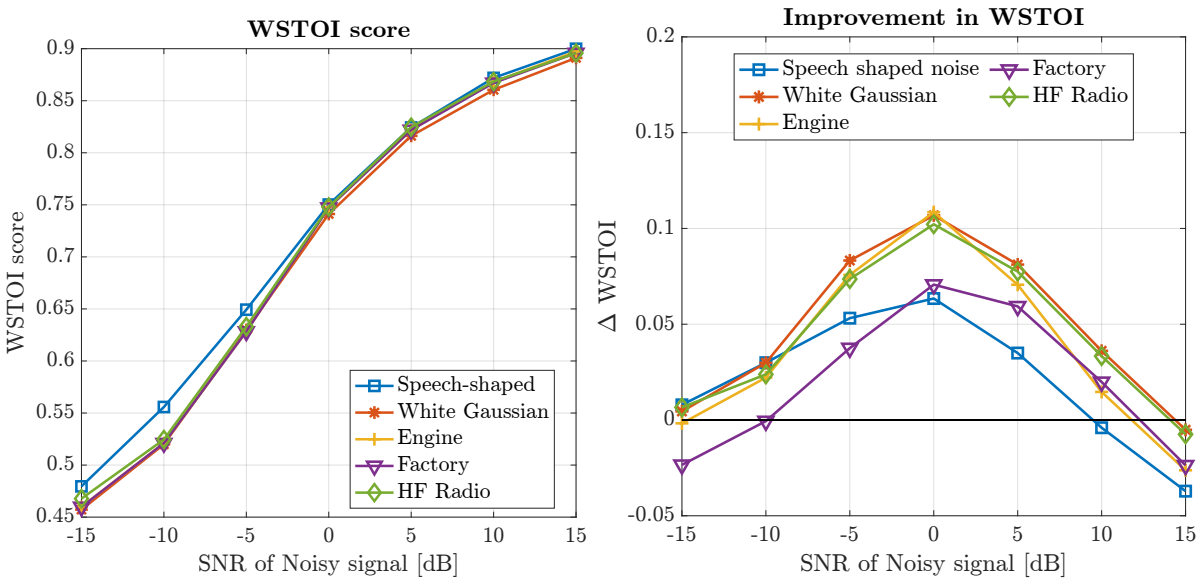
3.5 Results

Figure 3.2(a) shows the MBSTOI score and Fig. 3.2(b) shows the improvement in MBSTOI of the enhanced signals for different SNRs. The maximum improvement in MBSTOI was observed between -5 to 0 dB SNR for all noise cases. Similarly, Fig. 3.3(a) shows the better ear WSTOI score and Fig. 3.3(b) shows the improvement in better ear WSTOI of enhanced signals for different SNRs. In very noisy conditions, WSTOI score is higher than the MBSTOI score. MBSTOI is more susceptible to certain noise conditions such as modulating noise and results in a lower score [174]. WSTOI on the other hand is more robust and the use of language models is shown to correlate better with intelligibility [193]. At very noisy SNRs of under -10 dB, extraction of speech features from the signal to estimate an accurate mask becomes difficult, which explains lower improvements. For input SNRs higher than 5 dB, there is degradation in MBSTOI, as the input binaural speech signal has a higher MBSTOI score before processing and applying mask-based enhancement on these signals results in reducing the score. As all the signals above 5 dB SNR had a MBSTOI score above 0.7 after processing, the overall intelligibility of speech signals did not



(a) **(b)**

Figure 3.2: (a) MBSTOI score of the enhanced speech and (b) improvement in MBSTOI score at different input SNRs.



(a) **(b)**

Figure 3.3: (a) Better ear WSTOI score of the enhanced speech and (b) improvement in better ear WSTOI score at different input SNRs.

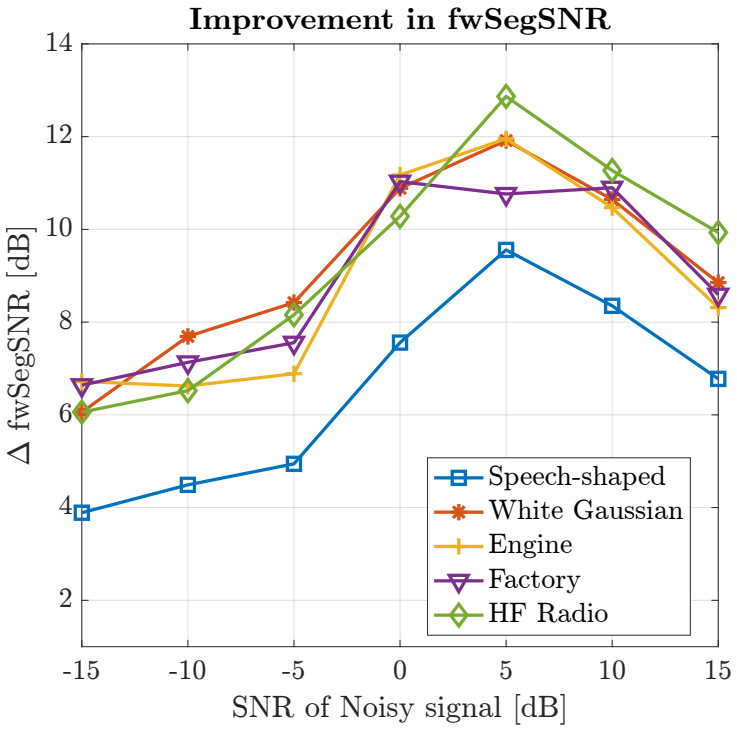
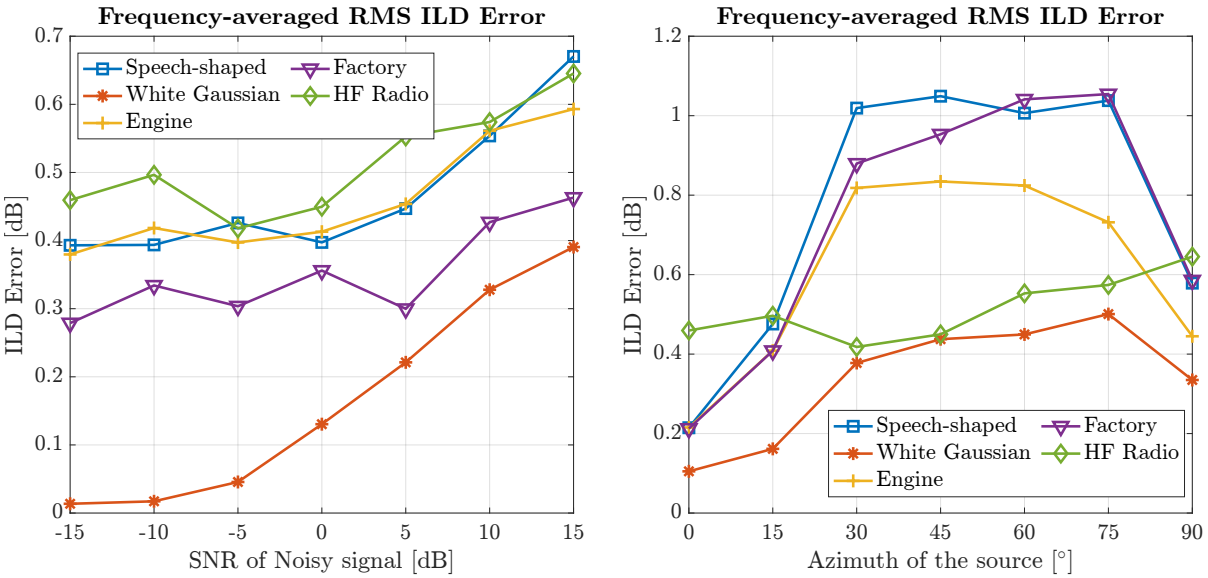


Figure 3.4: Improvement in frequency weighted segmental SNR for both channels

deteriorate to unintelligible levels. A similar trend is observed in the better-ear WSTOI score but with a lower level of degradation at higher input SNRs. In Fig. 3.4, which shows the improvement in fwSegSNR, even under very noisy cases of -15 dB SNR where the speech information in the extracted features is low, the method was able to provide at least 3 dB fwSegSNR improvement for all the noise cases on average and as the input SNR of the noisy speech improved, fwSegSNR improvements greater than 10 dB were also observed. From Fig. 3.4, it can be seen that a similar improvement in fwSegSNR can be observed for all noise cases with speech-shaped noise being the exception, and also from Fig. 3.2(b), the lowest improvement for MBSTOI can be seen for speech-shaped noise. Although speech-shaped noise is stationary noise, most of its energy is located in the TF bins shared with speech, thus affecting the estimation of SPP and hence the performance. For experiments with multiple azimuth angles, the source was simulated at azimuths from 0° to 90° since the other half of the frontal plane in the HRIRs database [211] is a reflection of the former and to avoid the front-back confusion caused in binaural signals. Figure 3.5(a) shows the frequency averaged RMS ILD error observed with the method for input SNRs of the noisy signal and Fig. 3.5(b) shows the error for different azimuth angles



(a) **(b)**

Figure 3.5: (a) Frequency averaged RMS ILD error with different SNRs, for a source placed 80 cm away at 30° azimuth and noise at 0° azimuth and (b) frequency averaged RMS ILD error for a source at 80 cm with 0 dB SNR for different azimuths of the source.

of the source. For all the noise cases, SNRs and azimuths the RMS ILD error observed is under 1 dB in the frontal plane. Studies have shown that human listeners can notice significant differences in ILD if the error exceeds more than 1.5 to 2 dB [7, 10, 11]. For all azimuths and SNRs, the proposed method performed the best for white Gaussian noise having the least error among all the noise cases. This low RMS ILD error after processing quantifies the binaural cue preservation of the method.

3.6 Conclusions

In this chapter, a new approach using STOI-optimal masking to enhance noisy binaural speech signals has been presented. This is a hybrid method which uses a combination of classical digital signal processing (DSP) techniques and DNNs. The method improves the SNR and intelligibility of the noisy signal while preserving the interaural cues. Instead of applying the TF masks as a gain function, SPP is calculated and used by an OM-LSA speech enhancer ensuring a smooth imposition of the estimated masks. The proposed method is able to provide significant improvements in fwSegSNR and MBSTOI score with

minimal distortion of interaural cues. In the next chapter, end-to-end complex-valued neural networks (CVNNs) using convolutional encoder-decoder (CED) architecture for binaural speech enhancement is introduced.

Chapter 4

Binaural Speech Enhancement Using Deep Complex Convolutional Networks

This chapter presents an end-to-end binaural speech enhancement method using complex convolutional neural networks with an encoder-decoder architecture. Two models are studied, firstly, a complex self-attention transformer block placed in between the encoder and decoder and secondly, a complex recurrent block. A novel loss function that focuses on the preservation of spatial information along with speech intelligibility and noise reduction. The network is trained to estimate individual complex ratio masks in the time-frequency domain for the left and right-ear channels of binaural hearing devices. In acoustic situations with a single target speaker and isotropic noise of various types, the proposed method significantly reduces the noise and improves the estimated speech intelligibility compared to the binaural TasNet [1] and the binaural STOI-optimal masking method from Ch. 3. An earlier version of the work presented in this chapter has been published in [214, 215].

4.1 Introduction

As previously discussed in Ch. 2, binaural speech enhancement has recently become the state-of-the-art approach for improving speech clarity in hearing aids and augmented/virtual reality devices [5, 216]. Binaural signals inherently contain the spatial characteristics of sounds, providing essential information for accurate sound source localization [183]. Additionally, binaural unmasking effects have been shown to significantly increase speech intelligibility, highlighting the importance of preserving interaural cues in binaural signals while reducing noise [8]. The primary cues beneficial for localizing sounds and improving speech intelligibility are ILD and ITD or interaural phase differences (IPD) [7].

Several methods for binaural speech enhancement have been proposed, including multichannel Wiener filters [33, 36], beamforming [216], and mask-based enhancement techniques [181, 186]. For instance, a time-domain CED model for binaural speech separation achieved state-of-the-art performance [1]. Unlike binaural methods, monaural speech enhancement approaches enhance each binaural channel independently, which often results in the degradation of essential binaural cues. However, monaural speech enhancement using deep learning techniques has yielded significant results in both the time domain [217, 218] and the TF domain [219, 220].

In the TF domain, spectrograms are commonly used as input to networks for speech enhancement [89, 181, 220]. Many TF domain methods focus solely on magnitude-based enhancement, relying on the noisy TF-domain phase for reconstructing the enhanced speech signal [90, 181]. To address optimal phase estimation for signal reconstruction, some approaches jointly estimate both magnitude and phase by utilizing complex-valued spectrograms. Monaural speech enhancement methods using complex-valued networks have shown promising results and can outperform real-valued networks [89,90]. The convolutional recurrrent network (CRN) introduced in [219] employed a CED architecture with long short-term memory (LSTM) blocks placed between the encoder and decoder. Additionally, attention-based transformer neural networks (TNN) have demonstrated state-of-the-art performance in natural language processing (NLP) problems compared to other DNN models [153]. Speech enhancement using attention models has shown promising results as well [90].

In [89], a deep complex CRN was trained to optimize the scale invariant SNR (SI-SNR) for monaural speech signals. However, applying a similar approach to binaural

signals could damage interaural cues. Specifically, for binaural signals, phase information is crucial for preserving IPD values, and the enhanced signals should retain ILD consistent with the original signal. Even if the model achieves significant noise reduction and improves speech intelligibility, altering the level and phase information would modify the spatial information of the target, compromising localization and spatial awareness [7, 8]. Binaural signals can exhibit different speech and noise levels in each channel. The models are trained to estimate separate complex ratio masks (CRMs) for each binaural channel, differing from the approach proposed in Ch. 3. While applying the same mask to both channels preserves interaural cues, it also results in a similar level of noise reduction across channels, which is not optimal for binaural signals with varying speech and noise levels.

This chapter proposes two novel models for binaural speech enhancement that aim to address these challenges:

- **Binaural Complex Convolutional Transformer Network (BCCTN)**: This model utilizes phase-aware training [89, 110] for binaural speech signals and uses a CED structure with a complex-valued attention-based transformer block between the encoder and decoder blocks.

- **Binaural Complex Convolutional Recurrent Network (BCCRN)**: This model leverages a complex-valued CED-based recurrent network architecture for binaural speech enhancement. Compared to BCCTN, BCCRN utilizes a smaller complex recurrent network, maintaining high performance with reduced computational complexity.

The advantages of using CVNNs for speech and audio processing applications are detailed in Sec. 2.4. For binaural signals, the relationship between magnitude and phase is vital to preserve spatial information. Both models emphasize the preservation of ILD and IPD cues to ensure that the spatial information remains intact, providing listeners with improved speech intelligibility and accurate sound source localization in noisy environments. These approaches represent significant advancements in binaural speech enhancement, leveraging the strengths of complex-valued networks to deliver superior performance.

This chapter is organized as follows, Sec. 4.2 introduces the architecture for both the models, Sec. 4.3 describes the signal model and loss function used to train the model. The

training setup, baselines and datasets are detailed in Sec. 4.4 and the results are discussed in Sec. 4.5 followed by the conclusion in Sec. 4.8.

4.2 Model Architecture

4.2.1 Binaural Complex Convolutional Transformer Network (BCCTN)

The proposed BCCTN model uses a CED structure with a transformer block between the encoder and decoder and is trained to estimate an individual CRM for each channel. The block diagram of the architecture is shown in Fig. 4.1. A CED architecture for monaural speech enhancement has been previously studied in [89,217,219]. The proposed architecture is newly modified to work with binaural signals by using individual encoder and decoder blocks for each channel. The STFT blocks transform the signals into the TF domain. The encoder block is made of 6 complex convolutional layers with parametric rectified linear unit (PReLU) activation and employs batch normalization. The convolutional encoder blocks help in identifying the local patterns in the input spectrogram [89, 217]. Individual encoder blocks are used for the left and right-ear channels as the network needs to estimate two individual CRMs. The encoded information from both channels is concatenated and supplied as the input to the transformer. The transformer block consists of multi-head attention layers based on the architecture proposed in [153]. The structure of the complex transformer is shown in Fig. 4.2. The real and imaginary output of the transformer H_{n+1} for the $(n+1)^{th}$ hidden state are given by

$$H^r_{n+1} = (H^r_n \circledast H^r_n) - (H^i_n \circledast H^i_n), \tag{4.1}$$

$$H^i_{n+1} = (H^r_n \circledast H^i_n) + (H^i_n \circledast H^r_n), \tag{4.2}$$

where H^r_n and H^i_n are the real and imaginary parts of the encoder output H_n. The multi-head attention operation is denoted by $\circledast$. The transformer processes input sequences in parallel which can potentially reduce the computation and inference times. Contrary to sequential processing of recurrent neural networks (RNNs), transformers have a fixed

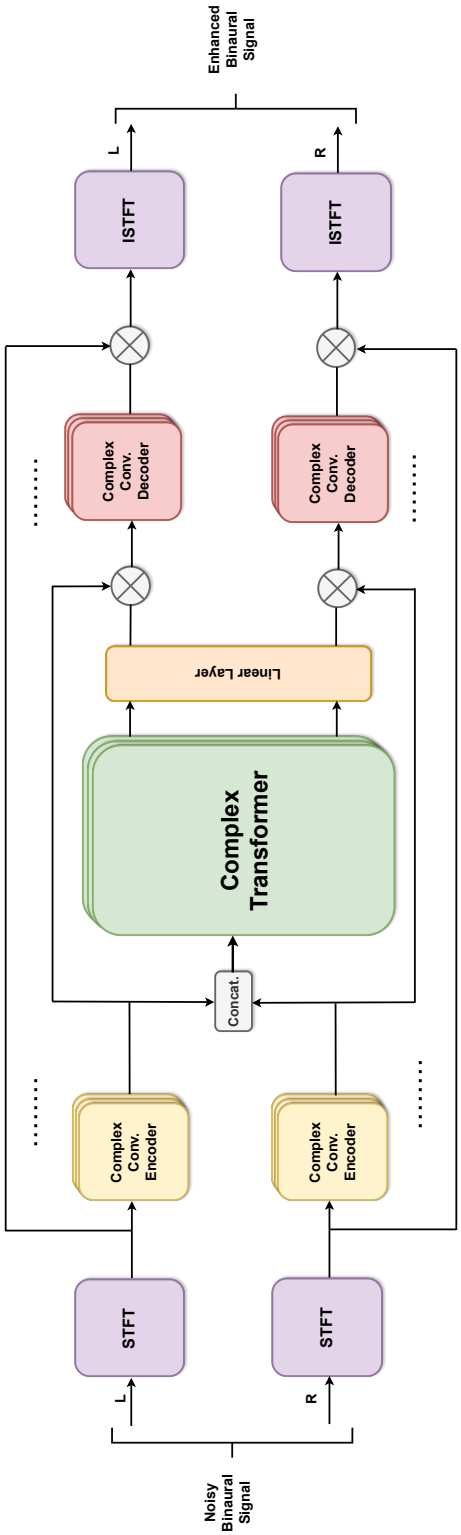


Figure 4.1: Architecture of the proposed model

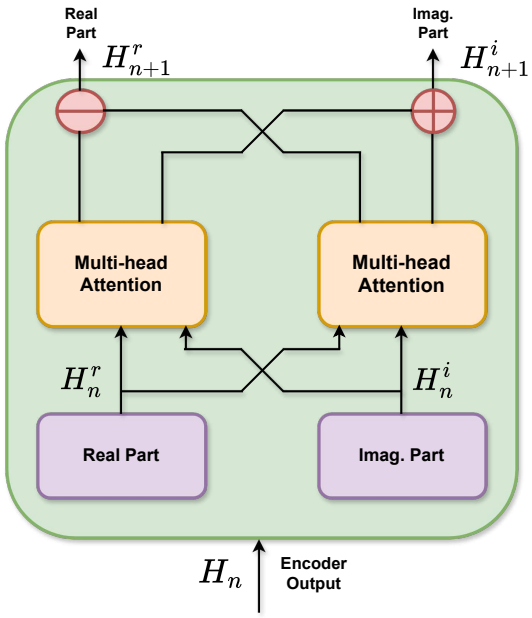


Figure 4.2: The complex transformer block which implements (4.1) and (4.2).

path length before computing the attention weights. This enables faster processing of long signal sequences and establishes relationships between sequences compared to RNNs based on LSTMs or gated recurrent units (GRUs) [153]. The fixed path length property was key to solving many NLP problems but was not very successful in dealing with the physical characteristics of audio signals which have closer correlated components. Hence, positional encoding is required when using transformer networks to ensure that less emphasis is placed on distant correlations [90]. The transformer block focuses on identifying relationships within the encoded information from both channels [90, 153]. The convolutional decoder consists of 6 transposed complex convolutional blocks which are symmetric in design to the convolutional layers of the encoder to reconstruct the signal to its original size using the processed feature information from the transformer. Skip connections are placed between each encoder and decoder layer, denoted by $\cdots\cdots$ in Fig. 4.1 based on the CRN architecture [219], which concatenates the output of each encoder block to the decoder layer. This improves the information flow and facilitates network optimization [219]. The left and right channel decoders output individual CRMs that are applied to the noisy binaural signal for enhancement. The ISTFT blocks in Fig. 4.1 transform the

enhanced TF domain signal back into the time domain. Implementation code is available online [221].

4.2.2 BCCRN

The proposed Binaural Complex Convolutional Recurrent Network (BCCRN) is trained to estimate an individual CRM for each channel. The block diagram of the model architecture is shown in Fig. 4.3. The encoder and decoder blocks are similar in structure and dimensions to those in the BCCTN model. The encoder extracts high-level features from the input spectrograms and reduces the resolution of the input. The decoder rebuilds the low-resolution features back to the original size. The encoded information from the left and right encoder blocks is concatenated and provided to the complex LSTM block. The complex-valued LSTM blocks are designed analogous to the transformer blocks on the basis of complex arithmetic. The LSTM layers are designed to model the frequency dependencies. Skip connections are placed between the encoder and decoder layers based on the CRN [63] and U-Net architecture [99], improves the information flow and facilitates network optimization [219]. The individual decoders output individual CRMs that are applied to the left and right channels of the noisy binaural signal for enhancement. The ISTFT blocks transform the signal back into the time domain.

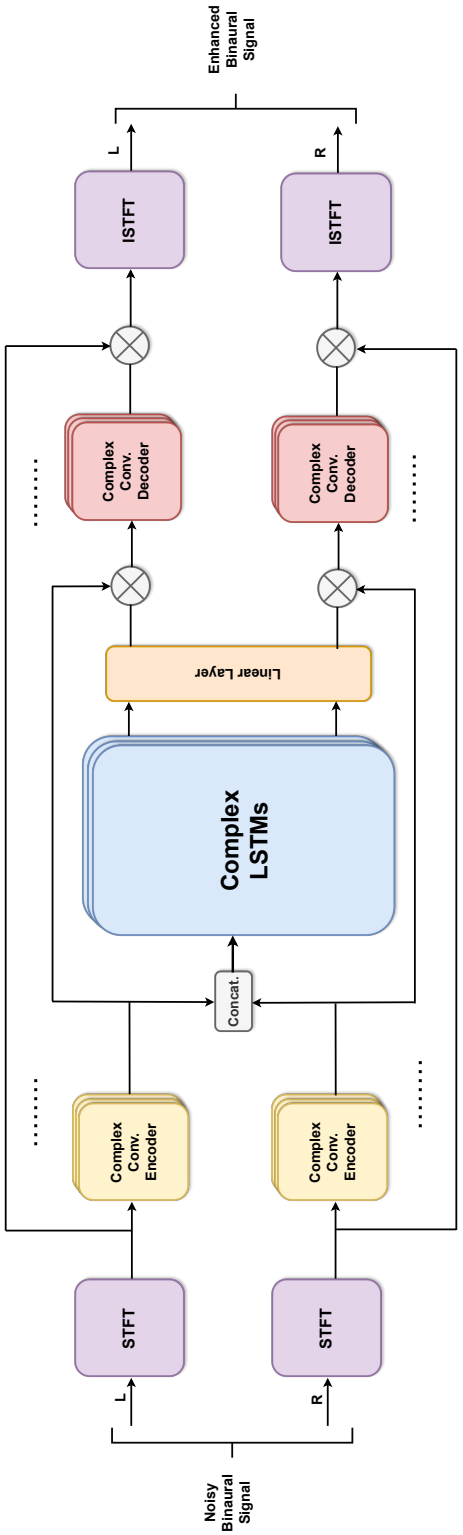


Figure 4.3: Architecture of the proposed BCCRN model

4.3 Signal Model and Loss Function

The noisy time-domain input signal for the right channel, y_R, is given by

$$y_R(t) = s_R(t) + v_R(t), \tag{4.3}$$

where s_R is the anechoic clean speech signal, v_R is the noise and t is the time index. The STFT is used to transform the signals into the TF domain and the respective TF representations are $Y_R(k,\ell)$, $S_R(k,\ell)$ and $V_R(k,\ell)$ with k and ℓ being the frequency and time frame indices respectively. The left channel is described similarly with an L subscript. For brevity, the L and R indices are omitted where possible. The estimated CRM, $M(k,\ell)$ is applied to the noisy signal $Y(k,\ell)$ to obtain the enhanced speech signal $\hat{S}$. Individual channels are enhanced by applying the estimated CRM, $(M_r + jM_i)$ to the complex-valued noisy signal $(Y_r + jY_i)$ in the TF domain (omitting k and ℓ indices for brevity),

$$\hat{S}_r + j\hat{S}_i = (M_r + jM_i) \cdot (Y_r + jY_i)\,, \tag{4.4}$$

where r and i indicate the real and imaginary parts. The computed CRM is given by

$$M_r + jM_i = \frac{\hat{S}_r + j\hat{S}_i}{Y_r + jY_i} = \frac{Y_r\hat{S}_r + Y_i\hat{S}_i}{Y_r^2 + Y_i^2} + j\frac{Y_r\hat{S}_i - Y_i\hat{S}_r}{Y_r^2 + Y_i^2}. \tag{4.5}$$

4.3.1 Loss Function

The loss function for model training consists of four terms and optimizes the network for noise reduction, intelligibility improvement, and interaural cue preservation. The loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \alpha\mathcal{L}_{\text{SNR}} + \beta\mathcal{L}_{\text{STOI}} + \gamma\mathcal{L}_{\text{ILD}} + \kappa\mathcal{L}_{\text{IPD}}, \tag{4.6}$$

where $\mathcal{L}_{SNR}$ is the SNR loss, $\mathcal{L}_{STOI}$ is the STOI [208] loss, and $\mathcal{L}_{ILD}$ and $\mathcal{L}_{IPD}$ are the proposed ILD and IPD error losses which are functions of both $\hat{S}_L$ and $\hat{S}_R$. The parameters α, β, γ, and κ are the weights applied to each term.

$\mathcal{L}_{\text{SNR}}$ is defined as the mean of the left and right-ear channel values and append a negative sign to maximize the SNR value, such that

$$\mathcal{L}_{\text{SNR}} = -\frac{(\text{SNR}_L + \text{SNR}_R)}{2}. \tag{4.7}$$

The SNR of the enhanced signal, $\hat{\mathbf{s}}$, is defined as

$$\text{SNR}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left(\frac{\|\mathbf{s}\|^2}{\|\mathbf{e}_{noise}\|^2} \right), \tag{4.8}$$

where $\mathbf{e}_{noise} = \hat{\mathbf{s}} - \mathbf{s}$ with $\mathbf{s}$ and $\hat{\mathbf{s}}$ being the clean and enhanced signal vectors respectively, and $\|.\|$ is the L2 norm.

To optimize the intelligibility of the enhanced speech signals, STOI [208, 222] is computed for left and right channels individually and averaged, and a negative sign is appended to maximize the STOI so that

$$\mathcal{L}_{\text{STOI}} = -\frac{(\text{STOI}_L + \text{STOI}_R)}{2}. \tag{4.9}$$

Including $\mathcal{L}_{\text{STOI}}$ helps the network optimize individual CRMs to maximize intelligibility in the enhanced signals.

To ensure the preservation of interaural cues in the enhanced binaural speech using two separate CRMs, minimizing the ILD and IPD errors of the target speech is enforced using the loss function. The ILD and IPD for the clean speech signal are given by

$$\text{ILD}_S(k, \ell) = 20 \log_{10} \left(\frac{|S_L(k, \ell)|}{|S_R(k, \ell)|} \right), \tag{4.10}$$

$$\text{IPD}_S(k,\ell) = \angle\left(\frac{S_L(k,\ell)}{S_R(k,\ell)}\right). \tag{4.11}$$

The ILD and IPD for the enhanced speech, $\hat{S}$, are calculated similarly. In order to restrict the ILD and IPD errors to the speech-active regions, an IBM [48] $\mathcal{M}$ is computed by selecting the TF bins which have energy above a threshold. The energy $E(k,\ell)$ of the clean signal is given by

$$E(k,\ell) = 10\log_{10}|S(k,\ell)|^2. \tag{4.12}$$

The IBM $\mathcal{M}(k,\ell)$ that defines the speech active TF tiles is then defined as,

$$\mathcal{M}(k,\ell)_{L/R} = \begin{cases} 1 & E(k,\ell) > \max\limits_{\ell}\left(E(k,\ell)\right) - \mathcal{T} \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

where $\max_l\left(E(k,\ell)\right)$ is the maximum energy computed for each frequency bin, k. Individual IBMs, $\mathcal{M}_L$ and $\mathcal{M}_R$ are computed for the left and right-ear channels. The final mask $\mathcal{M}$ is obtained by choosing the bins that have energy above the threshold, $\max_\ell\left(E(k,\ell)\right) - \mathcal{T}$, in both channels and is given by

$$\mathcal{M}(k,\ell) = \mathcal{M}_L(k,\ell) \odot \mathcal{M}_R(k,\ell), \tag{4.14}$$

where $\odot$ denotes the Hadamard product. As an example, Figure 4.4 shows the spectrograms of the clean speech signal and the corresponding target speech-based binary mask. For training and evaluation $\mathcal{T} =$ 10 dB was used [48]. The human auditory system interprets ILDs and IPDs differently based on the frequency range. Human spatial hearing relies primarily on interaural phase difference cues for frequencies below 1500 Hz, while

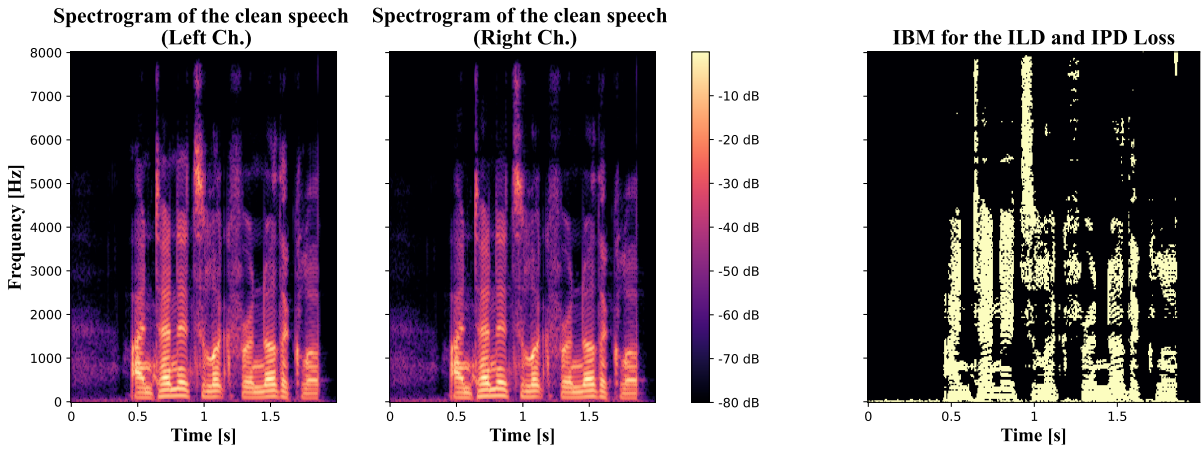


Figure 4.4: Spectrograms of the left and right-ear clean speech signals and the corresponding IBM computed for interaural cue error masking.

interaural level difference cues play a crucial role for frequencies above 1500 Hz [7]. From the estimated IBM $\mathcal{M}(k,\ell)$, separate masks for ILD and IPD cues are computed based on human spatial hearing. For choosing the IPD cues in the speech active bins below $f_p = 1500$ Hz, $\mathcal{M}(k,\ell)$ for $k \leqslant K_p$ is selected where

$$K_p = \frac{f_p \times N_{fft}}{f_s} \tag{4.15}$$

with N_{fft} and f_s being the FFT length and sampling frequency respectively. Similarly, for choosing the ILD cues $\mathcal{M}(k,\ell)$ for $k > K_p$ is selected. The $\mathcal{L}_{ILD}$ and $\mathcal{L}_{IPD}$ terms are given by

$$\mathcal{L}_{ILD} = \frac{1}{N_{ld}} \sum_{k>K_p,\ell} \mathcal{M}(k,\ell) \left(|ILD_S(k,\ell) - ILD_{\hat{S}}(k,\ell)| \right), \tag{4.16}$$

$$\mathcal{L}_{IPD} = \frac{1}{N_{pd}} \sum_{k\leqslant K_p,\ell} \mathcal{M}(k,\ell) \left(|IPD_S(k,\ell) - IPD_{\hat{S}}(k,\ell)| \right) \tag{4.17}$$

where

$$N_{ld} = \sum_{k>K_p,\ell} \mathcal{M}(k,\ell)$$
$$N_{pd} = \sum_{k\leqslant K_p,\ell} \mathcal{M}(k,\ell)$$

are the total number of speech-active frequency and time bins determined from the mask for ILD and IPD cues respectively. To preserve the interaural cues of the target speaker, the network optimization is guided by using masks based on the target speech. The errors in ILD and IPD are calculated in the time-frequency domain, while losses related to SNR and STOI are computed in the time domain through waveform synthesis using the ISTFT.

4.4 Simulations

This section details the generation of the training dataset and the network parameter dimensions of the BCCTN and BCCRN models. Subsequently, the training setup for these models is discussed. Following this, feature maps illustrating the flow and transformation of information within the models are presented. The baselines employed in the experiments are described in the ensuing section.

4.4.1 Datasets

To generate binaural speech data, monaural clean speech signals were taken from the CSTR VCTK corpus [212] and were spatialized using the measured HRIRs from [211]. The speech corpus [212] has around 13 hours of speech data uttered by 110 English speakers with various accents that were used to generate 2-second speech utterances and spatialized to have the left and right-ear channels. The dataset was made of 20000 speech utterances which were split into training (70%), validation (15%), and testing (15%) sets. Unseen speech data from the TIMIT corpus [209] were also used for evaluation. Reverberant speech signals for evaluation were generated using binaural room impulse responses (BRIRs) from [211] and were placed in isotropic noise fields for the anechoic

signals. Rooms with T_{60} varying from 0.3 to 1.2 s were used. Noise signals from the NOISEX-92 database [210] were used to generate diffused isotropic noise. Isotropic noise was generated using uncorrelated noise sources uniformly spaced every 5° in the azimuthal plane [186] using HRIRs from [211]. A sampling rate of 16 kHz was utilized for all signals. Binaural signals were generated with the target speech placed at a random azimuth in the frontal plane (−90° to +90°), using the HRIRs from [211] recorded using a HATS. For training, isotropic noise was added to the VCTK corpus [212] so that $(SNR_L + SNR_R)/2$ lies between -7 dB and 16 dB. The noise types used for training are white Gaussian noise (WGN), speech shaped noise (SSN), factory noise, and office noise and, for evaluation, an additional car engine noise was included. The datasets were generated in the anechoic condition for training. The evaluation set consists of speech signals from the VCTK corpus [212] (i.e, "matched' condition) and the TIMIT [209] (i.e, "unmatched condition') with random target azimuth and isotropic noise added at a random SNR between -6 dB and 15 dB. The speaker was always placed at 0° elevation and at a distance of either 80 cm or 300 cm chosen randomly for each signal.

4.4.2 BCCTN - Training setup

Table 4.1 shows the model parameter dimensions for the BCCTN model. An FFT length of 512, a window length of 25.6 ms, hop length of 6.25 ms and a Hanning window was used to compute the STFT of the speech signals. The number of channels used in the model's convolutional layers for the encoder and decoder blocks layers are $\{16, 32, 64, 128, 256, 256\}$, with a stride of 2 in the frequency and 1 in the time dimension with a kernel size of (5,1). All the convolutions in these layers are causal and in the frequency dimension. The Multihead attention block has an embedded dimension of 512 for real and imaginary blocks shown in Fig. 4.2, a hidden size of 128, and 32 heads. The model was implemented with Pytorch which provides native complex data support for most of the functions. The linear layer placed after the transformer block has an input and output feature size of 1024. The Pytorch model was trained using the Adam optimizer, an initial learning rate of 0.001, and a multi-step learning rate scheduler to modify the learning rate with the validation loss. The model has around 10 million parameters and was trained for 100 epochs with an additional early stopping condition of no improvement in the validation loss for three consecutive epochs. The loss functions

Block	**Ear Channel**	**Function**	**Dimension**
Input	L/R	STFT	$B \times C_{in} \times F \times T$
Encoder 1	L/R	Conv2D	$B \times 8 \times 128 \times T$
Encoder 2	L/R	Conv2D	$B \times 16 \times 64 \times T$
Encoder 3	L/R	Conv2D	$B \times 32 \times 32 \times T$
Encoder 4	L/R	Conv2D	$B \times 64 \times 16 \times T$
Encoder 5	L/R	Conv2D	$B \times 128 \times 8 \times T$
Encoder 6	L/R	Conv2D	$B \times 128 \times 4 \times T$
Attention Input	L&R	Concat.	$B \times 256 \times 4 \times T$
Flatten	L&R	Flatten	$B \times 1024 \times T$
Transformer	L&R	Attention	$B \times 1024 \times T$
Linear	L&R	Fully-Connected	$B \times 1024 \times T$
Unflatten	L&R	Unflatten	$B \times 256 \times 4 \times T$
Decoder Input	L&R	Transposed Conv2D	$B \times 256 \times 4 \times T$
Decoder Output	L/R	Transposed Conv2D	$B \times C_{in} \times F \times T$

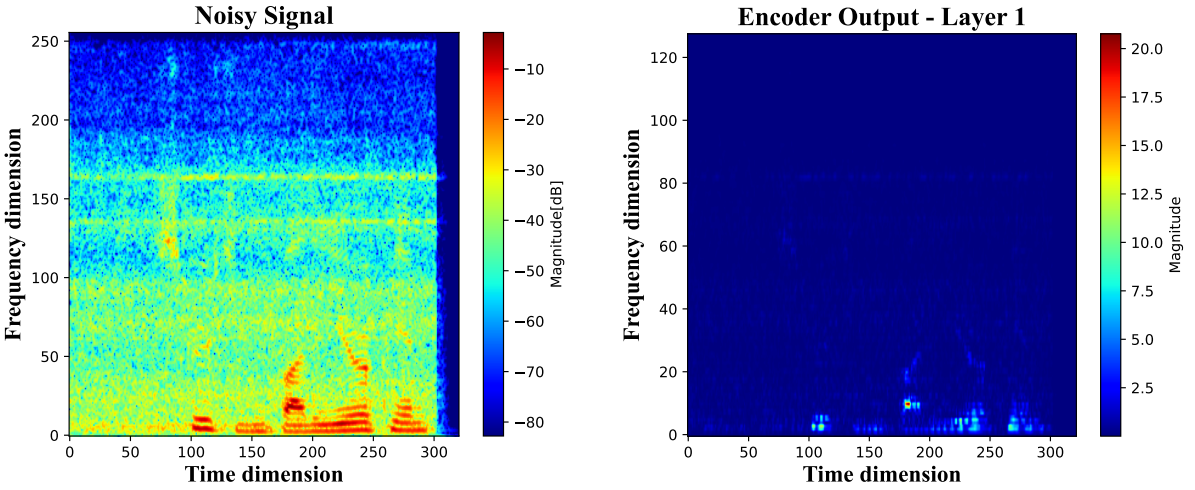
B - Batch size, C_{in} - input channel size of the encoder block, F - number of frequency bins, T - number of time frames, L/R - dimension shown for one channel i.e, left or right ear channel, L&R - dimension shown for both the ear channels combined. Individual decoder blocks have the same dimension as their respective encoder counterparts and skip connections are placed between them.

Table 4.1: BCCTN model parameter dimensions

weights $\alpha, \beta, \gamma, \kappa$, in (5.9), were set to $\{1, 10, 1, 10\}$ respectively. These weights were chosen experimentally to optimize loss function convergence and equalize the difference in the scale of the respective units of the individual loss function terms, where SNR and ILD are computed in dB, IPD is computed in radians and STOI is a bounded score between 0 and 1. The model was trained with the proposed loss function described in (5.9), named BCCTN-SILP, and for comparison, the model was trained to maximize the SNR, named BCCTN-S from (4.8).

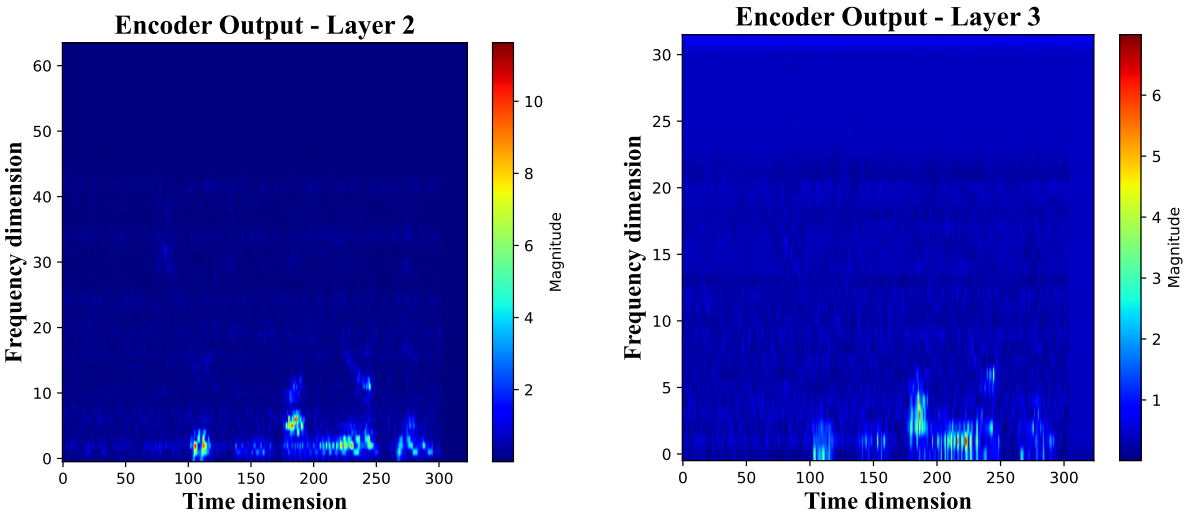
4.4.3 Feature maps of BCCTN

Figures 4.5 - 4.12 show the magnitude spectrum of the noisy signal, feature maps of the encoder, transformers and decoder layers, and the enhanced signal. The feature maps help in understanding the flow and transformation of the latent information in the model. Figure 4.5(a) shows the magnitude spectrum of the left channel of a typical



(a) (b)

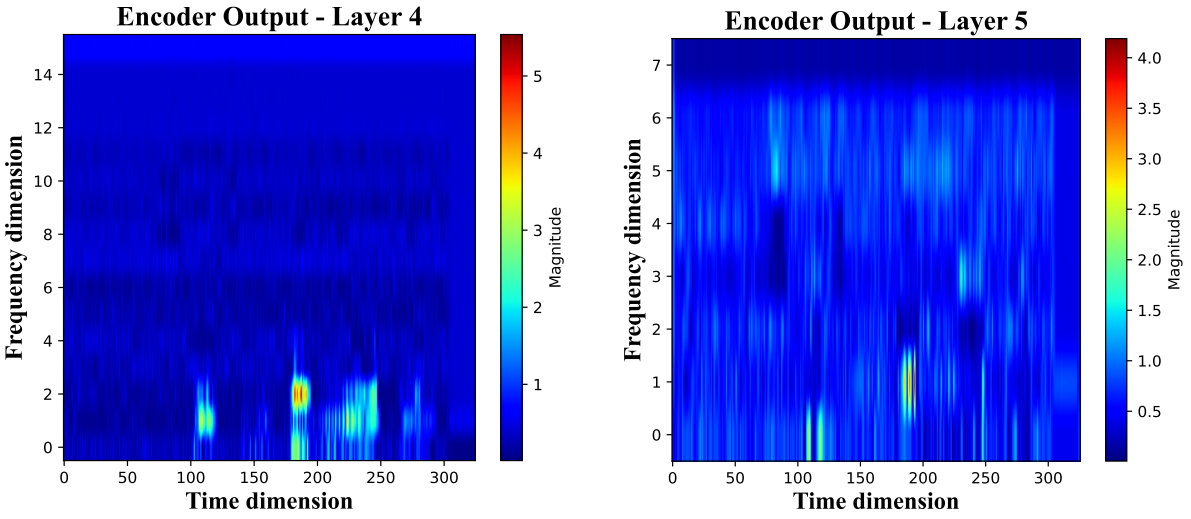
Figure 4.5: (a) Magnitude spectrum of the left channel of a noisy binaural signal, (b) magnitude spectrum of the first encoder layer output of the BCCTN model.



(a) (b)

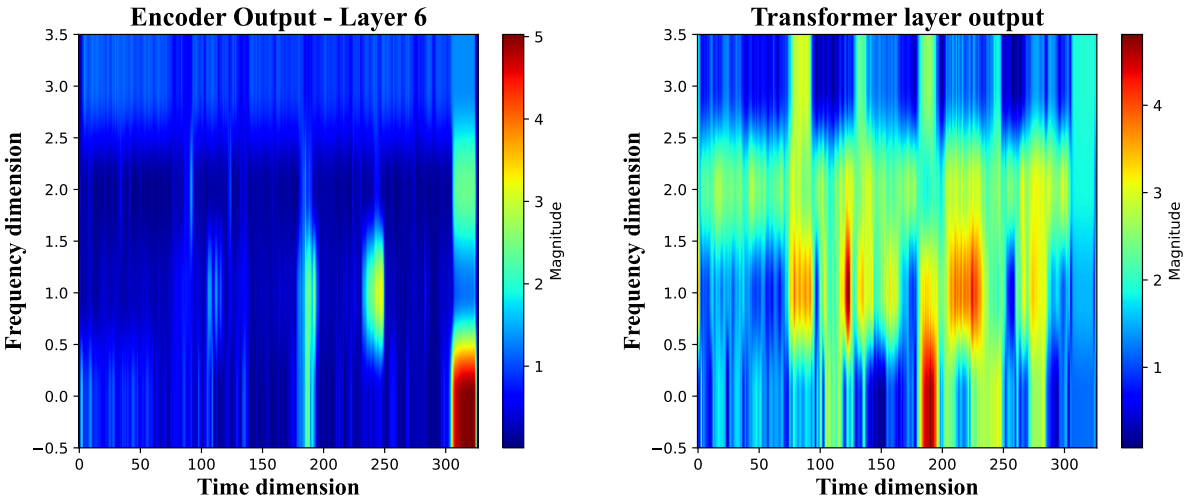
Figure 4.6: (a) Magnitude spectrum of the second and (b) third encoder layer output of the BCCTN model.

binaural noisy signal with a SNR of 3 dB. Figures 4.5(b),4.6(a), 4.6(b), 4.7(a),4.7(b) and 4.8(a) show the flow of information in the first channel of each layer of the encoder block. The 2D-convolutional layers in the encoder block reduce the frequency dimension



(a) **(b)**

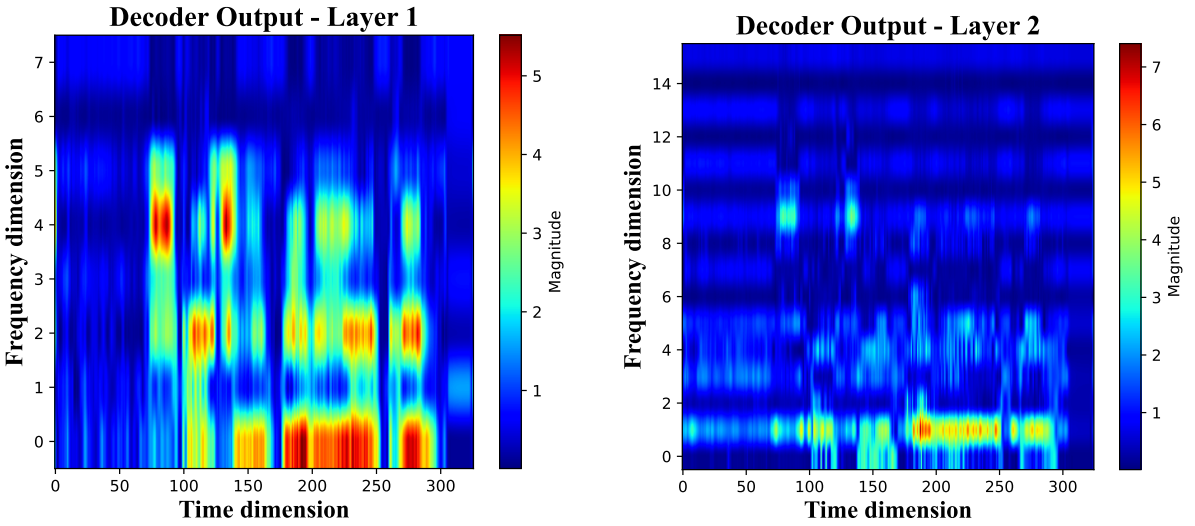
Figure 4.7: (a) Magnitude spectrum of the fourth and (b) fifth encoder layer output of the BCCTN model.



(a) **(b)**

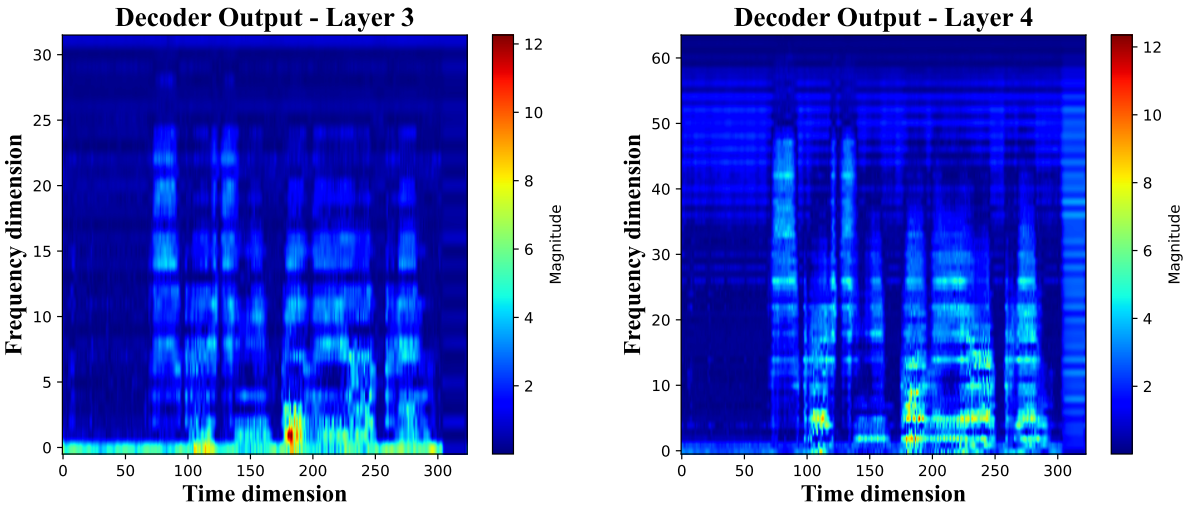
Figure 4.8: (a) Magnitude spectrum of the sixth encoder layer and (b) Transformer layer outputs of the BCCTN model.

of the input while identifying the local patterns in the spectrum. The outputs of the two encoder blocks from the noisy binaural signals are concatenated and provided to the transformer block. Figure 4.8(b) shows the magnitude spectrum of the output of the



(a) **(b)**

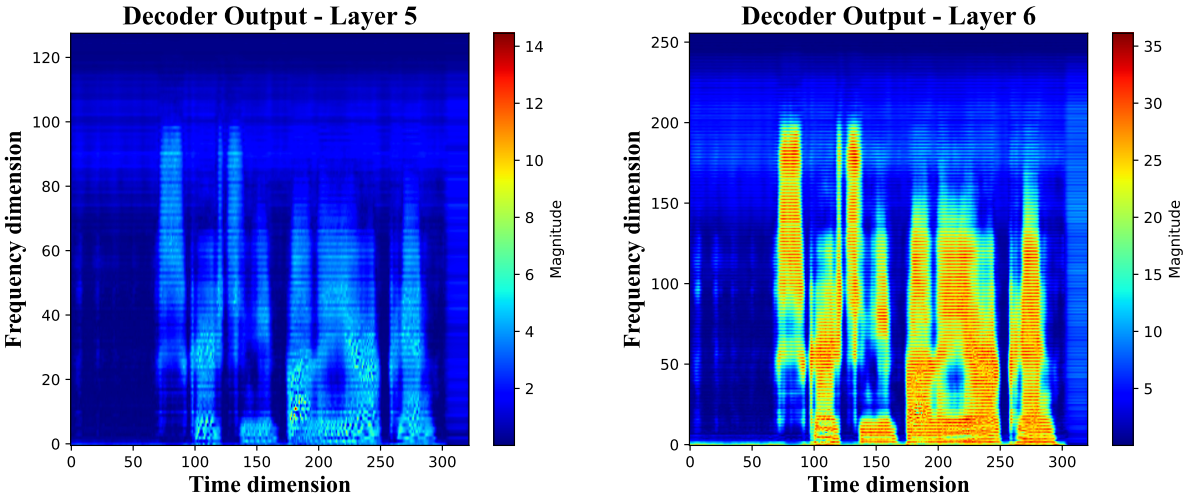
Figure 4.9: (a) Magnitude spectrum of the first and (b) second decoder layer output of the BCCTN model.



(a) **(b)**

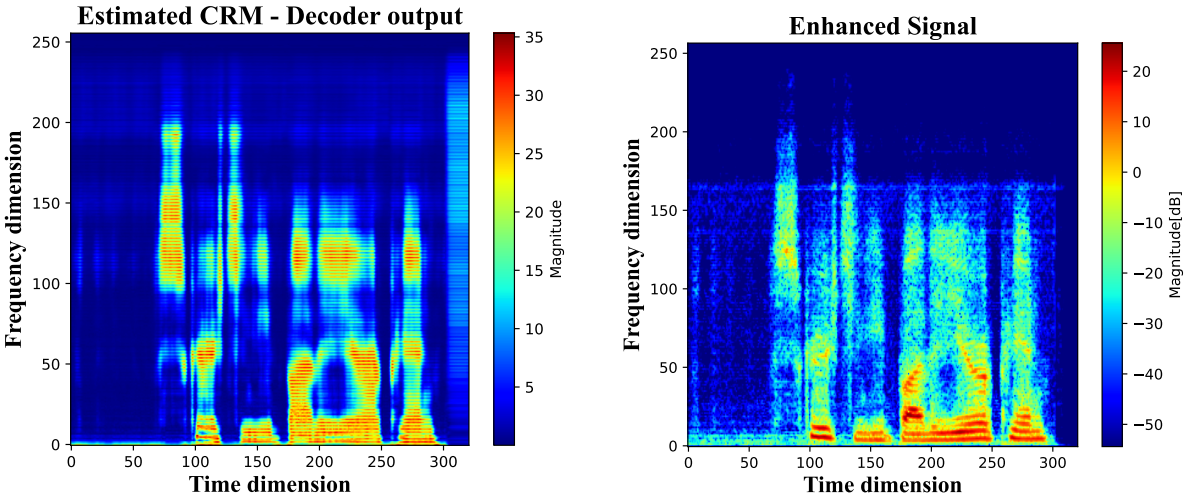
Figure 4.10: (a) Magnitude spectrum of the third and (b) fourth decoder layer output of the BCCTN model.

transformer block. The transformer output has a high magnitude in the TF-regions where there was speech presence which are the regions where the encoder blocks identified the local patterns. The multi-head attentions in the transformer block are causal and in the



(a) **(b)**

Figure 4.11: (a) Magnitude spectrum of the fifth and (b) sixth decoder layer output of the BCCTN model.



(a) **(b)**

Figure 4.12: (a) Magnitude spectrum of estimated CRM and (b) enhanced signal of the BCCTN model.

frequency dimension which can be observed in the magnitude spectrum in Fig. 4.8(b). The output of the transformer block is then provided to the decoder block along with the corresponding information from the the encoder block through the skip connection.

The output of the decoder blocks are shown in Figures 4.9(a) - 4.11(b). The decoders employ transposed Conv2D blocks which complement the encoder blocks and increase the frequency dimension of the information while simultaneously reducing the number of channels. Figures 4.9(a) - 4.11(b) show how the information is transformed to lead up to the desired CRM shown in Fig. 4.12(a) that is applied to the noisy signal to obtain the enhanced signal in Fig. 4.12(b).

4.4.4 BCCRN - Training setup

Table 4.2 shows the parameter dimensions of the BCCRN model. To compute the STFT, an FFT length of 512, a window length of 25.6 ms, and a hop length of 6.25 ms were employed similar to the BCCTN. The number of channels used in the model's convolutional layers for the complex-valued encoder and decoder block layers were $\{16, 32, 64, 128, 128, 128\}$, with a stride of 2 in the frequency and 1 in the time dimension with a kernel size of (5,1) and all the convolutions in these layers are causal and in the frequency dimension. 8 layers of complex-valued LSTMs with a hidden size of 128 were used. The model was implemented with Pytorch which provides native complex data support for most of the functions. The linear layer placed after the recurrent block had an input and output feature size of 1024. The Pytorch model was trained using the Adam optimizer, an initial learning rate of 0.001, and a multi-step learning rate scheduler to modify the learning rate with the validation loss. The model had around 5.7 million parameters and was trained for 100 epochs with an additional early stopping condition of no improvement in the validation loss for three consecutive epochs. The weights for the loss function $\alpha, \beta, \gamma, \kappa$ in (5.9) were assigned as {1,10,1,10} respectively. These weight values were selected to standardize the units of each individual loss function term. The terms involving SNR and ILD are computed in dB, IPD is calculated in radians and STOI is a bounded score ranging from 0 to 1. The model was trained with the proposed loss function described in (5.9), named BCCRN-SILP, and for comparison, the model was trained to maximize the SNR, named BCCRN-S from (4.8).

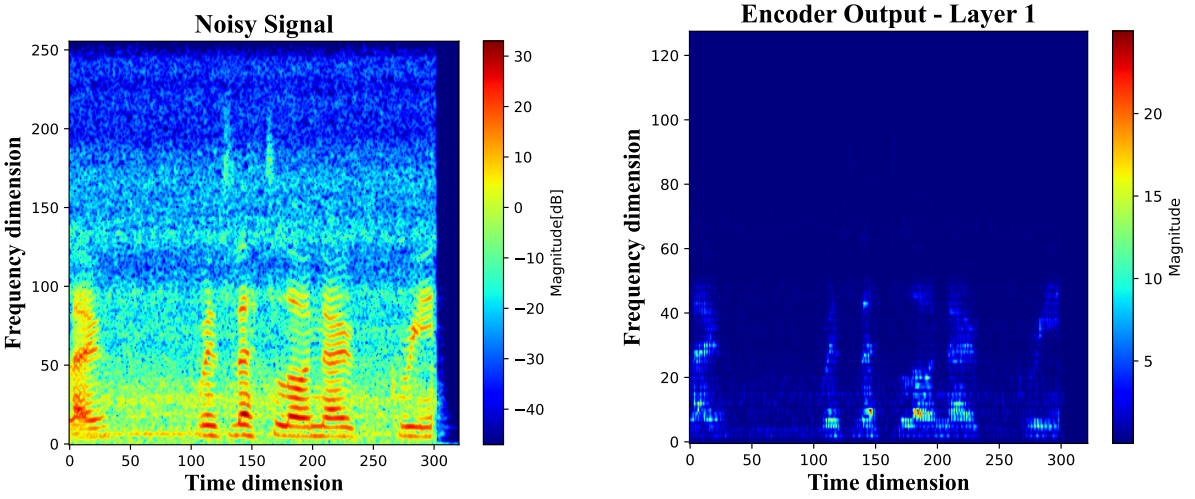
4.4.5 Feature maps

Figures 4.13 - 4.20 show the magnitude spectrum of the noisy signal, feature maps of the encoder, recurrent and decoder layers, and the enhanced signal. These maps are obtained

Block	**Ear Channel**	**Function**	**Dimension**
Input	L/R	STFT	$B \times C_{in} \times F \times T$
Encoder 1	L/R	Conv2D	$B \times 16 \times 128 \times T$
Encoder 2	L/R	Conv2D	$B \times 32 \times 64 \times T$
Encoder 3	L/R	Conv2D	$B \times 64 \times 32 \times T$
Encoder 4	L/R	Conv2D	$B \times 128 \times 16 \times T$
Encoder 5	L/R	Conv2D	$B \times 128 \times 8 \times T$
Encoder 6	L/R	Conv2D	$B \times 128 \times 4 \times T$
RNN Input	L&R	Concat.	$B \times 256 \times 4 \times T$
Flatten	L&R	Flatten	$B \times 1024 \times T$
LSTM	L&R	RNN	$B \times 1024 \times T$
Linear	L&R	Fully-Connected	$B \times 1024 \times T$
Unflatten	L&R	Unflatten	$B \times 256 \times 4 \times T$
Decoder Input	L&R	Transposed Conv2D	$B \times 256 \times 4 \times T$
Decoder Output	L/R	Transposed Conv2D	$B \times C_{in} \times F \times T$

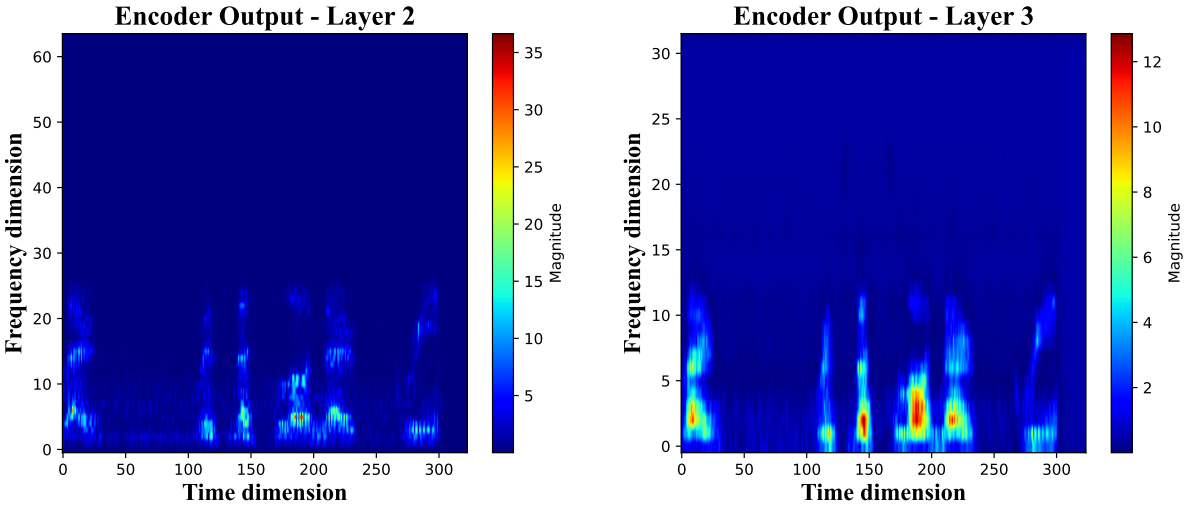
B - Batch size, C_{in} - input channel size of the encoder block, F - number of frequency bins, T - number of time frames, L/R - dimension shown for one channel i.e, left or right ear channel, L&R - dimension shown for both the ear channels combined. Individual decoder blocks have the same dimension as their respective encoder counterparts and skip connections are placed between them

Table 4.2: BCCRN model parameter dimensions



(a) **(b)**

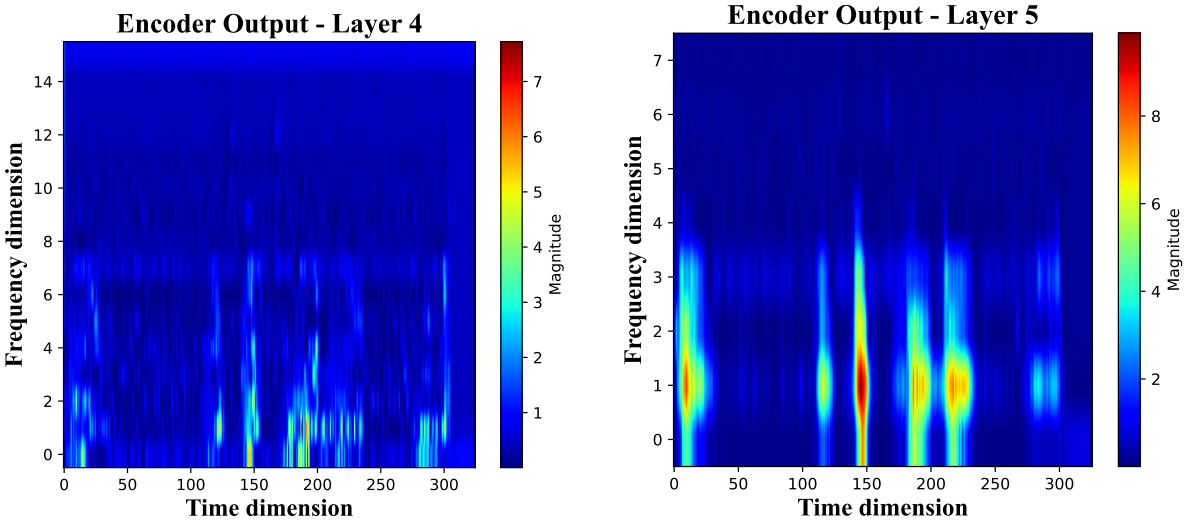
Figure 4.13: (a) Magnitude spectrum of the left channel of a noisy binaural signal, (b) magnitude spectrum of the first encoder layer output of the BCCRN model.



(a) **(b)**

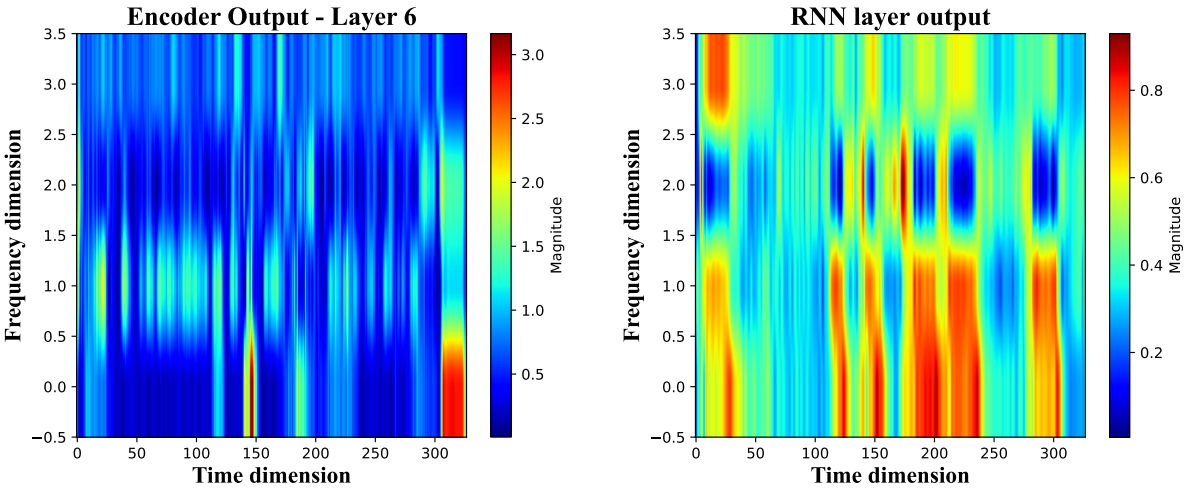
Figure 4.14: (a) Magnitude spectrum of the second and (b) third encoder layer output of the BCCRN model.

as previously described in Sec. 4.4.3. The feature maps for the encoders shown in Figures 4.13(b) - 4.16(a) are similar to the BCCTN model and reduce the frequency dimension while identifying the local patterns in the noisy input signal. Figure 4.16(b) illustrates



(a) **(b)**

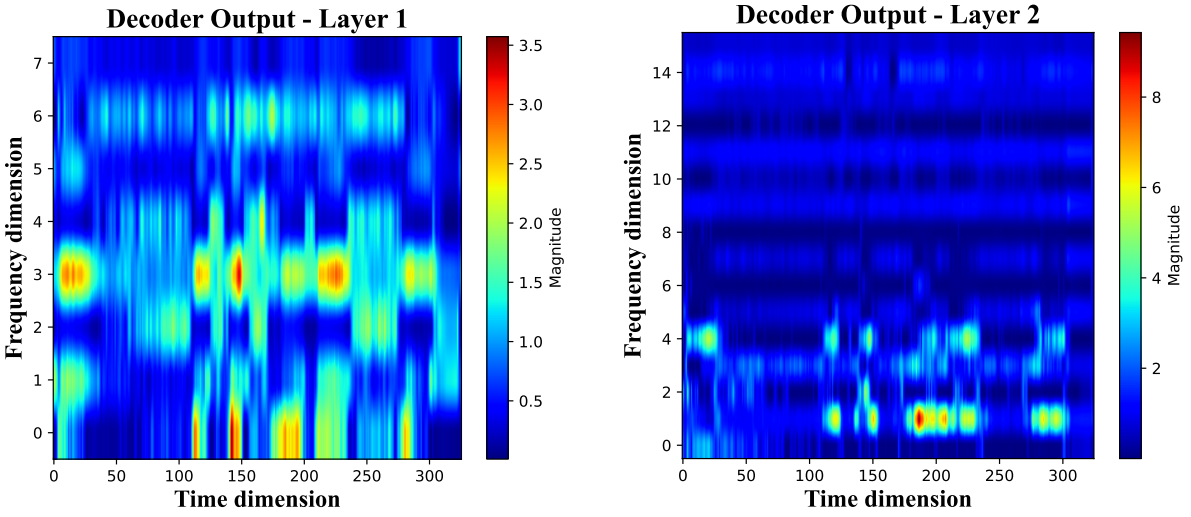
Figure 4.15: (a) Magnitude spectrum of the fourth and (b) fifth encoder layer output of the BCCRN model.



(a) **(b)**

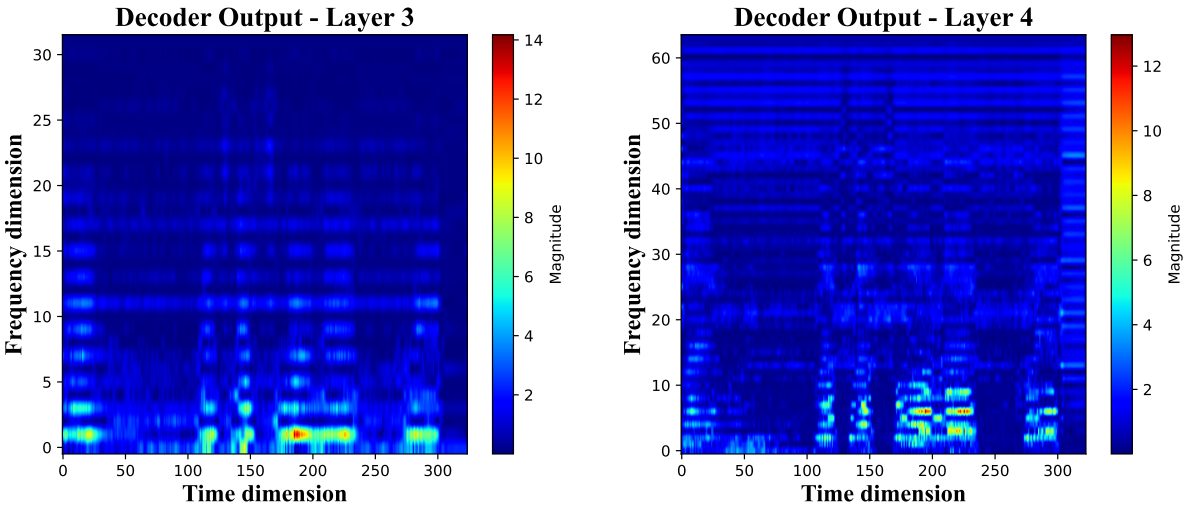
Figure 4.16: (a) Magnitude spectrum of the sixth encoder layer and (b) LSTM layer outputs of the BCCRN model.

the output generated by the LSTM layers within the recurrent block and demonstrates how frequency dependencies are captured. Figure 4.17(a) - 4.19(b) show the magnitude spectrums of the outputs of the decoder layers. Similar to BCCTN model the decoders



(a) **(b)**

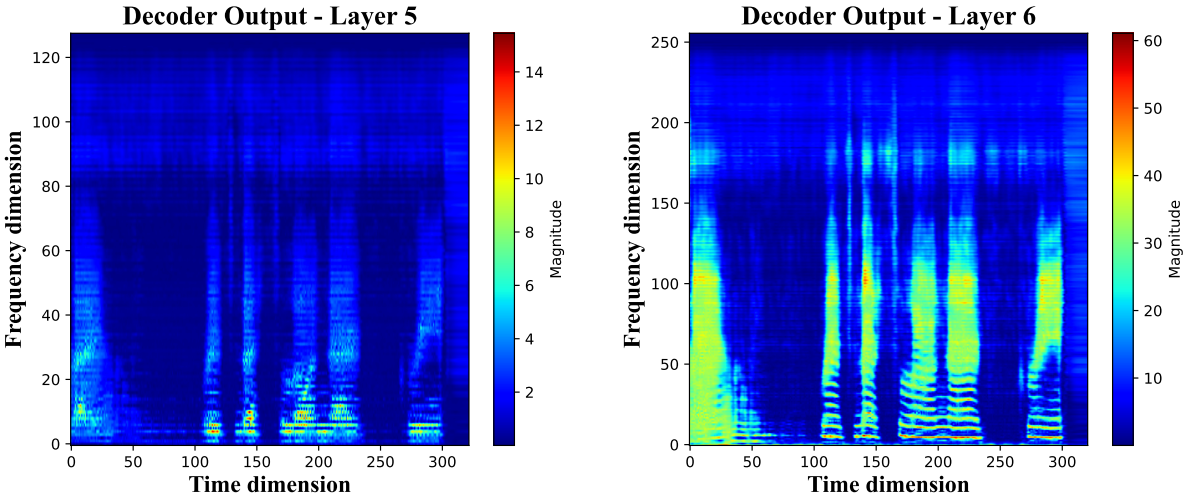
Figure 4.17: (a) Magnitude spectrum of the first and (b) second decoder layer output of the BCCRN model.



(a) **(b)**

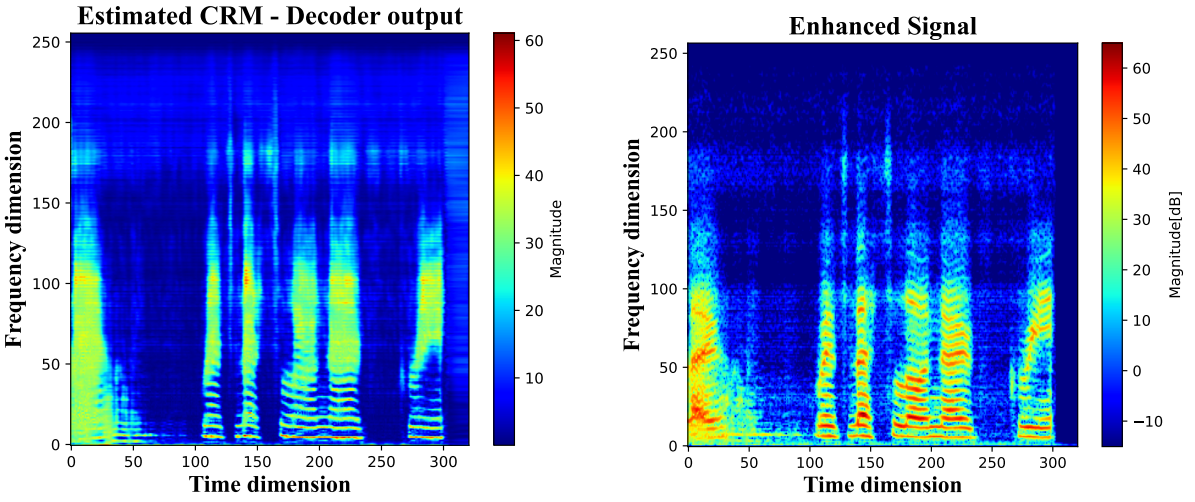
Figure 4.18: (a) Magnitude spectrum of the third and (b) fourth decoder layer output of the BCCRN model.

utilize transposed Conv2D blocks that work in tandem with the encoder blocks, expanding the frequency dimension of the information while decreasing the number of channels. The estimated CRM is illustrated in Fig. 4.20(a), and the enhanced signal resulting from



(a) **(b)**

Figure 4.19: (a) Magnitude spectrum of the fifth and (b) sixth decoder layer output of the BCCRN model.



(a) **(b)**

Figure 4.20: (a) Magnitude spectrum of estimated CRM and (b) enhanced signal of the BCCRN model.

the application of the CRM is depicted in Fig. 4.20(b).

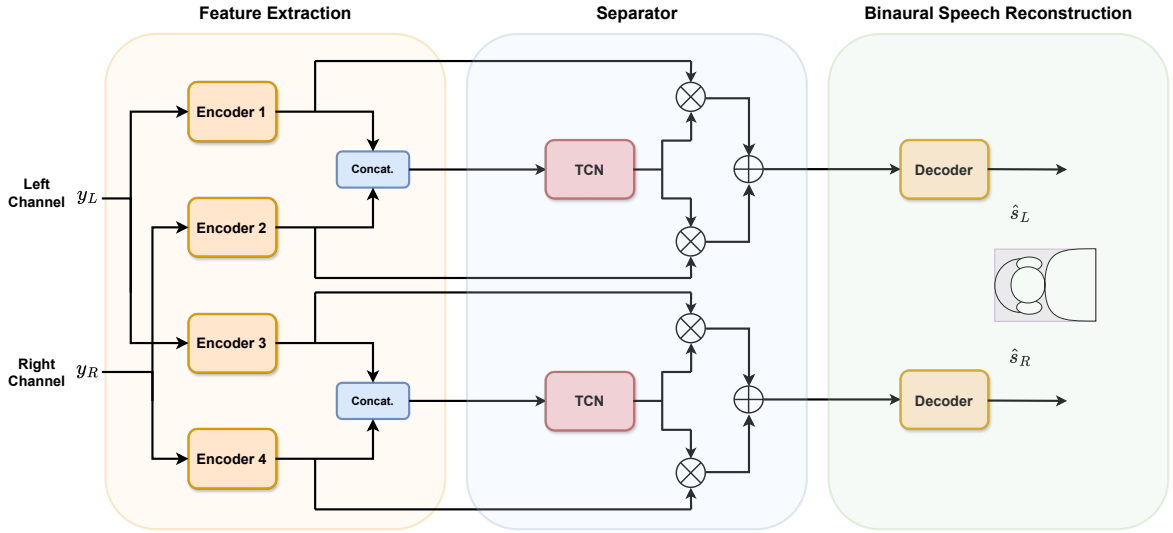


Figure 4.21: Block diagram of the model architecture of BiTasNet (taken from [1]).

4.4.6 Baselines

Binaural STOI-optimal masking (BSOBM)

A binaural speech enhancement method using STOI-optimal masks proposed in Chapter 3 which uses a feed-forward DNN was trained to estimate a STOI-optimal continuous-valued mask to enhance binaural signals using dynamically programmed HSWOBM as the training target [181]. To preserve the ILDs, a better-ear mask was computed by choosing the maximum of the two masks. The mask is used to supply SPP to an OM-LSA enhancer. The model was trained and evaluated on the same dataset as the proposed model.

Binaural TasNet (BiTasNet)

A time-domain CED-based network for binaural speech separation which was introduced in [1]. The block diagram of the BiTasNet is shown in Fig. 4.21 [1]. The BiTasNet is based on the temporal convolutional network (TCN) proposed in [218] for single channel enhancement. The best-performing version of the model, the parallel encoder with mask and sum, was modified and retrained for single-speaker binaural speech enhancement. The network was trained to maximize SNR [1]. The encoders and decoders in the model had a size of 128, a feature dimension of 128, kernel size of 3 and 12 layers. The encoders,

decoders and the separation TCN are majorly made of temporal Conv1D blocks. All other parameters were adapted from the original article [1] and the model has a size of 9.7 million parameters. The model was trained and evaluated on the same dataset used for the proposed method.

4.5 Results

In this section, the results for the BCCTN and BCCRN models are presented for binaural noisy speech in anechoic and reverberant conditions. The performance of the methods were assessed by evaluating 750 utterances from both VCTK [212] and TIMIT [209] datasets for each of the 8 noisy input SNRs. The noise reduction capability of the methods was demonstrated through improvement in fwSegSNR [161]. Objective binaural speech intelligibility of the enhanced signals was measured using the MBSTOI [174] metric. The preservation of interaural cues was evaluated by calculating the error in ILD and IPD after processing, using equations (4.16) and (4.17), respectively.

4.5.1 BCCTN - Results

Figures 4.22(a) and 4.22(b) shows the improvement in fwSegSNR for anechoic and reverberant conditions respectively. For noise reduction measured by the improvement (Δ) in frequency weighted fwSegSNR [161], BiTasNet has the best performance with fwSegSNR for almost all SNRs. However, the proposed method shows comparable performance to BiTasNet on the noise reduction task for both SNR-optimization and the proposed loss function. The proposed loss function had better noise reduction performance compared to the model with the SNR loss function. A possible explanation is that the addition of intelligibility and masked interaural cue terms in the loss function enables the network to better identify the active speech regions which results in better noise reduction performance. A maximum of 14 dB of fwSegSNR can be observed when the signal is very noisy at -6 dB SNR. Figures 4.23(a) and 4.23(b) show the MBSTOI scores for the anechoic and reverberant conditions respectively. Even though the BiTasNet has better noise reduction performance, it exhibits a lower MBSTOI binaural intelligibility score. Informal listening tests revealed that the BiTasNet produced more artefacts. Audio examples of all the methods can be found online [221]. The model provides an average of 0.15 to 0.25

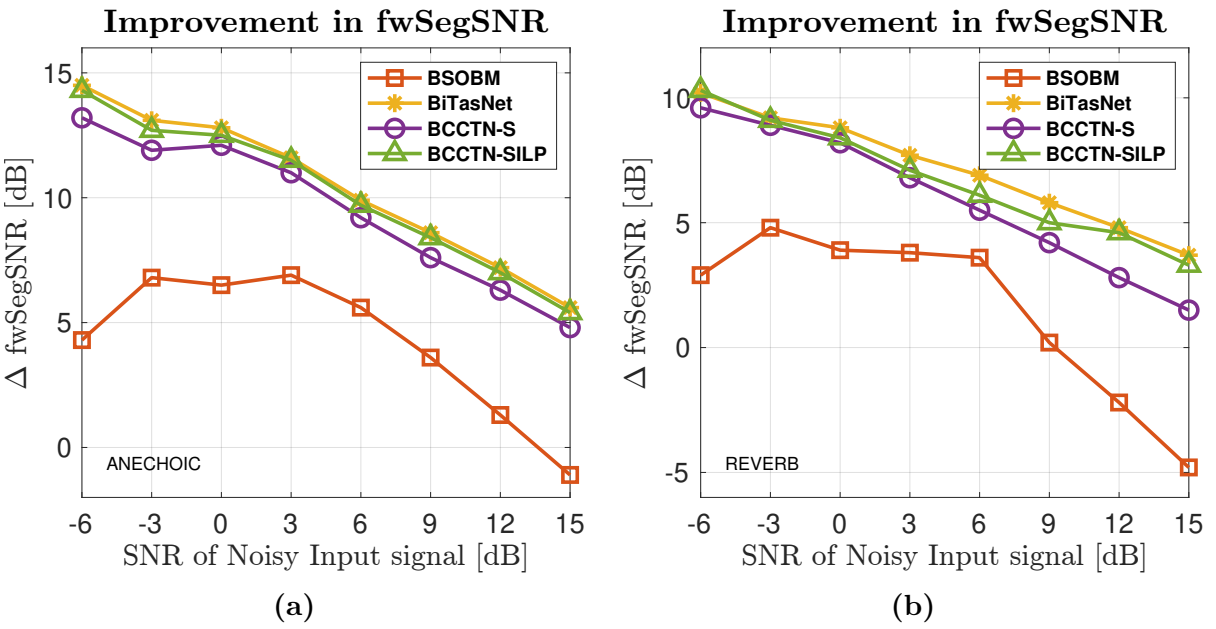


Figure 4.22: Comparison of improvement in channel averaged fwSegSNR for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

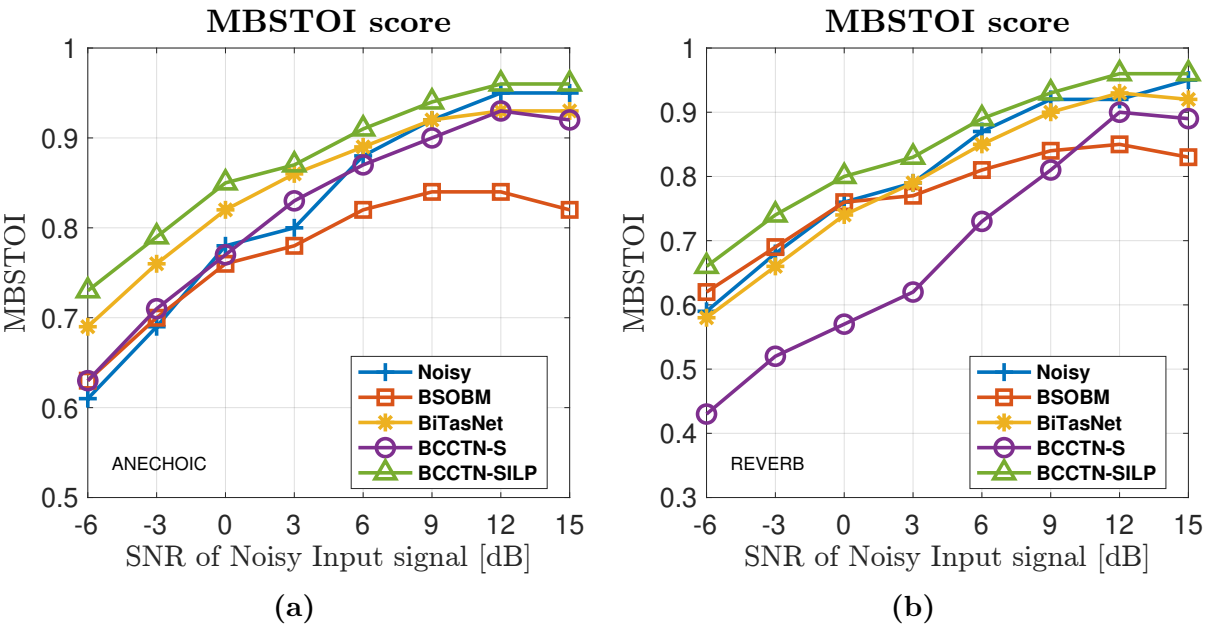


Figure 4.23: Comparison of MBSTOI for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

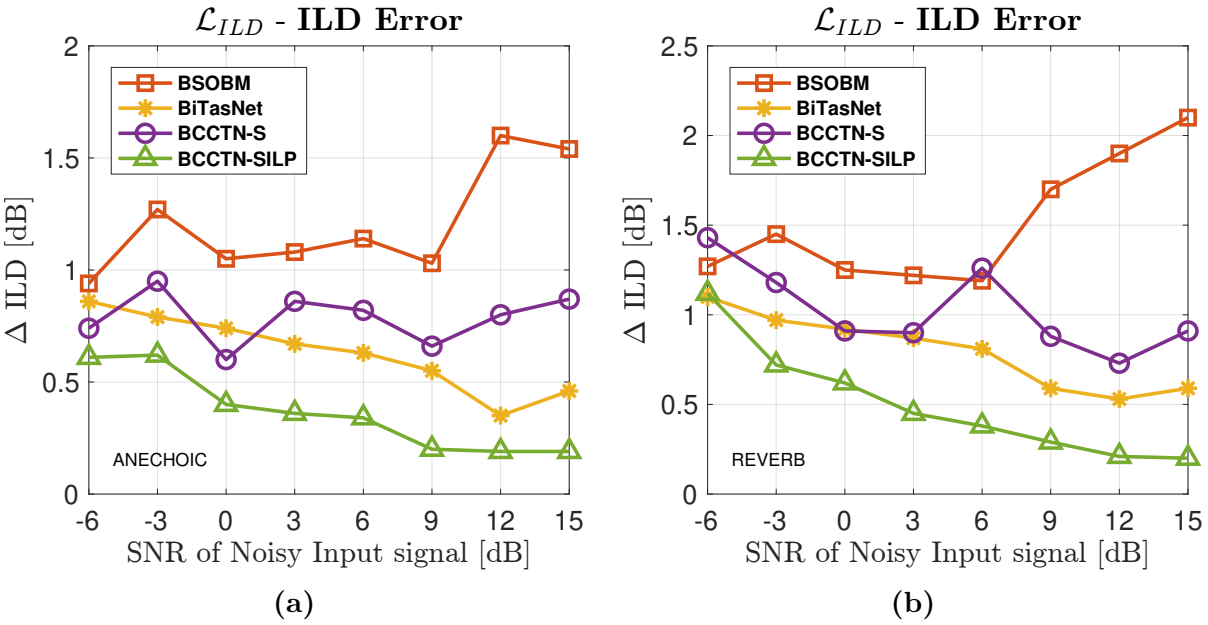


Figure 4.24: Comparison of ILD error (4.16) for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

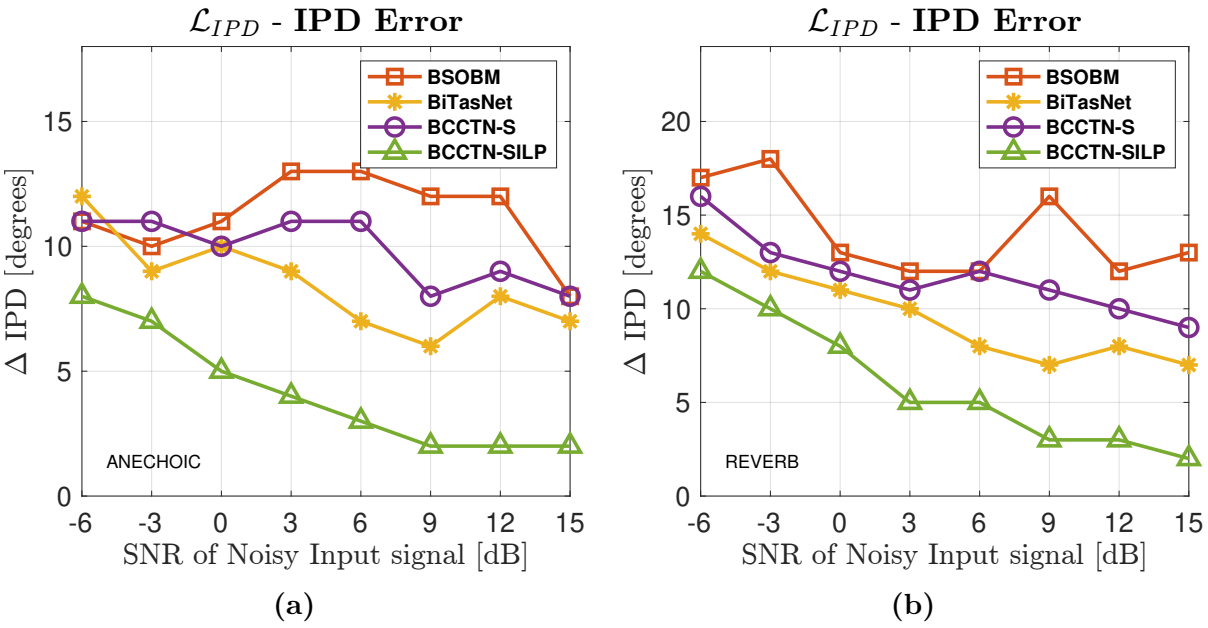


Figure 4.25: Comparison of IPD error (4.17) for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

improvement in MBSTOI scores over the noisy speech when the SNR is below 6 dB. As the input signal's SNR improves, the noisy signals inherently have a higher MBSTOI, and the proposed model provides a lower improvement. In cases with high input SNR, the BSOBM, BiTasNet and BCCTN-SNR methods degrade the MBSTOI score due to processing but the proposed method and loss function do not reduce the score or deteriorate the signal at high SNRs. The proposed model and loss function have the lowest ILD and IPD error for all input SNRs. The proposed model with SNR loss function performs similarly to the proposed loss function in noise reduction but does not focus on retaining the interaural differences and the additional terms in the loss function help the network in the preservation of interaural cues better. Similar performance trends for reverberant signals can be observed from all the methods. A maximum of 10 dB of fwSegSNR can be observed when the signal is very noisy and up to a maximum of 0.15 improvement in MBSTOI score. The ILD and IPD errors are slightly higher than the anechoic condition which could be due to the effects of reverberation [1].

4.5.2 Results - BCCRN

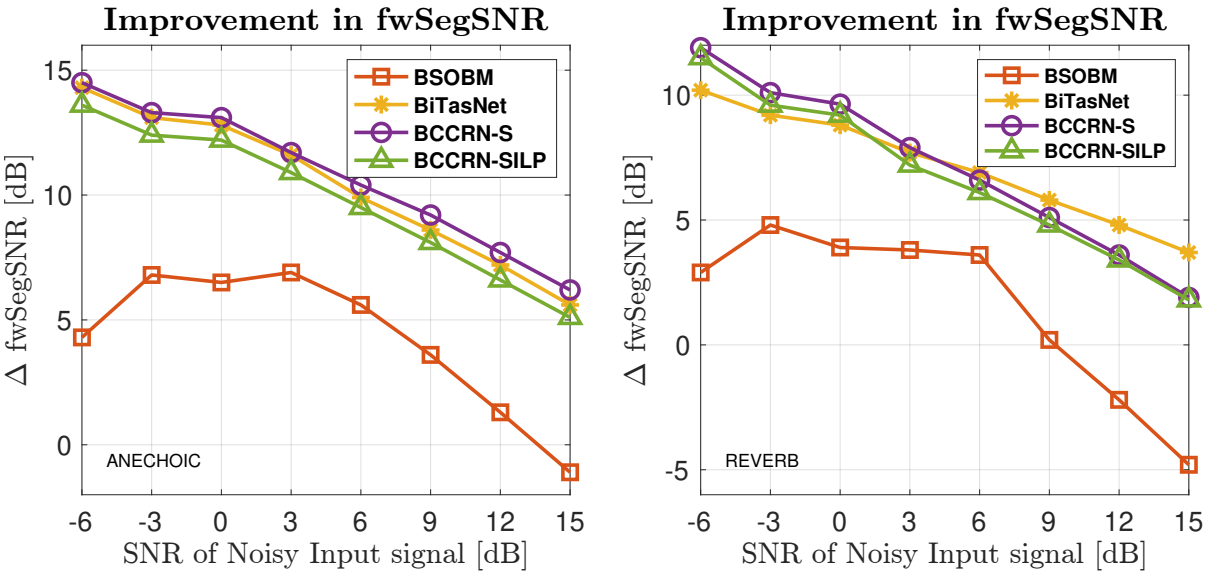


Figure 4.26: Comparison of improvement in channel averaged fwSegSNR for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

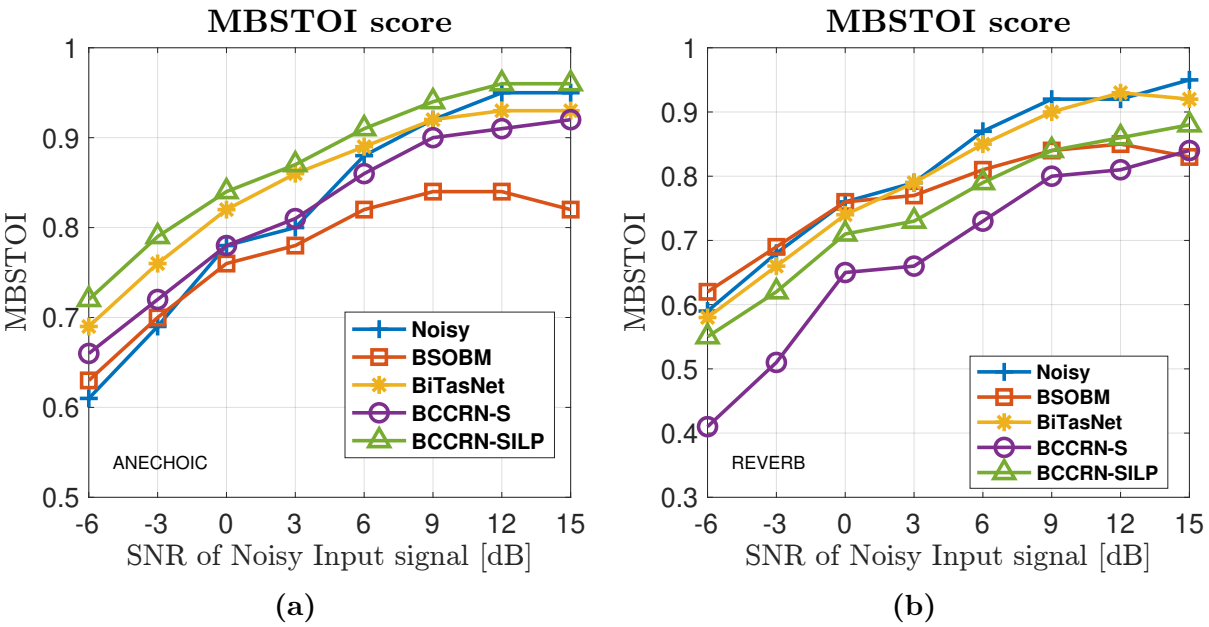


Figure 4.27: Comparison of MBSTOI for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

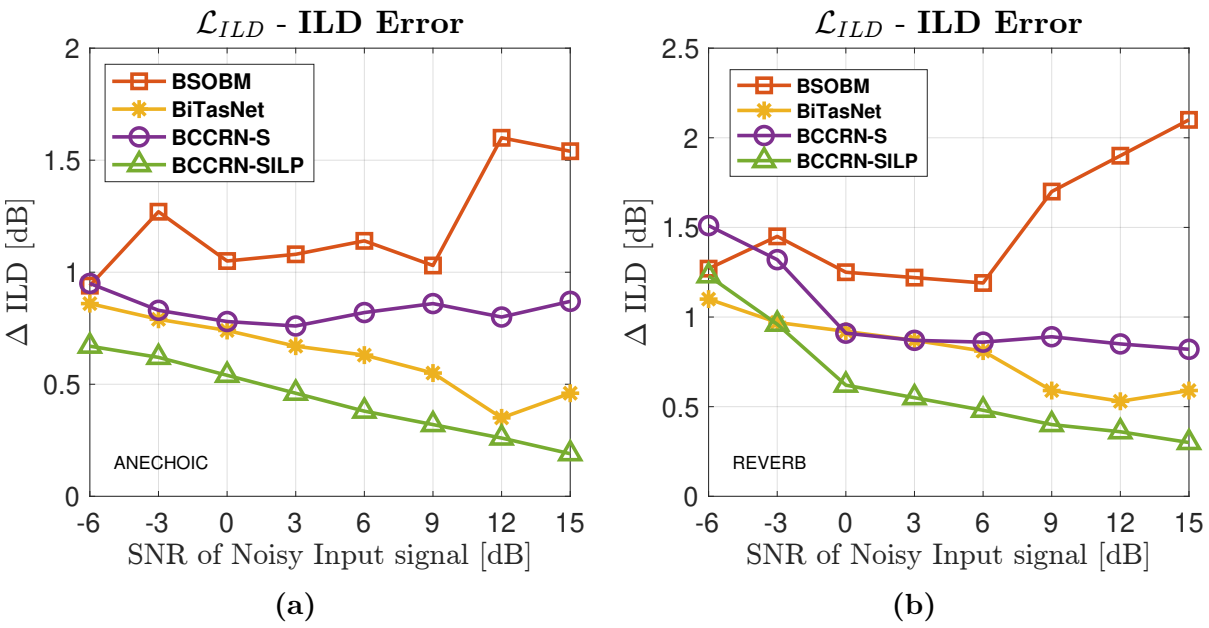


Figure 4.28: Comparison of ILD error (4.16) for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

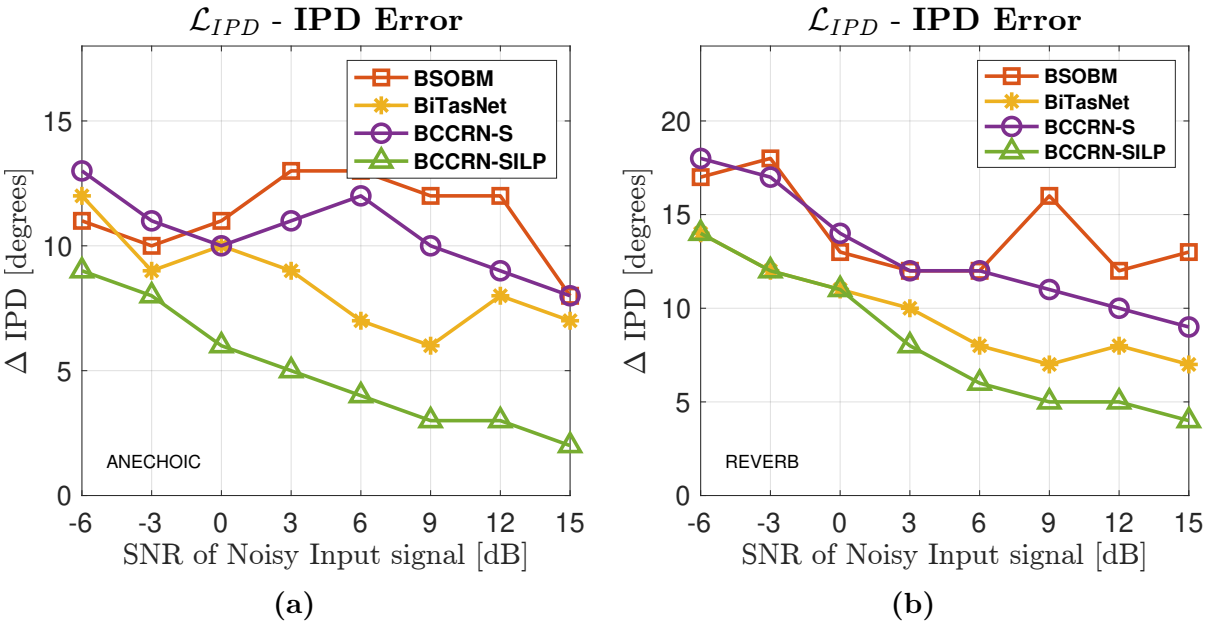


Figure 4.29: Comparison of IPD error (4.17) for noisy binaural speech in (a) anechoic conditions and (b) reverberant conditions.

Figure 4.26(a) and 4.26(b) shows the improvement in fwSegSNR [161] for different input SNRs for anechoic and reverberant conditions. BCCRN-S has the best performance for almost all SNRs, with noise reduction measured by the improvement in the fwSegSNR, while BSOBM has the lowest improvement. Nevertheless, the proposed model exhibits similar effectiveness to the BiTasNet and BCCRN-S in reducing noise. The model has better performance when trained to optimize the SNR compared to the proposed loss function and provides an additional 1 dB of improvement in fwSegSNR on average. A maximum improvement of about 14 dB fwSegSNR is observed in the noisy input SNRs and the amount of improvement observed tends to decrease as the input SNR of the noisy signal improves for both the BCCRN versions. A similar trend can be observed for reverberant conditions and maximum improvement of about 12 dB is observed. At higher SNRs the BiTasNet offers slightly higher improvement compared to the proposed model but has a lower performance noisy input SNRs.

Figures 4.27(a) and 4.27(b) show the MBSTOI of the enhanced signals for anechoic and reverberant conditions respectively. The proposed loss function had the best performance for all SNRs and provided an average of 0.15 to 0.25 improvement in the MBSTOI score. The BCCRN-S has a lower intelligibility score even though it has the best noise

reduction performance. Despite its improved ability to reduce noise, the BiTasNet model demonstrates a lower binaural intelligibility score as measured by MBSTOI shown in Figures 4.27(a) and 4.27(b). In the reverberant scenario, BCCRN-S, much like BCCTN-S, performs poorly. The enhanced MBSTOI is actually lower than that of the noisy signal, despite the model showing a significant improvement in fwSegSNR. The lower score can be attributed to the artefacts generated and the reverberant signals, making it challenging for MBSTOI to evaluate effectively. Informal listening tests revealed that the BiTasNet produced more artefacts and reduced intelligibility of the speech. Similar to fwSegSNR and the error in interaural cues, BSOBM has the lowest MBSTOI score for enhanced signals. A common trend observed in binaural speech enhancement methods is the degradation of the MBSTOI score at high input SNRs due to processing [181] as the input signals inherently have a higher MBSTOI score. However, the proposed loss function does not deteriorate the MBSTOI score, preserves the intelligibility, and provides an improvement at all SNRs.

Figures 4.28(a) and 4.28(a) show the ILD error after enhancement computed using (4.16) and IPD errors computed using (4.17) are shown in Figures 4.29(a) and 4.29(a). The proposed model and loss function have the lowest error for both cues. The suggested model utilizing the SNR loss function demonstrates comparable performance to the proposed loss function in reducing noise, but it does not prioritize preserving the interaural differences. The inclusion of additional terms in the loss function aids the network in better maintaining interaural differences. The observed ILD error was under 1 dB and the IPD error was under 10° for all input SNRs of the noisy signal in the anechoic conditions. For the reverberant conditions, the ILD error was under 1.5 dB and the IPD error was under 15° for all input SNRs. Also, the ILD and IPD errors for the proposed method tend to decrease with increasing input SNR while this is not observed in the model with SNR loss function and other methods. Audio examples of all the methods can be found online [223].

4.6 Comparison between BCCTN and BCCRN

The models are constructed based on the CED architecture yet they differ in the sizes of their encoders and decoders. Specifically, the BCCTN employs encoder and decoder sizes of {16, 32, 64, 128, 256, 256}, whereas the BCCRN utilizes sizes of {16, 32, 64, 128,

128, 128}. BCCRN model features an LSTM block designed to match the size of the transformer block in the BCCTN model. Despite these similarities, BCCTN comprises approximately 10 million parameters, compared to the BCCRNs 6 million parameters. Performance evaluations for binaural speech enhancement indicate that BCCTN consistently outperforms BCCRN across all metrics. Informal listening tests further reveal that BCCRN generates a greater number of artefacts than BCCTN. Analysis of the feature maps and estimated CRM for both models indicates that BCCRN exhibits irregularities in the non-speech regions of the magnitude spectrum, manifesting as horizontal lines or bands.

4.6.1 Complexity

During training and evaluation, both networks utilize the full 2-second spectrogram. This signal length was chosen as a compromise between hardware availability for training and network convergence speed. However, this approach is not optimal for real-world implementation, particularly in hearing devices with limited processing power. To address this limitation, the implementation can be adapted to process shorter signal frames, applying zero-padding as needed based on the hardwares computational capacity. The networks are designed to use only causal information and employ convolutions in the frequency dimension. Both networks achieved comparable performance with shorter frames and zero-padded signals without introducing artifacts. With appropriate processing, these models can be adapted for continuous streaming audio, ensuring that each processing block uses causal 2-second information at a time. While the hardware implementation of neural networks remains a challenging task, it lies beyond the scope of this thesis and is suggested as an area for future research.

4.7 MUSHRA Test

A MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test was conducted with 15 normal hearing participants who were not trained listeners. The participants were asked to listen to the reference signal and rate the test signals on a score of 0 - 100. The input SNRs (iSNRs) of the test signals were chosen randomly and a total of 36 test signals were presented to them for anechoic and reverberant conditions respectively. The

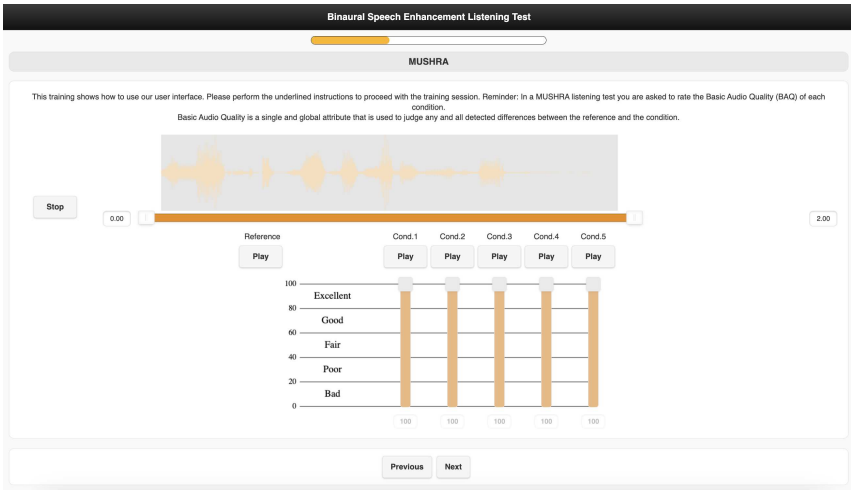


Figure 4.30: MUSHRA test GUI.

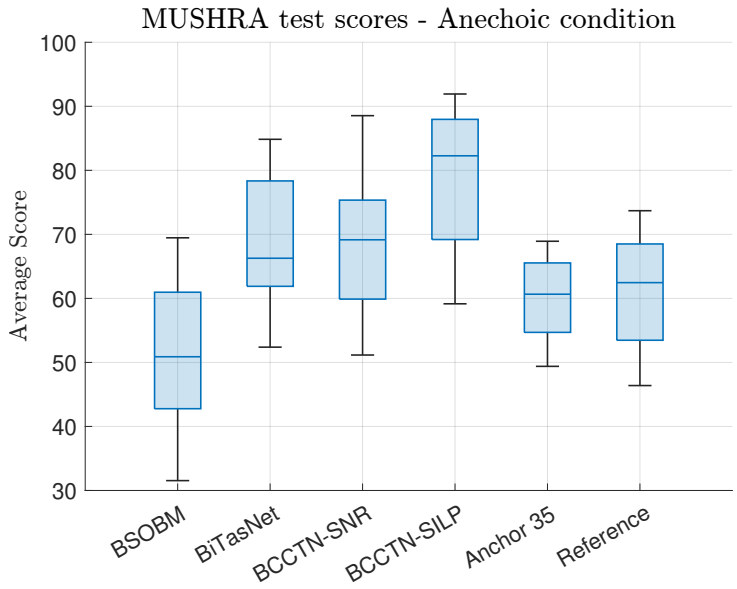


Figure 4.31: MUSHRA test scores for BCCTN compared with the baselines in anechoic conditions (box plot illustrating the median (central line), upper quartile (75%), lower quartile (25%), and the maximum and minimum outliers).

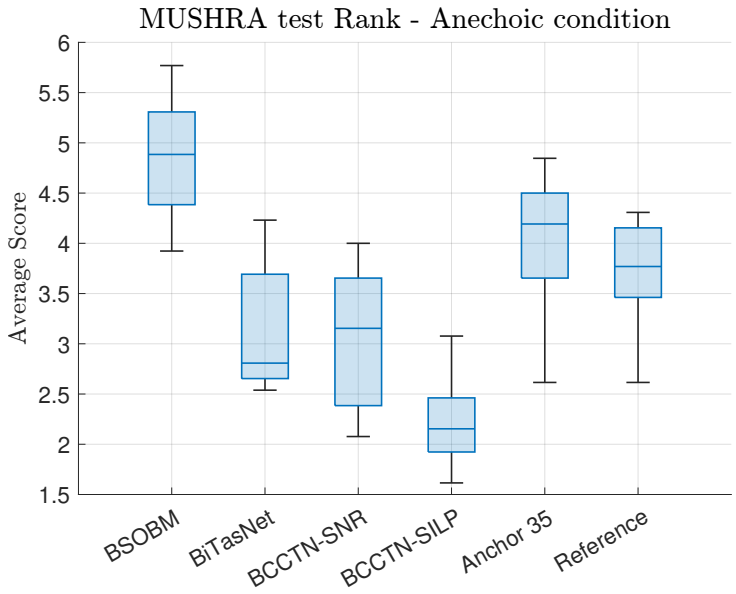


Figure 4.32: MUSHRA test ranks, ranging from 1 (best) to 6 (worst), for BCCTN compared with the baselines over various noisy iSNR levels in anechoic conditions.

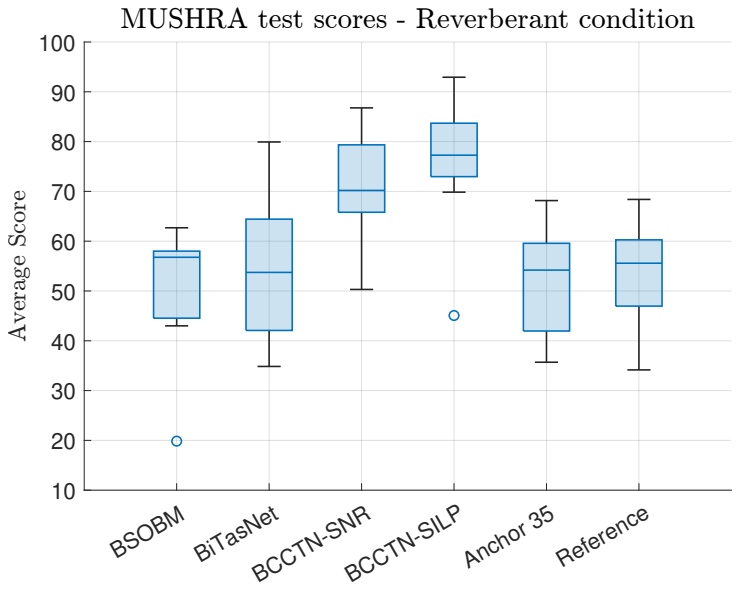


Figure 4.33: MUSHRA test scores for BCCTN compared with the baselines in reverberant conditions.

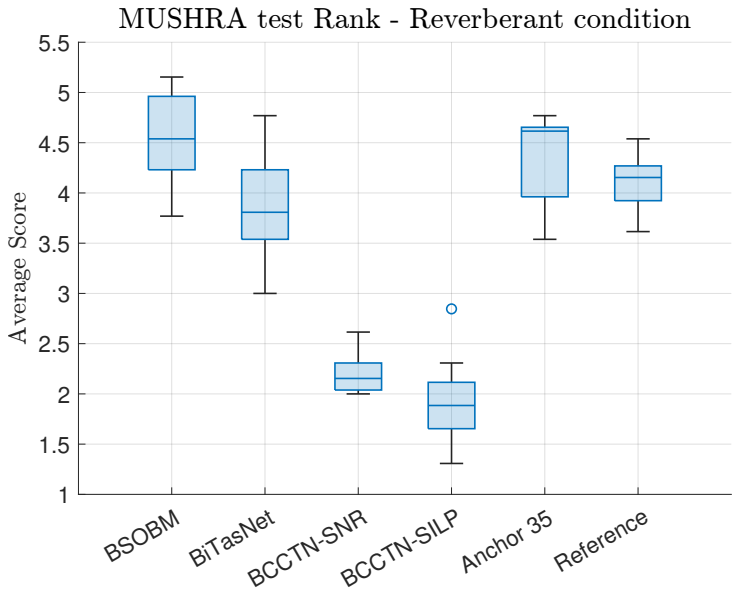


Figure 4.34: MUSHRA test ranks, ranging from 1 (best) to 6 (worst), for BCCTN compared with the baselines over various noisy iSNR in reverberant conditions..

signals were generated using the HRIRs from [211] and [224]. The noisy and reverberant signals were generated as previously described in Sec. 4.4.1. The participants listened to the signals on a Beyerdynamic DT1990 Pro headset and an RME UCX Fireface II was used to play the signals. Figure 4.30 shows the GUI of the web MUSHRA application. Figure 4.31 shows the score averaged over all the iSNRs and Fig. 4.32 shows the average rank obtained by the methods in anechoic conditions. Figure 4.33 shows the score averaged over all the iSNRs and Fig. 4.34 shows the average rank obtained by the methods in reverberant conditions. The participants preferred the BCCTN method with the proposed loss function indicated by BCCTN-SILP in the figures in both anechoic and reverberant conditions. The second most preferred method was the BCCTN method with the SNR loss function. BSOBM scored the least in the test and in some cases lower than the noisy reference and the downsampled Anchor 35 signal. BiTasNet scored better than BSOBM and ranked closer to BCCTN with the SNR loss function.

4.8 Conclusion

This chapter introduced complex-valued convolutional transformers and recurrent networks designed for binaural speech enhancement. A novel loss function was proposed to

optimize the network for noise reduction, speech intelligibility enhancement, and interaural cue preservation. Feature maps in the TF domain were presented which demonstrate the flow and transformation of data in the networks. In experiments, the proposed methods demonstrated significant improvements in MBSTOI and fwSegSNR values while maintaining minimal ILD and IPD errors. The BCCTN model showed promising results and consistently outperformed the BCCRN, highlighting its effectiveness in challenging acoustic environments. Additionally, MUSHRA listening tests supported the quantitative findings, indicating that the BCCTN generated fewer artefacts compared to the BCCRN. These results underscore the potential of the BCCTN model in practical applications. The next chapter will explore a multichannel extension of the proposed model, aiming to further enhance its performance and applicability in more complex audio scenarios.

Chapter 5

Multichannel Binaural Speech Enhancement

Many assistive listening devices such as modern binaural hearing aids or virtual reality headsets are now equipped with multiple microphones, which can be utilized for binaural speech enhancement. This chapter proposes two multichannel binaural speech enhancement methods designed to preserve the spatial information of speech while improving intelligibility and suppressing additive signal components due to a noise field. These comprise of (A) a complex convolutional neural network guided by a beamformer and, (B) an end-to-end convolutional neural network based on a convolutional encoder-decoder architecture. The networks estimate complex ratio masks in the time-frequency domain individually for the left and right channels. Evaluations using simulated data with objective metrics are presented. These show that the proposed methods significantly improve binaural speech intelligibility and reduce noise while preserving the binaural spatial cues of the speech signal in acoustic situations with a single target speaker and isotropic noise of various types. Further evaluations on real-world multichannel recordings of natural group conversations demonstrate that the proposed approaches achieve substantial noise reduction. An earlier version of this work has been submitted for publication [225].

5.1 Introduction

Binaural speech incorporates spatial cues which help listeners in noisy environments achieve better speech intelligibility and accurate sound source localization. Binaural speech enhancement is also used extensively in virtual/augmented reality devices. Binaural speech enhancement is crucial for effective operation of binaural assistive listening devices in adverse listening conditions such as noisy environments [216]. Since most hearing devices are nowadays equipped with multiple microphones at each ear, multichannel binaural speech enhancement algorithms that utilize the data from these multiple microphones are both feasible and essential. Previously, multichannel Wiener filters [34, 226] and beamforming-based approaches [216] have been proposed for multichannel speech enhancement. Multichannel binaural noise reduction techniques using multichannel Wiener filter (MWF) [33, 227] and beamformers [216, 227, 228] have been previously studied. A multichannel speech enhancement method composed of a microphone configuration-specific beamformer and a single channel enhancer was proposed in [229].

Recent advancements in deep learning techniques have led to remarkable improvements in speech enhancement. For optimal phase estimation and signal reconstruction, some recent approaches use complex-valued deep learning methods [89, 219] and use complex-valued spectrograms as the inputs to the network. In [219] the convolutional recurrrent network (CRN) was introduced which is based on the convolutional encoder-decoder (CED) architecture with long short-term memory (LSTM) blocks placed between the encoder and decoder. The complex-valued CRN introduced in [89] was trained to estimate a complex ratio mask (CRM) by optimizing the scale invariant SNR (SI-SNR) for monaural signals. Deep neural network (DNN) based methods typically demonstrate higher performance compared to traditional beamforming methods but typically require microphone-geometry-specific training. Recently, a neural network for monaural speech enhancement based on the beamformer-guided principle was proposed in [230] that is agnostic to microphone configuration. The device-agnostic neural network is said to be 'guided' by the beamformer and takes both a reference unprocessed microphone signal and the beamformer output as it's input. Binaural speech enhancement using a complex-valued CED, as presented in Ch. 4, is designed for two-channel binaural input. In Ch. 4, a CED architecture incorporating a complex-valued attention-based transformer and another CED architecture with LSTM blocks were implemented. These enhancement models

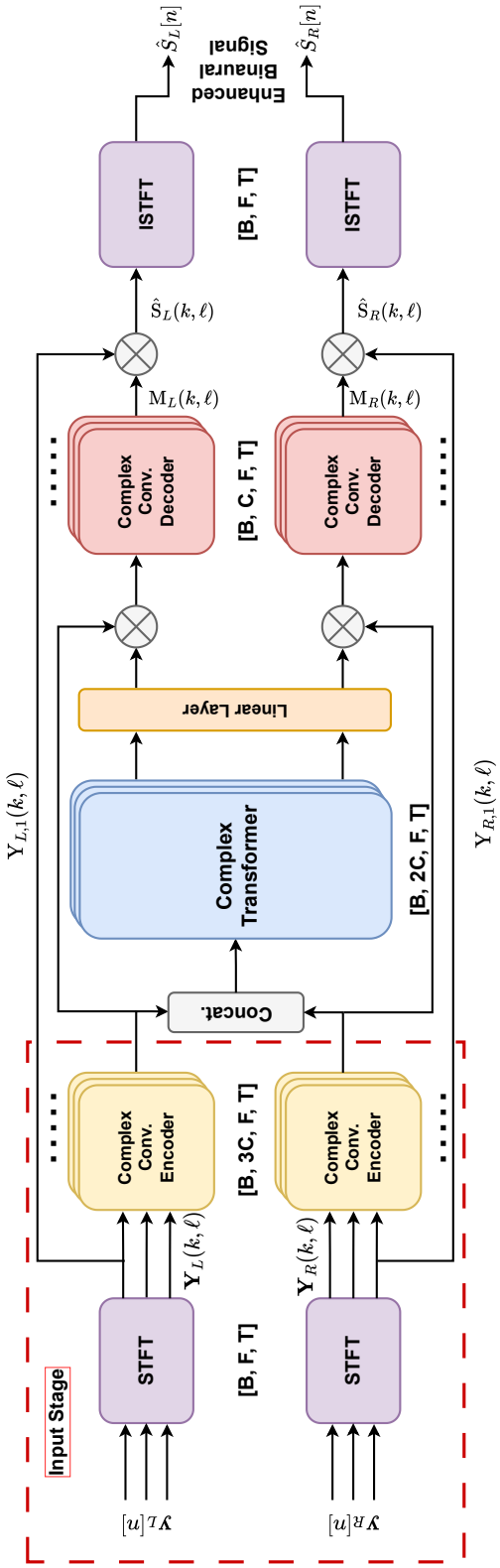


Figure 5.1: Block diagram of the model architecture of the MBCCTN enhancer.

utilize a loss function with terms that optimize the network for noise reduction, intelligibility improvement, and interaural cue preservation.

In this chapter, two multichannel binaural speech enhancers are proposed. The first is an end-to-end neural multichannel binaural speech enhancer based on the CED architecture. The second is a beamformer-guided multichannel enhancer, where the beamformer stage is either a minimum variance distortionless response (MVDR) [216] beamformer or a fixed cylindrical isotropic MVDR (FCIM) [231] beamformer.

5.2 Model Architecture

In Ch. 4, binaural speech enhancement was proposed using complex convolutional networks for 2-channel noisy binaural input. In this chapter, the methods are designed to work with multichannel data from each of the two hearing devices in a binaural setup. This section introduces the two model architectures for the proposed multichannel binaural speech enhancement algorithms: (A) an end-to-end complex-valued neural model based on the CED architecture and (B) a multichannel method utilizing the principle of beamformer-guided speech enhancement [230].

5.2.1 Multichannel binaural complex convolutional transformer network (MBCCTN)

Figure 5.1 shows the block diagram of the model architecture for the proposed end-to-end DNN based method denoted as the multichannel binaural complex convolutional transformer network (MBCCTN). This uses a CED architecture with an attention-based complex-valued transformer block between the encoder and the decoder. The microphone signals from both hearing devices are transformed using short time Fourier transform (STFT) blocks into the time-frequency (TF) domain. The encoder blocks for each ear hearing device input consist of 6 complex convolutional layers with parametric rectified linear unit (PReLU) activation and employ 2D batch normalization. In Fig. 5.1, the size of each block is shown in the form $[B, C, F, T]$ denoting batch size, number of channels, number of frequency bins, and number of time frames respectively. The encoder blocks identify local patterns in the input spectrograms across the encoder channels [89]. Individual encoder blocks are used for the left and right-ear channels of the multichannel input.

Each encoder block is analogous to a beamformer in that it transforms the multichannel input into a single-channel output. The complex transformer block receives the concatenated information from both left and right encoder blocks. It consists of complex-valued multi-head attention layers based on the architecture proposed in [153, 215].

The transformer block focuses on identifying relationships within the encoded information from both channels [90, 153]. The left and right decoder blocks have a similar design to the encoder blocks but differ in the number of input channels. They are designed to reconstruct the single-ear signal from each encoder and consist of 6 transposed complex convolutional blocks. Skip connections are placed between the encoder and decoder layers based on the CED architecture [219]. A separate CRM for each ear channel is output from the decoder and is applied to the spectrogram of the noisy reference channel input signal for the corresponding ear. The inverse STFT (ISTFT) blocks transform the enhanced spectrograms back into the time domain.

5.2.2 Guided OMVDR (GOMVDR) and Guided FCIM (GFCIM)

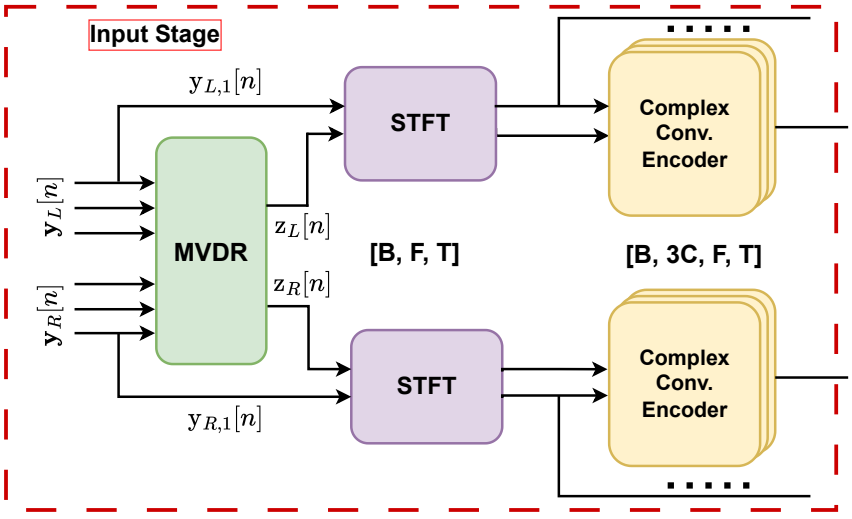


Figure 5.2: Block diagram of the modified input stage for the beamformer guided models, GOMVDR and GFCIM.

For the beamformer-guided enhancers proposed in this chapter, GOMVDR and GFCIM, the input stage of the model described in Sec. 5.2.1 is modified to include the beamformer that is independent of DNN training. Figure 5.2 shows the modified input stage of the

GOMVDR and GFCIM enhancers which replaces the input stage from the previous model shown in Fig. 5.1. The architecture for the beamformer-guided DNN proposed in [230] takes in noisy multichannel data as input and produces monaural enhanced speech as output. In this work, the architecture has been designed to enhance noisy binaural signals with multichannel data. The time-domain multichannel signal is provided as the input to a microphone-array-specific binaural beamformer. For each binaural channel, spectrograms of both the beamformer output and the unprocessed signal from the reference microphone are provided as the input to the encoder. The architecture of the model following the input stage is the same as for the MBCCTN enhancer as shown in Fig. 5.1.

5.3 Signal Model and Algorithm Formulations

Assuming a device that contains M microphones at each ear, the m^{th} microphone signal recorded at the left ear can be expressed in the STFT domain as

$$\mathrm{Y}_{L,m}(k,\ell) = \mathrm{S}_{L,m}(k,\ell) + \mathrm{V}_{L,m}(k,\ell), \tag{5.1}$$

where $\mathrm{S}_{L,m}$ is the speech component, $\mathrm{V}_{L,m}$ is the noise component, and k and ℓ are the frequency and time-indices, respectively. The m^{th} microphone signal at the right ear, $\mathrm{Y}_{R,m}$, is defined similarly. Without loss of generality, $Y_{L,1}$ and $Y_{R,1}$ are taken as the reference microphones. Stacking all microphone signals leads to $\mathbf{Y}(k,\ell) = \begin{bmatrix}\mathbf{Y}_L^T(k,\ell) & \mathbf{Y}_R^T(k,\ell)\end{bmatrix}^T \in \mathbb{C}^{2M\times 1}$, with $(\cdot)^T$ the transpose operator,

$$\mathbf{Y}_L(k,\ell) = \begin{bmatrix}\mathrm{Y}_{L,1}(k,\ell) & \dots & \mathrm{Y}_{L,M}(k,\ell)\end{bmatrix}^T \tag{5.2}$$

and $\mathbf{Y}_R(k,\ell)$, $\mathbf{S}(k,\ell)$, and $\mathbf{V}(k,\ell)$ defined similarly. The beamformer outputs in Fig. 5.2 are given by

$$\mathbf{Z}(k,\ell) \triangleq \begin{bmatrix} \mathrm{Z}_L(k,\ell) \\ \mathrm{Z}_R(k,\ell) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_L^H(k,\ell) \\ \mathbf{W}_R^H(k,\ell) \end{bmatrix} \mathbf{Y}(k,\ell), \tag{5.3}$$

with $(\cdot)^H$ the Hermitian transpose. The beamformer coefficients are $\mathbf{W}_L(k,\ell), \mathbf{W}_R(k,\ell) \in \mathbb{C}^{2M\times 1}$ and the beamformer output signals are $\mathrm{Z}_L(k,\ell)$ and $\mathrm{Z}_R(k,\ell)$ at the left and right ear respectively.

5.3.1 Binaural MVDR beamforming

For the beamformer coefficients in (5.3), a MVDR beamformer [216] is considered, with the left-ear weights provided by

$$\mathbf{W}_L(k,\ell) = \frac{\mathbf{R}_{\mathbf{V}}^{-1}(k,\ell)\mathbf{d}_L(k)}{\mathbf{d}_L^H(k)\mathbf{R}_{\mathbf{V}}^{-1}(k,\ell)\mathbf{d}_L(k)}, \tag{5.4}$$

where $\mathbf{R}_{\mathbf{V}}(k,\ell) \in \mathbb{C}^{2M\times 2M}$ is the Noise Covariance Matrix (NCM), and $\mathbf{d}_L(k) \in \mathbb{C}^{2M\times 1}$ is the relative transfer function (RTF) vector of the speech source through the array, normalised with respect to the left reference microphone. The right-ear weights, $\mathbf{W}_R(k,\ell)$, are defined similarly. The NCM is defined as $\mathbf{R}_{\mathbf{V}}(k,\ell) = \mathbb{E}[\mathbf{V}(k,\ell)\mathbf{V}^H(k,\ell)]$, with $\mathbb{E}[\cdot]$ the expectation operator. Two methods of determining $\mathbf{R}_{\mathbf{V}}(k,\ell)$ are used in this work. First, the Oracle MVDR (OMVDR) beamformer assumes oracle knowledge of the noise-only component and serves to provide an upper bound on the performance of binaural MVDR beamforming. The NCM is estimated as

$$\mathbf{R}_{\mathrm{OMVDR}}(k,\ell) = \alpha\, \mathbf{R}_{\mathrm{OMVDR}}(k,\ell{-}1) + (1-\alpha)\, \mathbf{V}(k,\ell)\mathbf{V}^H(k,\ell), \tag{5.5}$$

where α is a recursive smoothing parameter set to the time constant equivalent of 50 ms. Alternatively, the fixed cylindrical isotropic MVDR (FCIM) assumes a fixed cylindrically

isotropic noise field [231], and $\mathbf{R}_{\text{FCIM}}(k)$ can be obtained from [232] as

$$\mathbf{R}_{\text{FCIM}}(k) = \frac{1}{2\pi} \int_0^{2\pi} \mathbf{a}_\phi(k) \mathbf{a}_\phi^H(k) d\phi, \tag{5.6}$$

where $a_\phi(k)$ is the acoustic transfer function (ATF) from azimuth direction ϕ to the array. In practice, this is discretised to the 5° resolution of the measurements in [211], following [233]. In order to improve the robustness of the beamformer, diagonal loading is applied to the covariance matrices to limit their condition number to ≤ 100 [234]. The left- and right-ear steering vectors, $\mathbf{d}_L(k)$ and $\mathbf{d}_R(k)$, are obtained from [211] and normalised with respect to the front-most microphone at each ear.

5.3.2 Deep complex neural networks

The networks are trained to estimate separate CRMs, $M_L(k, \ell)$ and $M_R(k, \ell)$, for the left and right ear respectively. $M_L(k, \ell)$ is applied to the noisy reference $Y_L(k, \ell)$ to obtain the enhanced speech signal $\hat{S}_L$ for the left-ear signals. The right-ear signals are described similarly. For clarity, the L and R subscripts as well as the k and ℓ indices are omitted where possible for the remainder of this chapter. The enhanced speech is obtained for each ear by applying the estimated complex mask $(M_r + jM_i)$ to the complex-valued noisy signal $(Y_r + jY_i)$ in the TF domain. The enhanced speech STFT coefficients are given by

$$\hat{S}_r + j\hat{S}_i = (M_r + jM_i)(Y_r + jY_i) \tag{5.7}$$

where r and i indicate the real and imaginary parts. Equation 5.7 [63] can be written as

$$M_r + jM_i = \frac{\hat{S}_r + j\hat{S}_i}{Y_r + jY_i} = \frac{Y_r\hat{S}_r + Y_i\hat{S}_i}{Y_r^2 + Y_i^2} + j\frac{Y_r\hat{S}_i - Y_i\hat{S}_r}{Y_r^2 + Y_i^2}. \tag{5.8}$$

5.3.3 Loss function

To train the DNNs, the loss function that was designed for binaural speech enhancement in Sec. 4.3.1 is used. The loss function for model training consists of four terms and optimizes the network for noise reduction, intelligibility improvement, and interaural cue preservation. The loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \alpha\mathcal{L}_{\text{SNR}} + \beta\mathcal{L}_{\text{STOI}} + \gamma\mathcal{L}_{\text{ILD}} + \kappa\mathcal{L}_{\text{IPD}}, \tag{5.9}$$

where $\mathcal{L}_{\text{SNR}}$ is the signal-to-noise ratio (SNR) loss, $\mathcal{L}_{\text{STOI}}$ is the short-time objective intelligibility measure (STOI) [208] loss, and $\mathcal{L}_{\text{ILD}}$ and $\mathcal{L}_{\text{IPD}}$ are the proposed interaural level difference (ILD) and interaural phase differences (IPD) error losses with respect to the target speech signal and are functions of both $\hat{S}_L$ and $\hat{S}_R$. The parameters α, β, γ, and κ are weights chosen to approximately normalize the scale of each term in the loss function.

5.4 Experiments

5.4.1 Datasets

To generate the multichannel signals, speech signals from the CSTR VCTK corpus [212] were convolved with measured head related impulse responses (HRIRs) from [211]. These HRIRs were measured for two behind the ear (BTE) hearing aids, each containing $M = 3$ microphones mounted on a head and torso simulator (HATS). The speech corpus contains approximately 13 hours of spoken data from 110 English speakers with different accents. This data was used to create 2-second speech segments that were spatialized for the left and right-ear channels. The speech source was positioned at a random azimuth in the frontal horizontal plane ($-90°$ to $+90°$). The position of the speech source was set at $0°$ elevation and a random azimuth in the range $\pm 90°$, and at a distance of either 80 cm or 300 cm from the HATS chosen randomly. The dataset comprised 20,000 speech segments, which were divided into training (70%), validation (15%), and evaluation (15%) sub-

Table 5.1: Model Parameter dimensions

Block	**Ear Channel**	**Function**	**Dimension**
Input	L/R	STFT	$B \times C_{in} \times F \times T$
Encoder 1	L/R	Conv2D	$B \times 8 \times 128 \times T$
Encoder 2	L/R	Conv2D	$B \times 16 \times 64 \times T$
Encoder 3	L/R	Conv2D	$B \times 32 \times 32 \times T$
Encoder 4	L/R	Conv2D	$B \times 64 \times 16 \times T$
Encoder 5	L/R	Conv2D	$B \times 128 \times 8 \times T$
Encoder 6	L/R	Conv2D	$B \times 128 \times 4 \times T$
Attention Input	L&R	Concat.	$B \times 256 \times 4 \times T$
Flatten	L&R	Flatten	$B \times 1024 \times T$
Transformer	L&R	Attention	$B \times 1024 \times T$
Linear	L&R	Fully-Connected	$B \times 1024 \times T$
Unflatten	L&R	Unflatten	$B \times 256 \times 4 \times T$
Decoder Input	L&R	Transpose Conv2D	$B \times 256 \times 4 \times T$
Decoder Output	L/R	Transpose Conv2D	$B \times C_{in} \times F \times T$

B - Batch size, C_{in} - input channel size of the encoder block, F - number of frequency bins, T - number of time frames, L/R - dimension shown for one channel i.e, left or right ear channel, L&R - dimension shown for both the ear channels combined. Individual decoder blocks have the same dimension as their respective encoder counterparts and skip connections are placed between them.

sets. Noise signals from the NOISEX-92 [210] dataset were utilized to create cylindrically isotropic noise by distributing uncorrelated noise sources with the same power evenly every 5° in the horizontal plane [186], and convolving them with HRIRs from [211]. During training, isotropic noise was added to the anechoic binaural speech components such that $(SNR_L + SNR_R)/2$ fell between -6 dB and 16 dB. The noise types employed for training included white Gaussian noise, speech-shaped noise, factory background noise, and office noise. The evaluation dataset comprised unseen speech segments at a random azimuth between $-90°$ to $+90°$ and added isotropic noise. During evaluation along with the noise types used in training, car engine noise was additionally included. Although the models were trained using only anechoic binaural speech, reverberant speech segments were also considered for evaluation and were created using binaural room impulse responses (BRIRs) from [211] with additive isotropic noise. Various rooms with T_{60} values in the range between 0.3 to 1 s were employed for this purpose. A sampling rate of 16 kHz was used for all signals.

5.4.2 Training setup and baselines

End-to-end MBCCTN model

For the STFT computation, an FFT length of 512 (32 ms), a window length of 25 ms, and a hop length of 6.25 ms were used. Table 5.1 shows the model parameter dimensions for the network. The number of inputs used in the convolutional layers for the encoder and decoder blocks is $16, 32, 64, 128, 256, 256$, with a stride of 2 in the frequency dimension, 1 in the time dimension, and a kernel size of (5,1). The number of input channels is equal to 3 for the encoder and 1 for the decoder. All convolutions in these layers are causal. The multi-head attention block in the complex transformer has an embedded dimension of 1024, with a hidden size of 128 and 32 heads. The linear layer placed after the complex transformer block has an input and output feature size of 1024. The Pytorch model was trained using the Adam optimizer with an initial learning rate of 0.001, and a multi-step learning rate scheduler. The model has a size of around 10 million parameters. The model was trained for 100 epochs and terminated early if no improvement was observed for 3 consecutive epochs. The loss functions weights $\alpha, \beta, \gamma, \kappa$, in (5.9), were set to $\{1, 10, 1, 10\}$ respectively to normalize the differences in the scale of the respective units of the individual loss function terms [214,215]. The front microphones of the hearing

aids are used as the reference signals for loss computation and backpropagation and the MBCCTN model does not use any clean speech oracle information during evaluation.

Beamformer-guided models

The guided versions of the model have the same model parameters as the end-to-end model with the number of input channels to the encoder reduced to 2 by an OMVDR or an FCIM beamformer, respectively. The front microphones of the hearing aids were used as the reference channels for the beamformers and the loss function computation. The neural net was then trained with the same loss function as the end-to-end neural model. As GOMVDR and GFCIM have access to oracle information (respectively, both noise statistics and steering direction, and steering direction only), they can be treated as the upper limit of beamforming performance.

5.4.3 Reducing model size

Mini-versions of the end-to-end and the guided models were also considered where the number of channels in the encoder and decoder block layers is $\{8, 16, 32, 64, 64, 64\}$. The multi-head attention layers in the transformer block have an embedded dimension of 256, with a hidden size of 32 and 8 heads. The size-reduced models have around 3 million parameters. The models were trained for 100 epochs to obtain a similar loss function optimization as the larger models.

5.4.4 Mismatched steering vectors and SRP-based beamformers

To study the performance of the beamformer-guided methods for realistic scenarios involving imperfect direction of arrival (DOA) estimation algorithms, during evaluation, the beamformers were supplied with incorrect steering vectors that had an error introduced of $5°, 10°$ or $20°$ offset from the true target direction. The outputs from the beamformers with the offset steering vectors along with the reference channels were supplied to the trained neural net for enhancement.

Instead of using an oracle source direction, the steered response power (SRP)-based beamformer-guided models (GMVDR+SRP and GFCIM+SRP), the left and right ear channels of the binaural clean speech signals, S_L and S_R, are considered. The azimuth

θ of the target speaker's DOA in the frontal azimuth plane is then estimated using the SRP with phase transform (SRP-PHAT) algorithm [235, 236].

5.5 Results

The models presented in Sec. 5.4 were evaluated using 400 speech utterances at each input SNR (iSNR) ranging from -6 dB to 15 dB in steps of 3 dB. The speech enhancement performance was evaluated using frequency-weighted segmental SNR (fwSegSNR) [161]. The modified binaural STOI (MBSTOI) [174] score was computed to evaluate the binaural speech intelligibility. Table 5.2 shows the MBSTOI score and the improvement in fwSegSNR of all the models in anechoic conditions for all noisy iSNRs and Table 5.3 shows the performance in reverberant conditions. For the beamformer-guided methods along with the size-reduced versions, the beamformers with SRP-based steering vector estimation are shown. Figure 5.3 shows the fwSegSNR performance for the beamformer-guided models with an offset in the steering vector.

From Table 5.2 for noisy iSNRs under 6 dB, all proposed DNN-based methods outperform the OMVDR beamformer in fwSegSNR and provide a maximum improvement of up to 13 dB. The GOMVDR provides better noise reduction for iSNRs below 0 dB and the MBCCTN provides the maximum improvement for iSNRs above 0 dB. The beamformer stage of GOMVDR has access to clean speech while MBCCTN does not and this helps the GOMVDR's beamformer stage to provide a signal that has an improved SNR as the input to the DNN stage. In anechoic conditions, all the enhancers improved MBSTOI with OMVDR giving the highest increase. OMVDR with access to clean speech statistics and the target speaker direction was able to provide the maximum improvement at all iSNRs of up to 0.2. The GOMVDR and MBCCTN models were able to improve the MBSTOI by up to 0.19 but outperformed the OMVDR in noise reduction. In reverberant conditions, the MBCCTN model provides the best performance in MBSTOI and fwSegSNR improvement for almost all iSNRs with an improvement of up to 0.26 for MBSTOI and up to 12 dB of improvement in fwSegSNR as shown in Table 5.3. Compared to the GOMVDR, the GFCIM has a lower performance given that it assumes a cylindrically isotropic noise field. Considering the size-reduced mini models, trends are similar to the large models. However, up to 2-3 dB reduction in terms of fwSegSNR improvement and up to 0.2 reduction in terms of MBSTOI was observed for a reduction of around 65% in

Table 5.2: The table shows the average performance for anechoic conditions in terms of (a) MBSTOI and (b) Δ fwSegSNR relative to the noisy speech.

Input SNR	**-6 dB**		**-3 dB**		**0 dB**		**3 dB**	
Method	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**
Noisy signal	0.60	-	0.67	-	0.78	-	0.81	-
OMVDR	**0.80**	8.4	**0.85**	8.5	**0.91**	8.4	**0.92**	8.1
GOMVDR	0.79	**13.1**	0.82	**11.9**	0.87	11.6	0.89	9.9
GFCIM	0.73	12	0.79	10.8	0.85	10.9	0.87	9.2
GOMVDR mini	0.79	12.5	0.84	10.9	0.88	10.7	0.89	8.7
GFCIM mini	0.73	12.4	0.79	11.0	0.86	10.8	0.87	9.2
GMVDR + SRP	0.78	12.9	0.82	11.7	0.87	11.5	0.89	10.1
GFCIM + SRP	0.75	12.4	0.81	11.2	0.87	11.4	0.89	9.5
GMVDR mini + SRP	0.76	12.9	0.81	11.5	0.86	11.8	0.88	9.7
GFCIM mini + SRP	0.73	12.1	0.79	10.7	0.85	10.9	0.86	9.2
MBCCTN	0.76	**13.1**	0.81	11.7	0.87	**11.9**	0.89	**10.2**
MBCCTN mini	12.1	0.79	11.5	0.86	11.8	0.88	10.1	0.91

Input SNR	**6 dB**		**9 dB**		**12 dB**		**15 dB**	
Method	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**
Noisy signal	0.88	-	0.91	-	0.94	-	0.95	-
OMVDR	**0.95**	7.9	**0.96**	5.3	**0.97**	4.8	**0.97**	3.6
GOMVDR	0.92	8.5	0.94	6.4	0.96	6.1	0.97	4.5
GFCIM	0.91	8.1	0.93	6.2	0.95	5.6	0.96	3.5
GOMVDR mini	0.92	7.5	0.93	5.7	0.95	5.6	0.96	3.7
GFCIM mini	0.91	8.1	0.93	6.3	0.95	5.6	0.96	3.8
GMVDR + SRP	0.92	8.6	0.94	6.7	0.95	6.0	0.97	4.1
GFCIM + SRP	0.92	8.4	0.93	6.4	0.95	5.7	0.96	3.8
GMVDR mini + SRP	0.92	8.5	0.93	6.7	0.95	5.6	0.96	3.7
GFCIM mini + SRP	0.91	8.2	0.93	6.3	0.95	5.5	0.96	3.7
MBCCTN	0.93	**9.1**	0.94	**6.9**	0.97	**6.6**	0.97	**4.6**
MBCCTN mini	0.91	8.7	0.95	6.7	0.95	5.1	0.97	4.3

Table 5.3: The table shows the average performance for reverberant conditions in terms of (a) MBSTOI and (b) Δ fwSegSNR relative to the noisy speech.

Input SNR	**-6 dB**		**-3 dB**		**0 dB**		**3 dB**	
Method	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**
Noisy signal	0.34	-	0.40	-	0.47	-	0.52	-
OMVDR	0.60	6.6	0.66	5.7	0.67	4.7	0.68	3.1
GOMVDR	**0.63**	11.1	**0.67**	9.1	**0.70**	8.7	**0.71**	6.3
GFCIM	0.58	10.1	0.64	8.5	0.67	7.8	0.69	5.8
GOMVDR mini	0.62	10.7	0.66	8.9	0.69	8.5	0.70	6.1
GFCIM mini	0.57	9.8	0.61	8.3	0.66	7.5	0.67	5.6
GMVDR + SRP	0.61	10.6	0.65	8.9	0.68	8.5	0.70	6.2
GFCIM + SRP	0.58	10.5	0.64	8.8	0.68	8.3	0.69	6.2
GMVDR mini + SRP	0.60	10.4	0.64	8.6	0.67	8.4	0.69	6.2
GFCIM mini + SRP	0.57	10.0	0.62	8.4	0.66	7.8	0.67	5.9
MBCCTN	0.60	**12.3**	0.64	**10.7**	0.68	**9.9**	0.7	**8.7**
MBCCTN mini	0.58	11.7	0.63	10.6	0.66	9.7	0.68	8.2
Input SNR	**6 dB**		**9 dB**		**12 dB**		**15 dB**	
Method	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**	**MBSTOI**	**ΔfwSegSNR**
Noisy signal	0.58	-	0.66	-	0.67	-	0.69	-
OMVDR	0.71	1.5	0.75	-0.3	0.71	-2.2	0.72	-4.4
GOMVDR	0.74	5.2	**0.75**	4.1	**0.75**	3.0	**0.76**	2.1
GFCIM	0.73	4.7	0.74	3.4	0.75	2.4	0.76	1.6
GOMVDR mini	0.73	5.1	0.74	3.9	0.74	2.9	0.75	1.7
GFCIM mini	0.70	4.6	0.72	3.4	0.73	2.4	0.74	1.5
GMVDR + SRP	0.73	5.3	0.74	4.2	0.75	3.0	0.76	2.1
GFCIM + SRP	0.71	5.1	0.74	3.9	0.75	2.8	0.75	1.9
GMVDR mini + SRP	0.70	5.0	0.73	3.9	0.74	2.8	0.74	2.1
GFCIM mini + SRP	0.70	4.8	0.72	3.7	0.73	2.8	0.75	1.8
MBCCTN	**0.71**	**7.4**	0.73	**5.4**	**0.75**	**4.4**	0.74	**3.2**
MBCCTN mini	0.70	6.8	0.72	4.5	0.73	4.1	0.74	1.7

Table 5.4: Interaural Signal Phase Difference (ISPD) error in degrees for clean speech signals processed by the methods (model size is indicated in millions (M) of parameters.

ISPD Error (10M param)		**ISPD Error (3M param)**	
OMVDR	6°	-	-
GOMVDR	3°	GOMVDR mini	4°
GFCIM	3°	GFCIM mini	4°
GMVDR+ SRP	4°	GMVDR mini + SRP	6°
GFCIM + SRP	5°	GFCIM mini + SRP	6°
MBCCTN	2°	MBCCTN mini	3°

Table 5.5: Maximum ILD error in dB of the target speaker computed using (4.16) for the noisy speech signals processed by the models (model size is indicated in millions (M) of parameters.

ILD Error (dB) (10M param)		**ILD Error (dB) (3M param)**	
OMVDR	1.7	-	-
GOMVDR	0.5	GOMVDR mini	1.0
GFCIM	0.8	GFCIM mini	1.2
GMVDR+ SRP	0.6	GMVDR mini + SRP	0.9
GFCIM + SRP	0.8	GFCIM mini + SRP	0.9
MBCCTN	0.8	MBCCTN mini	1.1

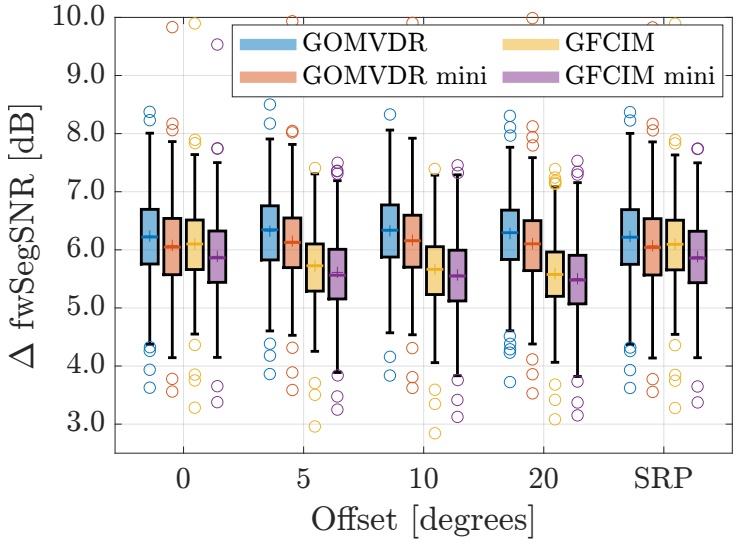


Figure 5.3: Improvement in frequency-weighted segmental SNR (fwSegSNR) for the beamformer-guided methods with mismatched steering vectors.

the model size.

For the beamformer-guided models with SRP, the target speaker direction is estimated using the SRP algorithm described in Section 5.4.4. In both anechoic and reverberant conditions both GMVDR+SRP and GFCIM+SRP perform similarly to the models with full access to clean speech oracle information. A similar observation can be made with the size-reduced mini models. Figure 5.3 presents the enhancement in fwSegSNR, while Fig. 5.4 illustrates the MBSTOI scores achieved when there is a deviation in the steering vector employed during the beamforming stage. The values depicted represent the averages across all iSNR levels, considering offsets of 5°, 10°, and 20°. For comparison purposes, the performance of the SRP based beamformer is evaluated alongside a beamformer utilizing the oracle clean speech information. Figures 5.3 and 5.4 demonstrate that mismatched steering vectors have a negligible impact on the enhancement performance of the beamformer-guided models, with a maximum reduction of only 0.5 dB in fwSegSNR improvement and a decrease of 0.05 in MBSTOI scores for a 20° offset. In Sec. 5.6, the proposed models will be further assessed using real data without access to oracle information.

Table 5.4 shows the Interaural Signal Phase Difference (ISPD) calculated by evaluating the models on clean speech signals and measuring the errors in IPD before and after processing. It can be observed that for all proposed DNN-based binaural speech

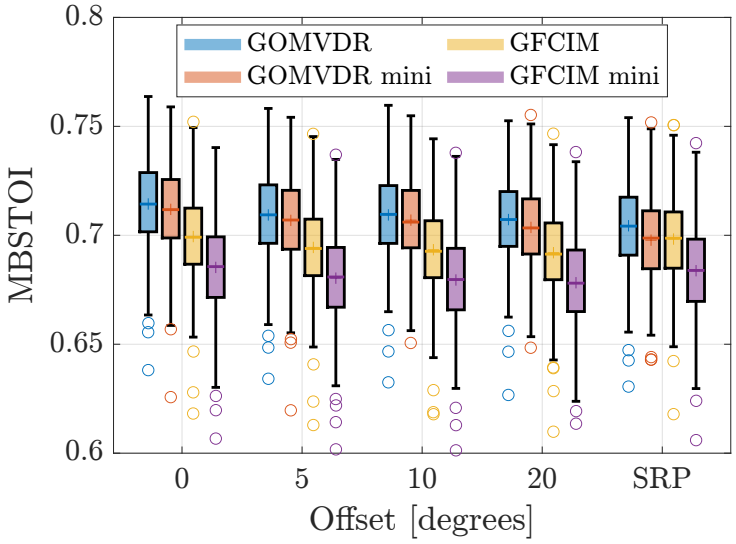


Figure 5.4: MBSTOI score for the beamformer-guided methods with mismatched steering vectors.

enhancement methods the Interaural Signal Phase Difference (ISPD) error is very low thereby preserving the phase information of the target speaker. Table 5.5 presents the maximum ILD error in dB for the target speaker, calculated using (4.16) for the noisy speech signals processed by the models. The results indicate that the maximum ILD errors for all the proposed methods are below 1.5 dB, which is below the threshold of noticeable human hearing perception [7]. When comparing the multichannel methods with the two-channel networks discussed in Ch. 4, similar performance is observed under anechoic conditions. The guided MVDR (GMVDR) exhibited fewer artefacts at iSNRs below 0 dB. In reverberant conditions, the multichannel methods demonstrated superior performance compared to the two-channel approaches, with size-reduced models achieving performance comparable to the two-channel methods. The additional channels provided by the multichannel methods enable the network to achieve improved performance in reverberant environments. The listening examples and implementation code can be found online [221].

5.5.1 Complexity

The multichannel methods, similar to the two-channel approaches outlined in Ch. 4, utilize the full 2-second spectrogram during training and evaluation. However, this approach

may not be practical for real-world implementation, particularly in hearing devices with constrained processing power. To improve feasibility, the implementation can be modified to process shorter signal frames with zero-padding applied as needed, depending on the hardwares computational capacity. For the beamformer-guided methods, latency is determined by the combination of the STFT frame length and the inference frame length. While this could be optimized for beamformers using techniques proposed in [237], such optimization lies beyond the scope of this thesis. Similarly, techniques like knowledge distillation [238] and hardware-specific optimization can be introduced for the neural networks while preserving the core binaural speech enhancement principles introduced.

5.6 Evaluation on real data

This section investigates the performance of the proposed methods on a dataset of recorded real conversations in restaurant noise. The benefits of multichannel enhancement over typical 2-channel enhancement are also discussed.

5.6.1 The GiN dataset

The Group in Noise (GiN) dataset [239] contains multimedia recordings of participants involved in natural group conversations in restaurant noise. One of the participants is equipped with binaural microphones and a 5-microphone array mounted on a pair of glasses. These devices, analogous to earphones and smart glasses in real-life scenarios, can be used jointly to enhance the conversation following the methodology presented in this chapter. The recordings in GiN were obtained in 3 different reverberant rooms and provided challenging test samples. Recording segments that included multiple overlapping speakers were discarded. The remaining segments were resampled to 16 kHz and segmented to create a total of 1048 real-world single-speaker recordings of 2 seconds and are referred to as tokens. The distribution of the considered GiN tokens as a function of fwSegSNR is shown in Fig. 5.5: a majority of recordings exhibit an input fwSegSNR around -15 dB; some are easier (higher fwSegSNR) and typically represent self-speech; and most signals are more challenging (lower fwSegSNR) and contain sections of the conversation where the target speaker is the 'Waiter', a distant and ambulant participant.

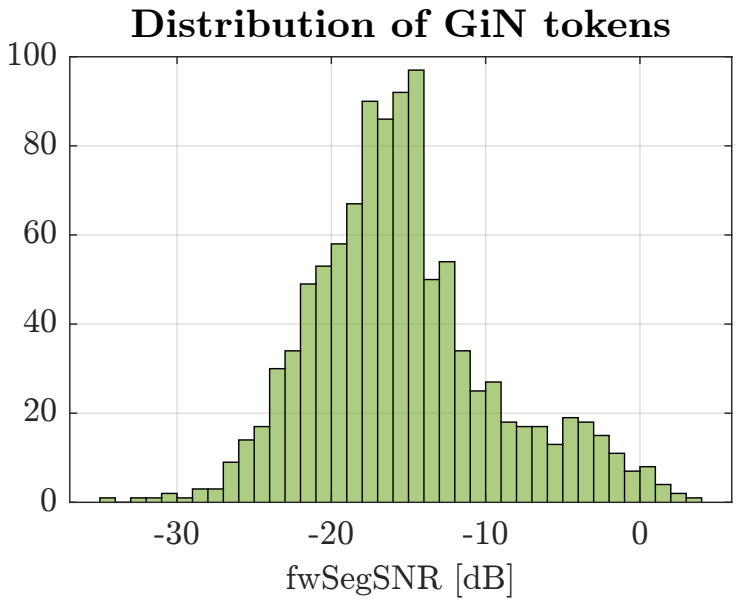


Figure 5.5: Distribution of the considered subset of recordings as a function of measured fwSegSNR.

5.6.2 Enhancement and evaluation

The performance of the FCIM, GFCIM, and MBCCTN are explored as in Sec. 5.4. The beamformer-based approaches assume a steering direction in front of the participant, motivated by the robustness to steering vector error discussed in Sec. 5.4.4. The covariance matrix in (5.6) is discretised to the 5° resolution of the ATF in GiN. Since the MBCCTN assumes a 6-channel input while GiN uses a 7-channel array (2 binaural and 5 glasses-mounted microphones), the microphone placed on the nose bridge of the glasses is not used by the fully neural approach. Finally, the 2-channel neural approach in Ch. 4, Binaural Complex Convolutional Transformer Network (BCCTN), is also evaluated to investigate the benefit of multichannel enhancement over typical binaural methods. Intrusive metrics are computed using the close-talking microphone signals provided in GiN as references.

5.6.3 Results

For each considered enhancement method, the cumulative density function of its improvement in fwSegSNR relative to the raw recordings at the binaural microphones (ΔfwSegSNR) is shown in Fig. 5.6, and Fig. 5.7 shows the improvement obtained in

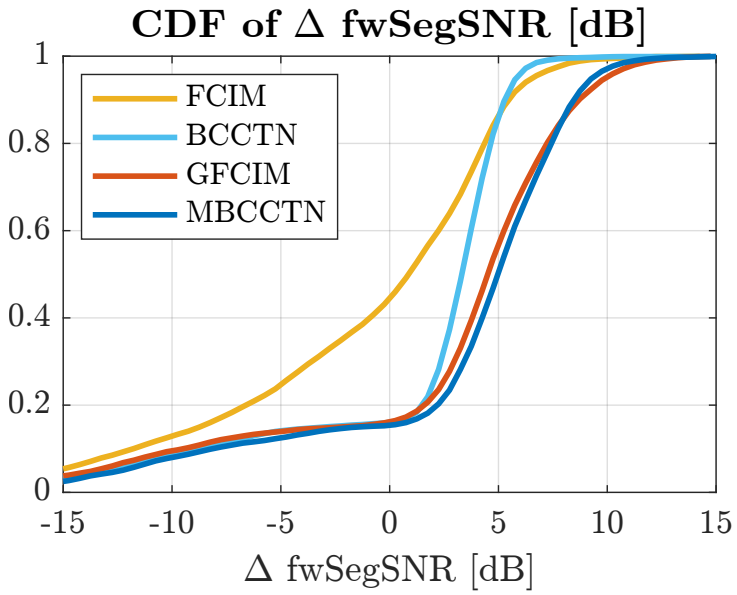


Figure 5.6: Cumulative Density Functions of fwSegSNR enhancement for various methods

each GiN token grouped as a function of input fwSegSNR. Similar results are obtained for STOI and are not included here for brevity. The traditional FCIM beamformer leads to an improvement in a little over 50% of the considered GiN tokens: this is expected for some of the challenging tokens in which the forward-steering assumption may not be valid (e.g. when the target speaker is distant and moving), and is confirmed by a bimodal distribution of the enhancement for low input fwSegSNR in Fig. 5.7. For tokens containing self-speech (high input fwSegSNR), the fixed beamformer is likely to introduce target signal cancellation. The GFCIM and MBCCTN exhibit very similar enhancement distributions, providing more than 5 dB improvement in 50% of the tokens. The GFCIM shows a slightly higher median enhancement at low input fwSegSNR than the MBCCTN, indicating a benefit to prior denoising using a fixed beamformer. The BCCTN yields an enhancement of smaller magnitude compared to GFCIM and MBCCTN: only approximately 10% of tokens experience an enhancement greater than 5 dB. This confirms the need to exploit all microphones available for enhancement. All neural methods fail to enhance the signal in approximately 15% of tokens, corresponding to the very challenging scenarios where the target speaker is far away and moving. Informal listening tests indicate that the neural methods introduce distortions in such tokens, suggesting they differ greatly from examples seen in training.

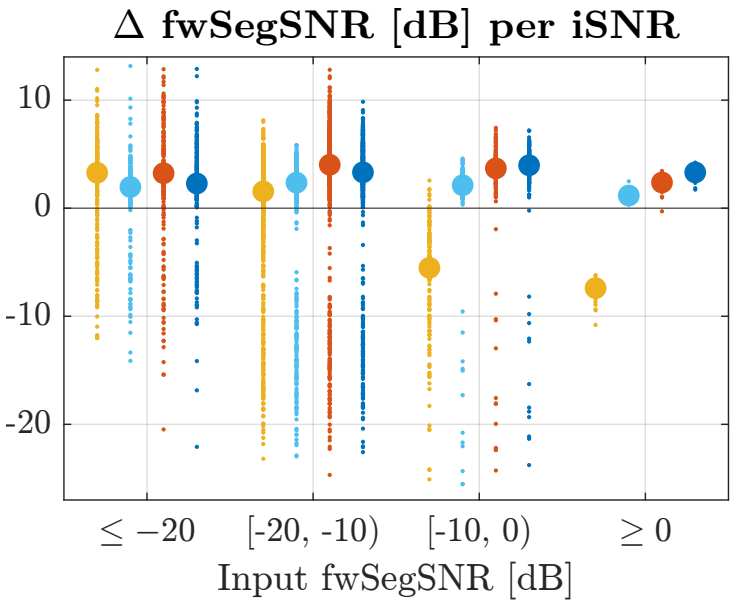


Figure 5.7: fwSegSNR improvement for individual GiN token grouped by input fwSegSNR, with large dots representing the median value for each iSNR range.

5.7 MUSHRA test

A MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test was conducted with 15 normal-hearing participants who were not trained listeners. The participants were asked to listen to a reference signal and rate the test signals on a scale of 0 to 100. The iSNRs of the test signals were selected randomly, and a total of 36 test signals were presented in both anechoic and reverberant conditions. The multichannel signals were generated using the HRIRs from [211] and [224]. The noisy and reverberant signals were generated as outlined in Sec. 5.4.1 for iSNRs from -6 dB to 15 dB. Participants listened to the signals using a Beyerdynamic DT1990 Pro headset, and the signals were played through an RME UCX Fireface II. The front microphone signals from the HRIRs were used as the clean and noisy signals for the headphone-based listening tests, while the enhanced signals were processed binaurally by the proposed methods. Figure 5.8 shows the average score across all iSNRs, and Fig. 5.9 displays the average rank of the methods in anechoic conditions. Figure 5.10 presents the average score over all iSNRs, and Fig. 5.11 shows the average rank of the methods in reverberant conditions. In the anechoic condition, listeners rated OMVDR and GOMVDR the highest, with the MBCCTN method

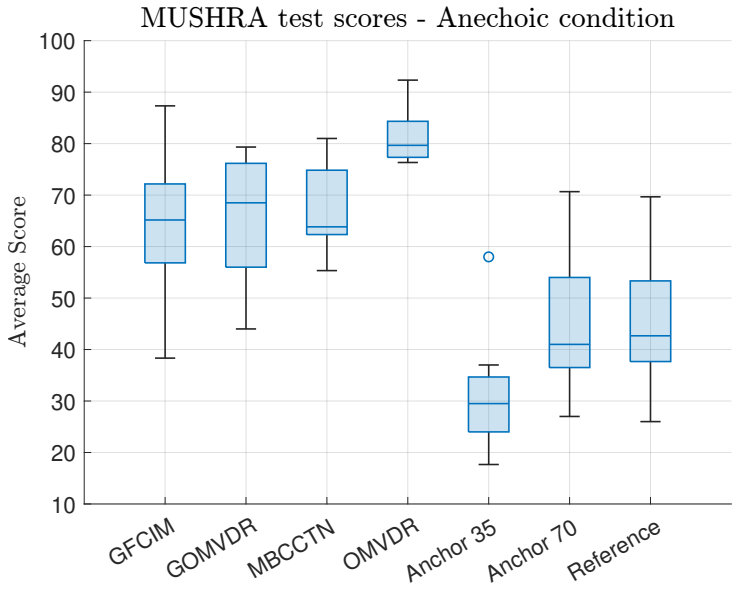


Figure 5.8: MUSHRA test scores for MBCCTN compared with the baselines (box plot illustrating the median (central line), upper quartile (75%), lower quartile (25%), and the maximum and minimum outliers).

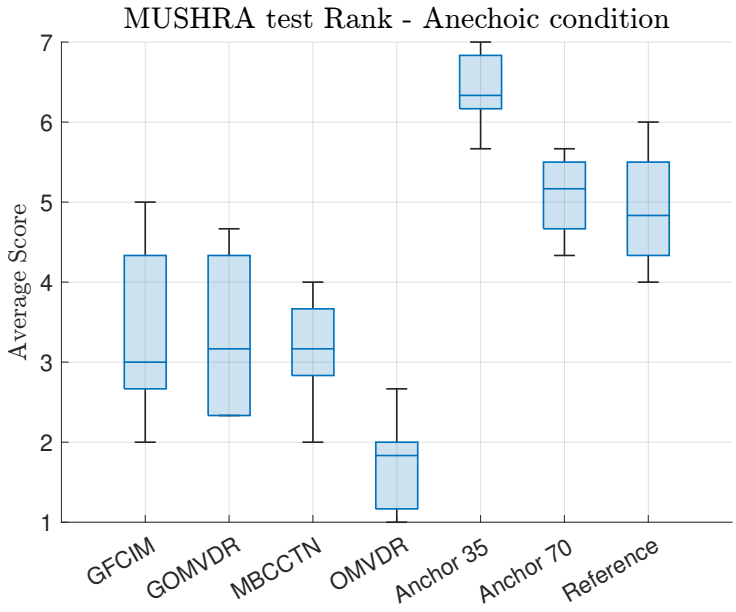


Figure 5.9: MUSHRA test ranks, ranging from 1 (best) to 7 (worst), for MBCCTN compared with the baselines over various noisy iSNR.

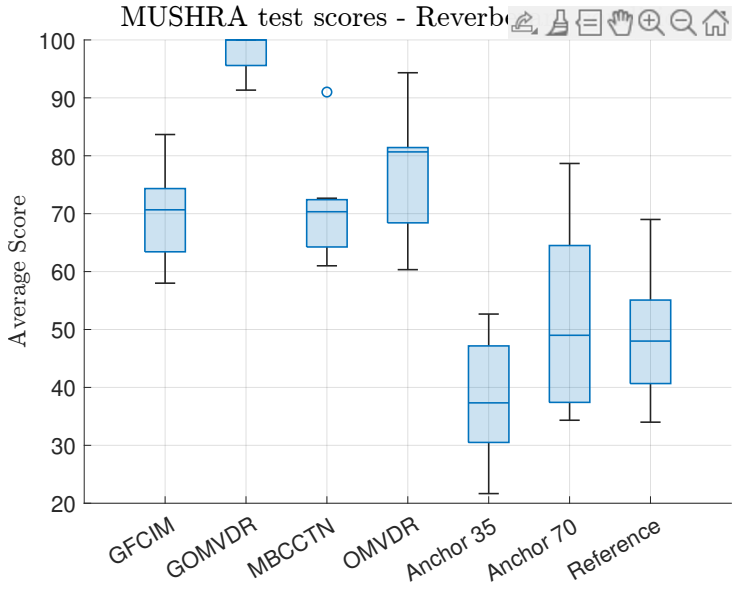


Figure 5.10: MUSHRA test scores for MBCCTN compared with the baselines.

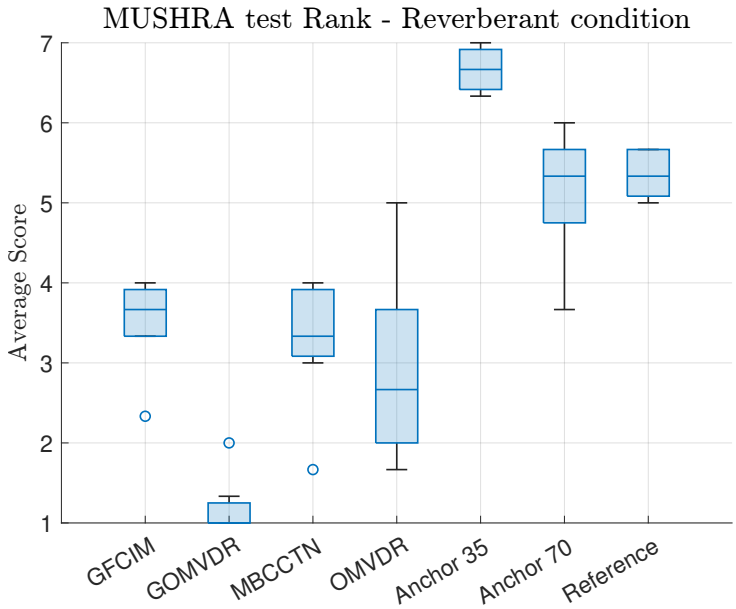


Figure 5.11: MUSHRA test ranks, ranging from 1 (best) to 7 (worst), for MBCCTN compared with the baselines over various noisy iSNR.

scoring close to GOMVDR despite not having access to clean speech oracle information. In the reverberant condition, GOMVDR received the highest scores, while OMVDR and MBCCTN followed with similar scores, although OMVDR exhibited greater variance. The listening tests demonstrated that, although MBCCTN did not rely on oracle information or beamformer support, its performance was comparable to the guided and oracle methods. GFCIM, which made assumptions about the NCM, performed lower than the oracle methods due to the beamformers suboptimal performance.

5.8 Conclusion

This chapter has introduced multichannel binaural speech enhancement methods based on complex convolutional networks. A fully end-to-end DNN-based method and a beamformer-guided DNN version were proposed for multichannel binaural speech enhancement. The beamformer-guided models were implemented using SRP for target DOA estimation and evaluated for mismatches in steering vector estimation. The proposed methods achieved a maximum fwSegSNR improvement of 13 dB and a MBSTOI score improvement of 0.2. Experimental results demonstrated that the proposed methods effectively reduced added noise, improved binaural intelligibility, and preserved interaural cues. These improvements are significant for real-world hearing applications, where maintaining speech clarity and spatial awareness is crucial for effective communication in noisy environments. The enhanced intelligibility and spatial preservation lead to a noticeable difference in hearing, allowing users to better distinguish speech from background noise and perceive sound sources more naturally. The model size of the networks can be significantly reduced with only a modest loss in performance. Mismatches in steering vectors for the beamformer-guided versions had minimal impact on the models performance. The performance was further evaluated with real recordings of natural group conversations in restaurant noise, showing significant improvement in fwSegSNR. The listening test results indicate that the MBCCTN method performed competitively with guided and oracle methods in both anechoic and reverberant conditions, despite lacking oracle information and beamformer support. Although GFCIM was less effective due to its reliance on assumptions about the NCM, the MBCCTN approach demonstrated robust performance close to that of the highest-rated oracle methods.

Chapter 6

Binaural Localization based on Human Auditory Perception

Binaural acoustic source localization is important to human listeners for spatial awareness, communication and safety. In this chapter, an end-to-end binaural localization model for speech-in-noise is presented, which is a significant step towards our final goal of predicting the effect of binaural speech enhancement algorithms on human sound localization. A lightweight convolutional recurrent network that localizes sound in the frontal azimuthal plane for noisy reverberant binaural signals is introduced. The localization performance of the model is compared with the steered response power algorithm and the use of the model as a measure of interaural cue preservation for binaural speech enhancement methods is studied. A listening test was performed to compare the performance of the model with human localization of speech in noisy conditions. An earlier version of the work presented in this chapter has been submitted for publication [240].

6.1 Introduction

Localizing sound sources using binaural input in noise and reverberation is a challenging problem with important applications in hearing aids, spatial sound reproduction, and virtual reality devices. Binaural localization has garnered significant attention in the field of computational auditory scene analysis (CASA), which is influenced by principles underlying the perceptual organization of sound by human listeners. The two primary cues for sound localization are the interaural time difference (ITD), also known as the time difference of arrival, and the ILD, which arises due to the influence of the head, torso, and outer ear. Differences between localization methods often stem from varying assumptions about environmental factors such as sound propagation, background noise, and microphone configuration.

Typical listening environments are often noisy and reverberant, which can mask signals and negatively affect both binaural and monaural spectral cues, leading to reduced sound localization accuracy and speech comprehension even for individuals with normal hearing [241–243]. Research has shown that localization accuracy declines as the SNR decreases. For example, [241] studied three normal hearing listeners who were asked to localize broadband click trains in an anechoic chamber under one quiet and nine noisy conditions with SNRs ranging from -13 to +14 dB. It was found that localization accuracy was poorest in the lateral horizontal plane and began to deteriorate at higher SNRs (+4 to +8 dB). Similarly in [242] the effect of SNR on localization ability in normal hearing listeners, finding that typical environments characterized by both noise and reverberation can further degrade localization cues and impair performance. In [244], it is suggested that the combined effects of noise and reverberation could further reduce localization accuracy.

A well-known method for localisation using ITD estimation is the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) approach [245], which assumes ideal single-path propagation. Although GCC and similar methods can be applied to any setup with two or more microphones, some recent research has focused on localization models specifically designed for binaural systems [246,247]. Recent efforts have integrated azimuth-dependent models of ITD and ILD, demonstrating that jointly considering both cues enhances azimuth estimation compared to using ITD alone [246–248]. However, these models often require prior training or calibration with the binaural input due to the

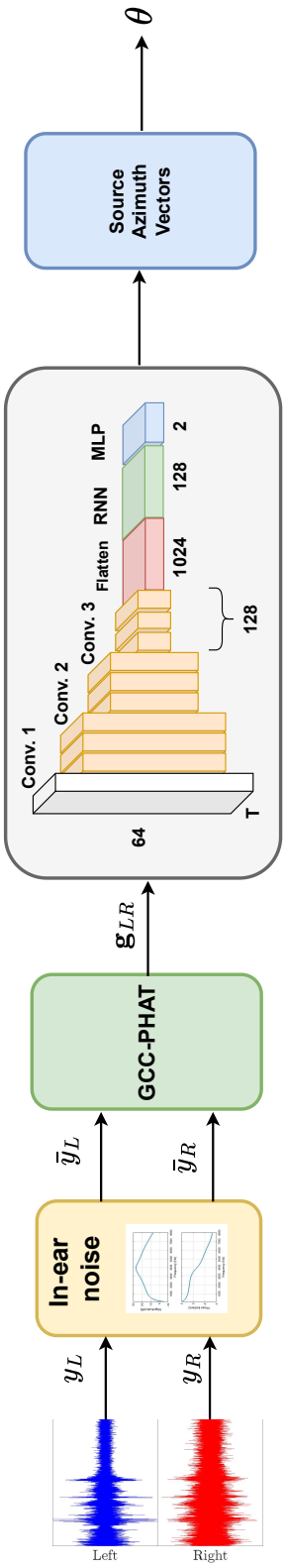


Figure 6.1: Block diagram of the model architecture.

significant variability in the frequency-dependent patterns of ITDs and ILDs across individuals, which can lead to performance degradation in different binaural setups. Methods also differ in how they integrate interaural information across time and frequency, with these variations largely reflecting different assumptions about source activity and interaction. The authors of [246], proposed a framework that determines the likelihood of each source location based on a Gaussian Mixture Model (GMM) classifier, which learns the azimuth-dependent distribution of ITDs and ILDs from joint analysis of both binaural cues. However, many binaural localization methods have focused on scenarios with minimal reverberation or background noise. One approach to improving localization in more complex environments involves using model-based information about the spectral characteristics of sound sources in the acoustic scene to selectively weight binaural cues. This involves estimating models for both target and background sources during a training stage, using spectral features derived from isolated source signals [247]. In [2], an end-to-end binaural localization algorithm that estimates the azimuth using convolutional neural network (CNN)s to extract features from the binaural signal was introduced.

Human auditory cognition includes complex neurological processes for localization. Although ILDs and ITDs are widely accepted to be the primary interaural cues that strongly influence human sound source localization [241], there is no standardized way to compute them. Precedence effect, spectral cues, head movement and other psychoacoustical processes affect sound localization in humans. There is no universally accepted method of measuring the correlation between human sound localization and the frequency-varying interaural cues. In Ch. 4, to demonstrate the preservation of spatial cues, the error in interaural cues of the enhanced speech was computed using a ideal binary mask (IBM) that selects the speech-active regions in the signal.

A relevant approach to measuring the accuracy with which spatial information is preserved, and the subsequent accuracy of localization of speech sources in noisy and enhanced speech signals, would be to employ a model that predicts the localization of the speech-in-noise in a manner highly correlated to a human listener. This chapter sets out to research methods of DOA estimation that are not necessarily the best performing but specifically follow the performance of the human listener in terms of binaural localization. The research will focus on an end-to-end binaural localization model for speech in noisy and reverberant conditions, introducing a lightweight CRN that utilizes input features based on GCC-PHAT, which is a first step towards this goal. The model adds synthetic

internal ear noise onto an audio signal to simulate the effects of the frequency-dependent hearing threshold of a normal listener. The model is trained on binaural speech data to predict the source azimuth directly without assuming a predetermined azimuth-dependent distribution of interaural cues as previously done in [246]. The proposed model is referred to as BIL. The approach is evaluated using a listening test that was conducted using 15 normal hearing listeners, in which the participants were tasked to localize a target speaker in a simulated noisy and reverberant condition.

6.1.1 Signal model

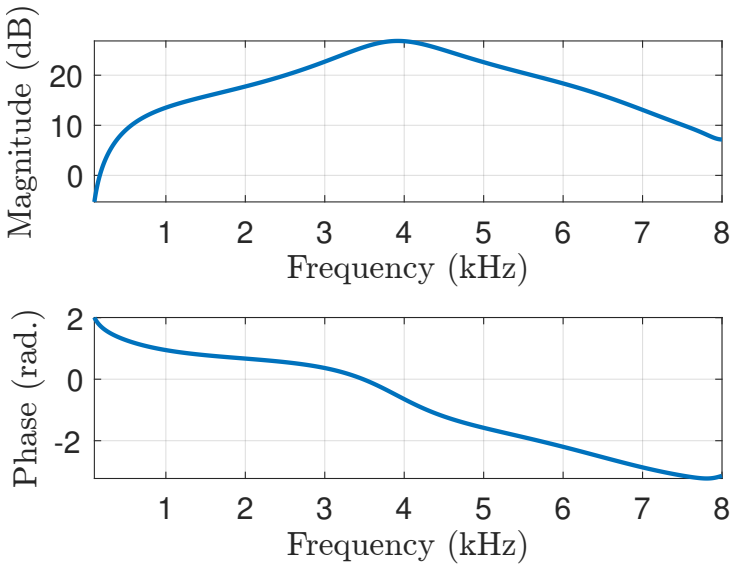


Figure 6.2: Magnitude and phase response of filter used to simulate the listener's hearing threshold.

A binaural system is composed of a left and right channel; these form the inputs to the proposed model architecture shown in Fig. 6.1. The time-domain signal y_L received by the left channel is modeled as

$$y_L(n) = s_L(n) + v_L(n), \tag{6.1}$$

where s_L is the anechoic clean speech signal, v_L is the noise and n is the discrete-time

index. The right channel is described similarly with an R subscript. The model adds artificially generated internal ear noise onto the audio signal to simulate the effects of the frequency-dependent hearing threshold of a normal hearing listener. At each frequency, the spectrum of this noise equals the pure-tone hearing threshold minus $10\log_{10}(C)$ dB where C is the critical ratio [249, 250]. The critical ratio is approximately independent of level and equals the power of a pure tone divided by the power spectral density of white noise that just masks it. To avoid having to add very high noise levels at low and high frequencies, the input signal, $y_L(n)$, is filtered by, $h_e(n)$, whose frequency response, shown in Fig. 6.2, is the inverse of the internal ear noise spectrum [161, 249] before adding white noise with 0 dB power spectral density. The in-ear noise added signal $\bar{y_L}$ is given by

$$\bar{y}_L(n) = h_e(n) * y_L(n) + e_L(n) \tag{6.2}$$

where $h_e(n)$ is the impulse response of the filter depicted in Fig. 6.2 and $e_L(n)$ is the white noise added to the filtered noisy signal [249]. Hearing loss can readily be taken into account here by modifying the filter to additionally attenuate the signal level by the hearing loss at each frequency. The online implementation of the ear noise filter can be found in [161]. The in-ear noise-added signal is then used as the input to the neural network which determines the target azimuth in the frontal azimuthal plane.

6.1.2 Localization network

Input Feature Set

The input feature of the proposed network consists of the GCC-PHAT for the pair of microphone signal frames $(\bar{\mathbf{y}}_\mathbf{L}, \bar{\mathbf{y}}_\mathbf{R})$, defined as

$$\mathbf{g}_{LR} = \text{IDFT}\left(\frac{\bar{\mathbf{Y}}_L}{|\bar{\mathbf{Y}}_L|} \odot \frac{\bar{\mathbf{Y}}_R^*}{|\bar{\mathbf{Y}}_R|}\right), \tag{6.3}$$

the Inverse Discrete Fourier Transform (IDFT) of the element-wise product of the normalized frequency-domain frames $\bar{\mathbf{Y}}_L$ and $\bar{\mathbf{Y}}_R$, where $\bar{\mathbf{Y}} = \text{DFT}(\bar{\mathbf{y}})$ and $|\bar{\mathbf{Y}}|$ is the element-wise magnitude.

Network architecture

As shown in Fig. 6.1, the network is composed of a set of convolutional blocks, followed by an operation of flattening of the frequency and channel dimension. The resulting tensor is then used as input for a gated recurrent unit (GRU). Finally, a linear layer is applied to produce, a 2-D output vector, $\hat{\mathbf{v}}$, representing the direction of the source's azimuth.

6.1.3 Loss function

The proposed model is trained using a modification of the cosine similarity given by

$$\mathcal{L}(\mathbf{v}, \hat{\mathbf{v}}) = 1 - \frac{|\mathbf{v} \cdot \hat{\mathbf{v}}|}{\|\mathbf{v}\|\|\hat{\mathbf{v}}\|} \tag{6.4}$$

where $\mathbf{v}$ and $\hat{\mathbf{v}}$ are the vectors for the true and estimated angles. The loss function (6.4) was designed so that the absolute value of the cosine similarity between the vectors is maximized, therefore not penalising the effects caused by the front-back ambiguity, which are expected when employing only two microphones in binaural configuration.

6.2 Experiments

6.2.1 Dataset

To generate binaural speech data, monaural clean speech signals were obtained from the CSTR VCTK corpus [212] and spatialized using the measured BRIRs from [224] for training which were recorded in a listening room with a T_{60} of 460 ms. The VCTK corpus contains approximately 13 hours of speech data from 110 English speakers with various accents. These recordings were used to create 2 s speech utterances, which were spatialized to produce left and right ear channels. The resulting dataset comprised 20,000 speech utterances, which were divided into training (70%), validation (15%), and testing (15%) sets. Diffuse isotropic speech-shaped noise was generated using uncorrelated noise sources uniformly distributed every 5° in the azimuthal plane [186]. The binaural signals were generated with the target speech positioned at a random azimuth in the frontal plane (−90° to +90°) with the source positioned at a distance of 100 cm. For the training

process, isotropic noise was added so that the average in dB of (SNR_L, SNR_R), ranged between -25 dB and 25 dB. The evaluation set comprised speech signals spatialized with BRIRs from [211] with random target azimuths in the range of ±90°, and isotropic noise added at random SNRs between -25 dB and 25 dB. The speaker was positioned at a 0° elevation and at a distance of 3 m. This ensured that training and evaluation sets contained binaural signals generated using different BRIRs to verify that the network generalised to different heads.

6.2.2 Training Setup

The two-second input signals were sampled at 16 kHz and a window size of 512 was used to generate the signal frames with a 25% overlap for a hop size of 24 ms. The parameters for the localization network are detailed in Fig. 6.1, which includes the tensor output shapes for each layer of the network. Convolutional layers employed a kernel size of (3,3) throughout. Max pooling with a kernel size of 2 was applied to all convolutional layers except the last one. The PReLU activation function was utilized in all layers of the network, except for the recurrent neural network (RNN) and multi-layer perceptrons (MLP) output layers, which used hyperbolic tangent (tanh) activation, and the output layer, which employed sigmoidal activation. This architecture was taken from [251] and modified to work for binaural signals. The network has 850K parameters and is implemented using the PyTorch library, and the Adam optimizer was used for backpropagation. The network was trained for 80 epochs with early stopping enabled. The code for implementation is available online [252].

6.2.3 Listening Tests

In the listening tests, 15 participants with normal hearing were tasked with localizing a target speaker within the frontal azimuthal plane. The noisy binaural signals were generated using BRIRs from [211] and isotropic noise was added as described in Sec. 6.2.1. Using Beyerdynamic DT1990 Pro open-back headphones, the audio signals were delivered in a soundproof booth through an RME Fireface UCX II audio interface. The participants were required to listen to the noisy speech utterances and select the perceived azimuth using a MATLAB-based GUI shown in Fig. 6.4. The azimuths were quantized to 15° intervals. Each participant listened to 36 speech utterances, which were evenly distributed

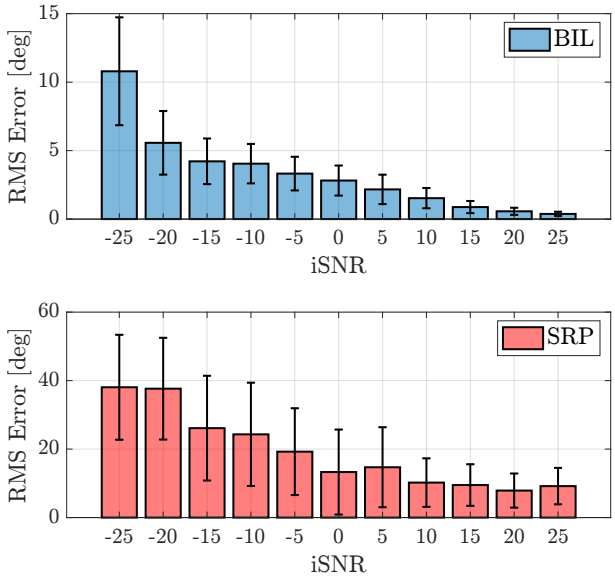


Figure 6.3: The localization error in noisy reverberant conditions for the proposed method (BIL).

across different SNRs and randomly assigned azimuths in the frontal azimuth plane. Three conditions of iSNR were used in the test: a very noisy condition at -15 dB, a noisy condition with 0 dB and the low noise condition at 15 dB iSNR.

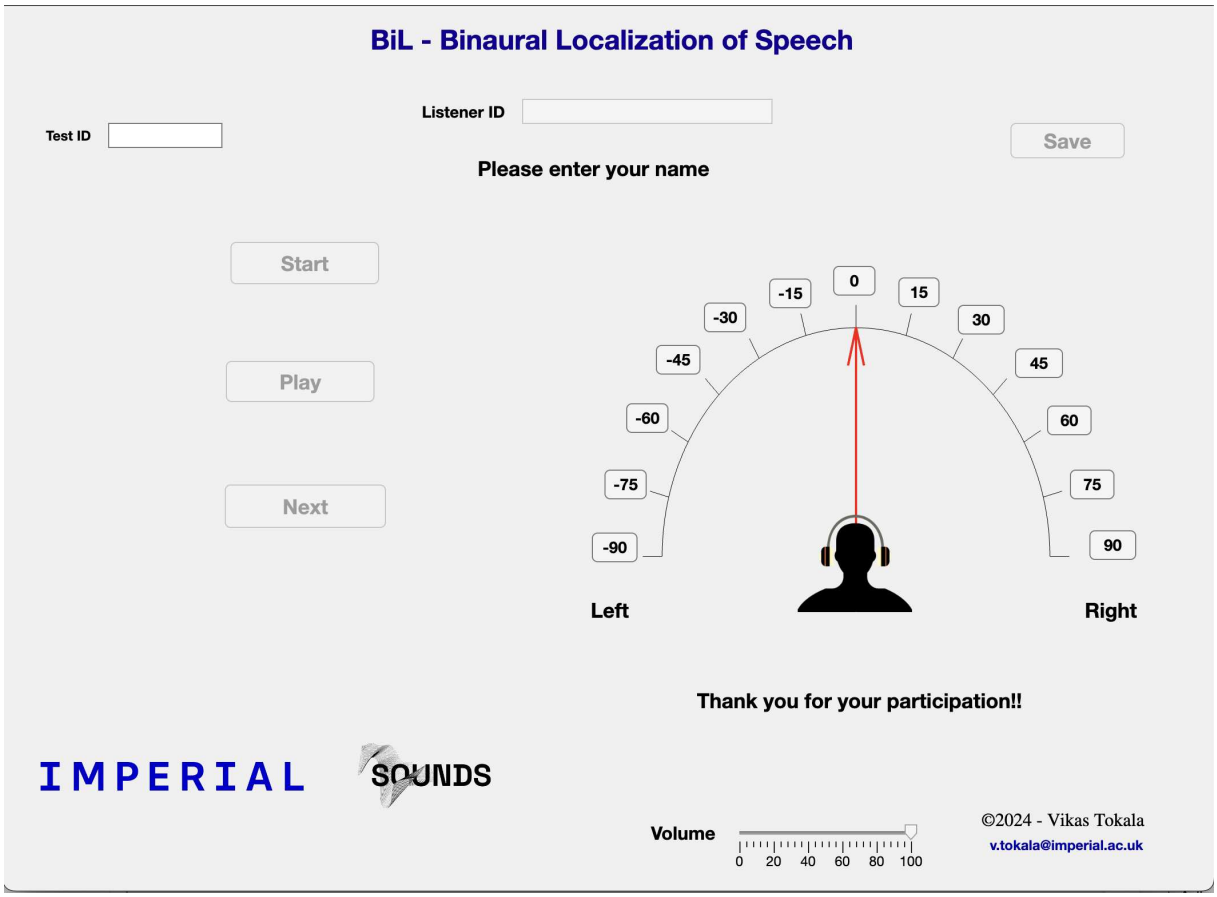


Figure 6.4: MATLAB GUI of the localization test

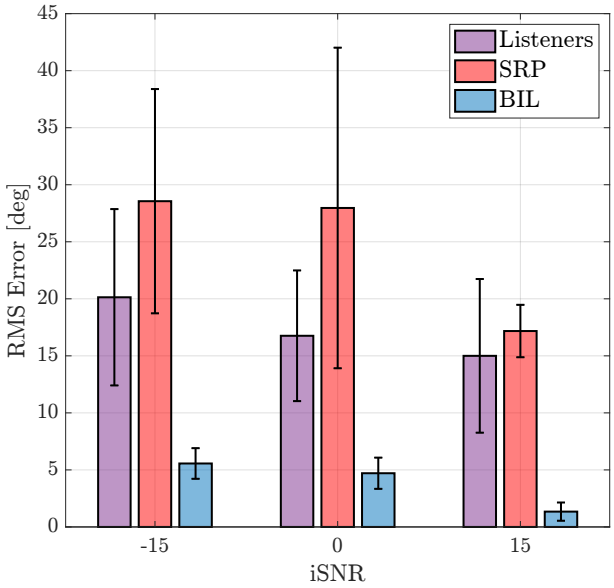


Figure 6.5: The plots show the localization error for listeners compared with the proposed method

6.3 Results and Discussion

The model was evaluated using 275 speech utterances for each noisy iSNR ranging from -25 dB to +25 dB in steps of 5 dB. The localization error for the proposed method, referred to as BIL, is shown in the upper plot of Fig. 6.3 for different iSNRs. The azimuth θ of the target speaker's DOA in the frontal azimuth plane is estimated using the SRP-PHAT algorithm [235, 236] is shown in the lower plot of Fig. 6.3. In extremely noisy conditions, such as -25 dB, the proposed method achieves a localization error of approximately 15°. Under similar iSNR conditions, the localization error for SRP is considerably higher, around 40°. As the iSNR improves, the localization error for the proposed method decreases to below 5°, eventually reaching just under 1° at 25 dB iSNR. In contrast, the

Method	Localization Error
SRP-PHAT	10.2°
WaveLoc-GTF [2]	3.0°
WaveLoc-CONV [2]	2.3°
BIL	1.2°

Table 6.1: Localization error compared to WaveLoc [2] methods.

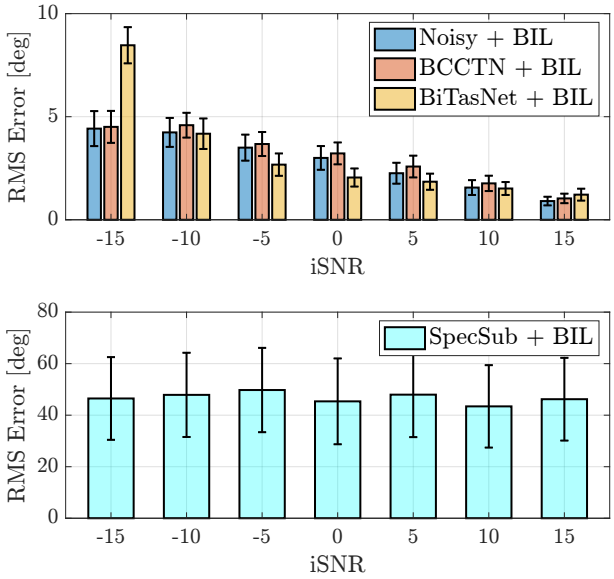


Figure 6.6: The plots show the localization error for signals processed by different enhancement methods evaluated by BIL.

SRP method maintains an error between 10° and 20° even at high iSNRs. The reduced performance of SRP at higher iSNRs can be attributed to reverberation, which causes multiple peaks in the correlation [251]. Table 6.1 shows the comparison of localization error with the WaveLoc methods proposed in [2]. These methods are also evaluated on BRIRs from [224] recorded in the listening rooms without the addition of isotropic noise and the values shown are taken from [2]. For similar conditions, the proposed method has a lower error than both versions of WaveLoc methods.

Figure 6.5 shows the localization error of human listeners compared with the proposed method and SRP for the three conditions of noisy signals as described in Sec. 6.2.3. The proposed method has significantly lower localization error for all the iSNR conditions. Listeners had an average error of 20° in the very noisy condition of -15 dB and an average error of 15° in the low noise condition of 15 dB given that there was no head movement to assist them. SRP-based localization had the highest localization error and standard deviation for the test samples. Previous studies have shown that human localization of speech and tones can have a localization error of up to 40° when noise and reverberation are present [241–243]. If the signals processed by enhancement methods produce a low localization error with the proposed method, the expected explanation is that the inter-aural cues of the signal are preserved and human listeners will still localize the target

speaker in the same azimuth as the original noisy signal.

Figure 6.6 demonstrates how the proposed method can be used to assess the performance of binaural speech enhancement methods in preserving the interaural cues and the spatial information of the target speaker. While there are well-known objective measures to evaluate noise reduction, speech intelligibility and quality, there are no standardised measures to assess the preservation of binaural cues after they are processed by enhancement algorithms. The upper plot in Fig. 6.6 shows the localization error for noisy signals at the iSNRs from -15 dB to 15 dB and the signals processed by BCCTN [214] and binaural TasNet (BiTasNet) [1] at the same iSNRs measured by the proposed model. The binaural enhancement algorithms are generally designed to preserve the interaural cues while enhancement and they show a low localization error. At -15 dB the BiTasNet shows a higher error compared to the noisy input signal, which indicates disruption in the interaural cues, and this is expected as the method was not designed to perform enhancement at -15 dB. As the iSNR improves, all the binaural enhancement methods show localization error under 5° which signifies the preservation of interaural cues. From Fig 6.3 - Fig. 6.6, it is evident that the proposed model has a monotonic relationship to SNR, i.e., the localization error decreases with increasing iSNR. Furthermore, other studies including [241–243] show that human localization capability is monotonically proportional to SNR. Hence, the proposed method has been seen to be, as desired, highly correlated with human binaural localization - a conclusion which is supported by the subjective listening tests conducted. The lower plot in Fig. 6.6 shows the localization error obtained when the noisy signals are processed with bilateral spectral subtraction (SpecSub) [190], where no attempt is made to preserve binaural cues. The localization error is obtained around 45° as the testset contains signals which have azimuths distributed randomly between ±90°. If the binaural enhancement methods are being used for purposes other than human listening, the addition of in-ear noise can be omitted before performing localization which would further improve the localization performance of the proposed model.

6.4 Conclusion

This chapter presented an end-to-end binaural localization model for speech in noisy and reverberant conditions. A CRN network utilizing GCC-PHAT features was introduced, and a listening test with 15 normal-hearing listeners showed that the model closely aligns

with human perception, albeit with lower localization error. The model effectively evaluates the localization error of binaural speech enhancement algorithms, correlating with spatial information preservation and interaural cue retention. The key objective was to develop a DOA estimation method that mirrors human binaural localization rather than purely optimizing accuracy. The proposed method demonstrated significantly lower localization errors across all iSNR conditions. Listeners had average errors of 20° at -15 dB and 15° at 15 dB without head movement. SRP-based localization showed the highest error and variability and as iSNR improves, all binaural enhancement methods exhibit localization errors below 5°, confirming interaural cue preservation. The models localization error follows a monotonic relationship with SNR, aligning with human performance trends.

Chapter 7

Conclusions

This chapter summarises the work presented in this thesis and outlines some future research.

7.1 Summary of Thesis Contributions

The research goals of this thesis were to research and develop binaural speech enhancement methods for hearing devices. Accordingly, its primary contributions can be broadly classified into the following categories.

1. **Extension of STOI-Optimal Masking to Binaural Speech Enhancement**

   A novel approach for enhancing noisy binaural speech signals using STOI-optimal masking was introduced. This hybrid method integrates classical digital signal processing (DSP) techniques with DNNs to enhance the SNR and intelligibility of noisy signals while maintaining the interaural cues. Instead of directly applying TF masks as gain functions, speech presence probability (SPP) is computed and incorporated into an optimally-modified log spectral amplitude (OM-LSA) speech enhancement framework, ensuring the smooth integration of the estimated masks. The proposed approach achieves notable improvements in fwSegSNR, MBSTOI scores, and preserves interaural cues with minimal distortion.

2. **Binaural speech enhancement using complex-valued neural networks**

This work presented complex-valued convolutional transformers and recurrent networks for binaural speech enhancement, introducing a novel loss function to optimize the network for noise reduction, speech intelligibility enhancement, and interaural cue preservation. Feature maps in the TF domain were utilized to illustrate the flow and transformation of data within the networks. Experimental evaluations demonstrated that the proposed methods achieved significant improvements in MBSTOI and fwSegSNR metrics while maintaining minimal ILD and IPD errors. Among the models, the BCCTN consistently outperformed the Binaural Complex Convolutional Recurrent Network (BCCRN), demonstrating enhanced robustness in challenging acoustic environments. Furthermore, MUSHRA listening tests validated the quantitative results, showing that the BCCTN produced fewer artefacts than the BCCRN. These findings emphasize the practical potential of the BCCTN for binaural speech enhancement applications.

3. **Development of Multichannel Binaural Speech Enhancement Methods**

Multichannel binaural speech enhancement methods based on complex convolutional networks, proposing both a fully end-to-end DNN-based approach and a beamformer-guided DNN variant for multichannel binaural speech enhancement were introduced in Ch. 5. The beamformer-guided models utilized SRP for target DOA estimation and were evaluated under conditions of steering vector mismatch. The proposed methods achieved a maximum improvement of 13 dB in fwSegSNR and a 0.2 improvement in MBSTOI scores, demonstrating significant noise reduction, enhanced binaural intelligibility, and effective preservation of interaural cues. Notably, the model sizes could be reduced substantially with minimal performance loss. Experimental results showed that mismatches in steering vector estimation had negligible effects on the performance of the beamformer-guided models. Evaluation with real recordings of natural group conversations in restaurant noise further confirmed significant fwSegSNR improvements. Listening test results revealed that the MBCCTN method performed competitively with guided and oracle methods under both anechoic and reverberant conditions, despite operating without oracle information or beamformer support. While GFCIM proved less effective due to its reliance on NCM assumptions, the MBCCTN approach delivered robust performance, closely matching the highest-rated oracle methods. These findings highlight

the MBCCTNs potential for practical multichannel binaural speech enhancement applications. The choice of the multichannel method can be decided by the available hardware, processing power and the application. In scenarios where the microphone configuration is unknown during training or subject to change during usage, the beamformer-guided methods provide a robust solution due to their hardware-agnostic nature. These methods can adapt to any microphone configuration and remain functional even when a subset of microphones fails, eliminating the need for retraining the neural network. Conversely, in applications where the microphone configuration is fixed and known, fully neural models deliver superior performance but require training specific to the given configuration. This distinction makes the beamformer-guided methods highly versatile for dynamic environments, while fully neural models excel in static setups.

4. **End-to-End Binaural Speech Source Localization**

   Chapter 6 introduced an end-to-end binaural localization model for speech in noisy and reverberant environments. A CRN network utilizing input features derived from GCC-PHAT was proposed for binaural localization. A listening test conducted with 15 normal-hearing participants demonstrated that the models performance closely aligns with human localization capabilities, albeit with lower error rates. The proposed model was further employed to assess localization errors introduced by binaural speech enhancement algorithms, highlighting its ability to correlate with the spatial information and interaural cue preservation of these methods.

7.2 Discussion and Areas for Future Research

This thesis focused on the research and development of binaural speech enhancement methods. The proposed approaches successfully enhanced noisy binaural speech signals, demonstrating significant advancements in noise reduction, speech intelligibility, and preservation of interaural cues. Despite these achievements, several opportunities remain for future research to further improve and expand upon the methods presented in this work.

The STOI-optimal masking method introduced in Ch.3 could benefit from replacing the DNNs with complex-valued neural networks (CVNNs), offering a potential avenue to

further enhance performance. Additionally, the models discussed in Ch.4 are characterized by their large model sizes, posing challenges for deployment on devices with limited computational power. Future research should focus on reducing the model size while minimizing performance degradation. Techniques such as knowledge distillation [238], which transfers the learned parameters of a large neural network to a smaller network, hold promise. This approach has demonstrated the ability to achieve performance levels comparable to larger networks, making it a compelling direction for optimizing resource-constrained implementations. Although model reduction was demonstrated for MBCCTN in Ch. 5, there remains an opportunity to explore knowledge distillation for these models. This technique could further optimize the models by transferring knowledge from larger networks to smaller ones, potentially achieving similar performance with reduced computational requirements. Such exploration would enhance the practicality of deploying MBCCTN models in resource-constrained environments.

For the models presented in this thesis, a detailed analysis of computational complexity has not yet been explored. The methods were designed with a focus on incorporating causal convolutions and enabling online implementations. However, conducting a thorough complexity analysis would be highly beneficial, particularly when considering the implementation of these models in binaural hearing devices. Such an analysis could provide insights into the latency of the methods optimise computational efficiency and ensure real-time performance in practical applications. The next critical step toward the practical application of these models is their deployment on hardware platforms. Hardware implementation introduces various challenges within the product development cycle. Depending on the target hardwarewhether hearing aids, headphones, or virtual/augmented reality devicesthe models will require retraining and size modifications to align with the computational power constraints of the specific device. Furthermore, these devices cater to diverse end-user groups, necessitating tailored data collection and training processes to address the unique requirements of each use case. This adaptation is essential to ensure optimal performance and user satisfaction across different applications.

Conducting comprehensive listening tests for the models is essential to better understand listener needs and refine the data and training processes to optimize the methods for specific use cases. Listening tests with trained listeners can provide valuable insights for fine-tuning the methods to achieve superior binaural speech enhancement. A significant challenge in the field of binaural speech enhancement is the absence of reliable objective

measures that accurately quantify enhancement. While MBSTOI [253] is a widely used metric, it has limitations, particularly in environments with modulating noise and high reverberation. The localization method introduced in Ch. 6 represents a foundational step toward developing a metric capable of localizing binaural speech and objectively evaluating interaural cue preservation in enhancement methods. Future research should aim to further regularize this model by incorporating psychoacoustic data and listening test results to more effectively mimic human speech localization in complex acoustic environments. Such advancements could lead to more robust and reliable objective metrics for binaural speech enhancement evaluation.

The methods presented in this thesis operate under the assumption that listeners heads remain stationary. However, in real-world scenarios, humans utilize head movements to enhance their ability to localize binaural audio accurately. Incorporating head movement into binaural enhancement methods could significantly improve performance, aligning the models more closely with natural listening behaviours. A major limitation in advancing this research area is the lack of publicly available datasets that account for head movement. Future efforts should focus on developing or acquiring such datasets to enable further exploration and integration of dynamic listener behaviours into binaural speech enhancement models. This advancement could pave the way for more realistic and effective solutions in practical applications.

Bibliography

[1] C. Han, Y. Luo, and N. Mesgarani, "Real-Time Binaural Speech Separation with Preserved Spatial Cues," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020, pp. 6404–6408.

[2] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end Binaural Sound Localisation from the Raw Waveform," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019, pp. 451–455.

[3] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. John Wiley & Sons, Inc., 2010, pp. 269–302.

[4] R. W. Lindeman, H. Noma, and P. G. De Barros, "Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio," in *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE, 2007, pp. 173–176.

[5] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An Introduction to the Speech Enhancement for Augmented Reality (Spear) Challenge," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022, pp. 1–5.

[6] H. Davis, D. Schlundt, K. Bonnet, S. Camarata, F. H. Bess, and B. Hornsby, "Understanding Listening-Related Fatigue: Perspectives of Adults with Hearing Loss," *International Journal of Audiology*, vol. 60, no. 6, pp. 458–468, Jun. 2021.

[7] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: The MIT Press, 1997.

[8] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, pp. 331–342, 2006.

[9] L. Rayleigh, "On our perception of the direction of a source of sound," *Proceedings of the Musical Association*, vol. 2, pp. 75–84, 1875.

[10] M. M. Van Wanrooij and A. J. Van Opstal, "Contribution of head shadow and pinna cues to chronic monaural sound localization," *Journal of Neuroscience*, vol. 24, no. 17, pp. 4163–4171, 2004.

[11] J. Schnupp, I. Nelken, and A. King, *Auditory Neuroscience: Making Sense of Sound*. MIT press, 2011.

[12] D. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. John Wiley & Sons, Inc., 2006.

[13] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls," *J Acoust Soc Am*, vol. 98, no. 2, pp. 988–1007, 1995.

[14] M. Tohyama and A. Suzuki, "Interaural cross-correlation coefficients in stereo-reproduced sound fields," *J Acoust Soc Am*, vol. 85, no. 2, pp. 780–786, 1989.

[15] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *J. Acoust. Soc. Am.*, vol. 127, p. 823, 2010.

[16] T. Brookes, R. Mason, and F. Rumsey, "Integration of measurements of interaural cross-correlation coefficient and interaural time difference within a single model of perceived source width," in *Proc. Audio Eng. Soc. (AES) Conv.*, 2004.

[17] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *J. of Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[18] N. L. Aaronson and W. M. Hartmann, "Interaural coherence for noise bands: Waveforms and envelopes," *J Acoust Soc Am*, vol. 127, no. 3, pp. 1367–1372, 2010.

[19] T. Hirvonen and V. Pulkki, "Interaural coherence estimation with instantaneous ILD," in *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006.* IEEE, 2006, pp. 122–125.

[20] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *J Acoust Soc Am*, vol. 83, no. 4, pp. 1508–1516, 1988.

[21] N. I. Durlach, C. L. Thompson, and H. S. Colburn, "Binaural interaction in impaired listeners: A review of past research," *Audiology*, vol. 20, no. 3, pp. 181–211, 1981.

[22] B. Cornelis, S. Doclo, T. V. dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[23] D. W. Batteau, "The role of the pinna in human localization," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 168, no. 1011, pp. 158–180, 1967.

[24] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J Acoust Soc Am*, vol. 56, no. 6, pp. 1829–1834, 1974.

[25] A. D. Musicant and R. A. Butler, "The influence of pinnae-based spectral cues on sound localization," *J Acoust Soc Am*, vol. 75, no. 4, pp. 1195–1200, 1984.

[26] E. H. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *J Acoust Soc Am*, vol. 112, no. 4, pp. 1583–1596, 2002.

[27] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J Acoust Soc Am*, vol. 85, no. 2, pp. 868–878, 1989.

[28] ——, "Headphone simulation of free-field listening. I: Stimulus synthesis," *J Acoust Soc Am*, vol. 85, no. 2, pp. 858–867, 1989.

[29] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *J Acoust Soc Am*, vol. 61, no. 6, pp. 1567–1576, 1977.

[30] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, pp. 1633–1654, 1999.

[31] T. Van den Bogaert, T. J. Klasen, M. Moonen, L. Van Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 515–526, 2006.

[32] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 484–497, 2008.

[33] T. J. Klasen, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural multi-channel Wiener filtering for hearing aids: Preserving interaural time and level differences," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. V, 2006, pp. 145–148.

[34] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[35] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[36] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Binaural multichannel Wiener filter with directional interference rejection," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015.

[37] ——, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[38] ——, "Comparison of binaural multichannel Wiener filters with binaural cue preservation of the interferer," in *Proc. IEEE Int. Conf. on the Science of Elect. Eng. (ICSEE)*, 2016, pp. 1–5.

[39] N. Gößling, D. Marquardt, and S. Doclo, "Perceptual evaluation of binaural MVDR-based algorithms to preserve the interaural coherence of diffuse noise fields," *Trends Hear.*, vol. 24, 2020.

[40] N. Bisgaard and S. Ruf, "Findings from EuroTrak surveys from 2009 to 2015: Hearing loss prevalence, hearing aid adoption, and benefits of hearing aid use," *American journal of audiology*, vol. 26, no. 3S, pp. 451–461, 2017.

[41] A. Sochenkova and N. Podzharaya, "People with Disabilities and AR/VR Usage Trends: Overview, Estimation, Forecasts," in *2021 10th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 2021, pp. 1–5.

[42] J.-Y. Zeng, Y. Xing, and C.-H. Jin, "The Impact of VR/AR-Based Consumers' Brand Experience on Consumer–Brand Relationships," *Sustainability*, vol. 15, no. 9, p. 7278, 2023.

[43] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing, J. of American Auditory Soc. (AAS)*, vol. 27, pp. 480–492, 2006.

[44] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[45] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[46] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[47] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, pp. 2336–2347, 2009.

[48] D. Wang, "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Springer US, 2005, pp. 181–197.

[49] U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, and D. Wang, "Speech intelligibility of ideal binary masked mixtures," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 1909–1913.

[50] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990.

[51] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proc. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, Oct. 2001, pp. 79–82.

[52] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[53] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[54] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[55] X. Wang, F. Bao, and C. Bao, "IRM estimation based on data field of cochleagram for speech enhancement," *Speech Communication*, vol. 97, pp. 19–31, Mar. 2018.

[56] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101–116, Sep. 2005.

[57] J. Barker, L. Josifovski, M. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition." in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, 2000, pp. 373–376.

[58] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, USA: CRC Press, 2007.

[59] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec. 2014, pp. 577–581.

[60] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, pp. 2604–2612, 2016.

[61] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 708–712.

[62] A. Sugiyama and R. Miyahara, "Phase randomization - A new paradigm for single-channel signal enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2013, pp. 7487–7491.

[63] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, 2016.

[64] D. J. Santry, *Demystifying Deep Learning: An Introduction to the Mathematics of Neural Networks | IEEE eBooks | IEEE Xplore*. Wiley-IEEE Press, 2024. [Online]. Available: https://ieeexplore.ieee.org/book/10357779

[65] D. Xanthidis, C. Manolas, O. K. Xanthidou, and H.-I. Wang, Eds., *Handbook of Computer Programming with Python*. New York: Chapman and Hall/CRC, Dec. 2022.

[66] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958. [Online]. Available: https://psycnet.apa.org/journals/rev/65/6/386/

[67] B. Widrow, J. R. Glover, Jr, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, Jr, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[68] W. Wirtinger, "Zur formalen Theorie der Funktionen von mehr komplexen Vernderlichen," *Math. Ann.*, vol. 97, no. 1, pp. 357–375, Dec. 1927.

[69] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *Proc. IEEE*, vol. 130, no. 1, Parts F and H, pp. 11–16, Feb. 1983.

[70] N. Benvenuto and F. Piazza, "On the complex backpropagation algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 967–969, 1992. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/127967/

[71] H. Leung and S. Haykin, "The complex backpropagation algorithm," *IEEE Transactions on signal processing*, vol. 39, no. 9, pp. 2101–2104, 1991. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/134446/

[72] G. M. Georgiou and C. Koutsougeras, "Complex domain backpropagation," *IEEE transactions on Circuits and systems II: analog and digital signal processing*, vol. 39, no. 5, pp. 330–334, 1992. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/142037/

[73] A. Hirose, "Complex-valued neural networks: The merits and their origins," in *International Joint Conference on Neural Networks*, Jun. 2009, pp. 1237–1244.

[74] A. Hirose and S. Yoshida, "Generalization Characteristics of Complex-Valued Feedforward Neural Networks in Relation to Signal Coherence," *IEEE Trans. Neural Netw.*, vol. 23, no. 4, pp. 541–551, Apr. 2012.

[75] J. Bassey, L. Qian, and X. Li, "A Survey of Complex-Valued Neural Networks," Jan. 2021.

[76] T. Nitta, "N-Dimensional Vector Neuron and Its Application to the N-Bit Parity Problem," in *Complex-Valued Neural Networks*. John Wiley & Sons, Ltd, 2013, ch. 3, pp. 59–74.

[77] C. Lee, H. Hasegawa, and S. Gao, "Complex-Valued Neural Networks: A Comprehensive Survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1406–1426, Aug. 2022.

[78] M. F. Amin and K. Murase, "Learning Algorithms in Complex-Valued Neural Networks using Wirtinger Calculus," in *Complex-Valued Neural Networks*. John Wiley & Sons, Ltd, 2013, ch. 4, pp. 75–102.

[79] R. Savitha, S. Suresh, N. Sundararajan, and P. Saratchandran, "A new learning algorithm with logarithmic performance index for complex-valued neural networks," *Neurocomputing*, vol. 72, no. 16, pp. 3771–3781, Oct. 2009.

[80] J. Deng, N. Sundararajan, and P. Saratchandran, "Complex-Valued Minimal Resource Allocation Network for Nonlinear Signal Processing," *Int. J. Neur. Syst.*, vol. 10, no. 02, pp. 95–106, Apr. 2000.

[81] D. Jianping, N. Sundararajan, and P. Saratchandran, "Communication channel equalization using complex-valued minimal radial basis function neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 687–696, 2002.

[82] M. F. Amin and K. Murase, "Single-layered complex-valued neural network for real-valued classification problems," *Neurocomputing*, vol. 72, no. 4-6, pp. 945–955, 2009.

[83] M. Peker, B. Şen, and D. Delen, "Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm," *Journal of healthcare engineering*, vol. 6, no. 3, 2015.

[84] A. Hirose, "Proposal of fully complex-valued neural networks," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, vol. 4. IEEE, 1992, pp. 152–157.

[85] M. A. Dedmari, S. Conjeti, S. Estrada, P. Ehses, T. Stöcker, and M. Reuter, "Complex Fully Convolutional Neural Networks for MR Image Reconstruction," in *Machine Learning for Medical Image Reconstruction*, F. Knoll, A. Maier, and D. Rueckert, Eds. Cham: Springer International Publishing, 2018, vol. 11074, pp. 30–38.

[86] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Trans. Geosci. Electron.*, vol. 55, no. 12, pp. 7177–7188, 2017.

[87] M.-B. Li, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "Fully complex extreme learning machine," *Neurocomputing*, vol. 68, pp. 306–314, 2005.

[88] S. Scardapane, S. Van Vaerenbergh, A. Hussain, and A. Uncini, "Complex-valued neural networks with nonparametric activation functions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 2, pp. 140–150, 2018.

[89] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, Sep. 2020, pp. 2472–2476.

[90] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-Weighted Self-Attention for Speech Enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020, pp. 6649–6653.

[91] B. Ding, H. Qian, and J. Zhou, "Activation functions and their characteristics in deep neural networks," in *2018 Chinese Control And Decision Conference (CCDC)*, Jun. 2018, pp. 1836–1841.

[92] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, "Noisy activation functions," in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2016, pp. 3059–3068.

[93] A. D. Rasamoelina, F. Adjailia, and P. Sinčák, "A Review of Activation Function for Artificial Neural Network," in *IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Jan. 2020, pp. 281–286.

[94] T. Kim and T. Adalı, "Approximation by fully complex multilayer perceptrons," *Neural computation*, vol. 15, no. 7, pp. 1641–1666, 2003.

[95] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Machine Learning (ICML)*, 2010, pp. 807–814.

[96] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[97] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, and J. Dean, "On rectified linear units for speech processing," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3517–3521.

[98] H. He, T. Yang, and J. Chen, "On time delay estimation from a sparse linear prediction perspective," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 1044–1047, Feb. 2015.

[99] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[100] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[101] L. Trottier, P. Giguere, and B. Chaib-Draa, "Parametric exponential linear unit for deep convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*. IEEE, 2017, pp. 207–214.

[102] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," Feb. 2019. [Online]. Available: https://arxiv.org/abs/1803.08375

[103] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Machine Learning (ICML)*, vol. 30. Atlanta, GA, 2013, p. 3.

[104] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying ReLU and Initialization: Theory and Numerical Examples," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1671–1706, Jun. 2020.

[105] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolutional Network," Nov. 2015. [Online]. Available: http://arxiv.org/abs/1505.00853

[106] Y. Özbay, "A New Approach to Detection of ECG Arrhythmias: Complex Discrete Wavelet Transform Based Complex Valued Artificial Neural Network," *J Med Syst*, vol. 33, no. 6, pp. 435–445, Dec. 2009.

[107] T. Kim and T. Adali, "Fully complex backpropagation for constant envelope signal processing," in *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No. 00TH8501)*, vol. 1. IEEE, 2000, pp. 231–240. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/889414/

[108] K. Tachibana and K. Otsuka, "Wind prediction performance of complex neural network with ReLU activation function," in *Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. IEEE, 2018, pp. 1029–1034.

[109] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2016, pp. 1120–1128.

[110] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep Complex Networks," Feb. 2018.

[111] C. K. Chui and X. Li, "Approximation by ridge functions and neural networks with one hidden layer," *Journal of Approximation Theory*, vol. 70, no. 2, pp. 131–141, 1992.

[112] M. Tanaka, "Weighted sigmoid gate unit for an activation function of deep neural network," *Pattern Recognition Lett.*, vol. 135, pp. 354–359, 2020.

[113] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 2010, pp. 249–256.

[114] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade*, G. Goos, J. Hartmanis, J. Van Leeuwen, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, vol. 1524, pp. 9–50.

[115] S. K. Roy, S. Manna, S. R. Dubey, and B. B. Chaudhuri, "LiSHT: Non-parametric Linearly Scaled Hyperbolic Tangent Activation Function for Neural Networks," in *Computer Vision and Image Processing*, D. Gupta, K. Bhurchandi, S. Murala, B. Raman, and S. Kumar, Eds. Cham: Springer Nature Switzerland, 2023, vol. 1776, pp. 462–476.

[116] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," Feb. 2016. [Online]. Available: http://arxiv.org/abs/1511.07289

[117] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[118] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions," Oct. 2017. [Online]. Available: http://arxiv.org/abs/1710.05941

[119] D. Misra, "Mish: A Self Regularized Non-Monotonic Activation Function," Aug. 2020. [Online]. Available: http://arxiv.org/abs/1908.08681

[120] P. Virtue, X. Y. Stella, and M. Lustig, "Better than real: Complex-valued neural nets for MRI fingerprinting," in *Proc. Int. Conf. Image Process.* IEEE, 2017, pp. 3953–3957.

[121] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2015, pp. 448–456.

[122] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing Overfitting in Deep Networks by Decorrelating Representations," Jun. 2016. [Online]. Available: http://arxiv.org/abs/1511.06068

[123] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[124] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951. [Online]. Available: https://www.jstor.org/stable/2236626

[125] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952. [Online]. Available: https://www.jstor.org/stable/2236690

[126] P. Netrapalli, "Stochastic Gradient Descent and Its Variants in Machine Learning," *J Indian Inst Sci*, vol. 99, no. 2, pp. 201–213, Jun. 2019.

[127] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6d-a separate, adaptive learning rate for each connection," *Slides of lecture neural networks for machine learning*, vol. 1, pp. 1–31, 2012.

[128] M. C. Mukkamala and M. Hein, "Variants of rmsprop and adagrad with logarithmic regret bounds," in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, 2017, pp. 2545–2553.

[129] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[130] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017.

[131] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.

[132] Z. Zhang, "Improved adam optimizer for deep neural networks," in *International Symposium on Quality of Service (IWQoS)*. Ieee, 2018, pp. 1–2.

[133] R. Zaheer and H. Shaziya, "A Study of the Optimization Algorithms in Deep Learning," in *International Conference on Inventive Systems and Control (ICISC)*, Jan. 2019, pp. 536–539.

[134] I. K. M. Jais, A. R. Ismail, and S. Q. Nisa, "Adam optimization algorithm for wide and deep neural network." *Knowl. Eng. Data Sci.*, vol. 2, no. 1, pp. 41–46, 2019.

[135] K.-S. Kim and Y.-S. Choi, "HyAdamC: A new adam-based hybrid optimization algorithm for convolution neural networks," *Sensors*, vol. 21, no. 12, p. 4054, 2021.

[136] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," Dec. 2012.

[137] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks," Feb. 2020. [Online]. Available: http://arxiv.org/abs/1908.06281

[138] D. A. Cirovic, "Feed-forward artificial neural networks: Applications to spectroscopy," *TrAC Trends in Analytical Chemistry*, vol. 16, no. 3, pp. 148–155, 1997.

[139] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[140] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," 2001.

[141] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[142] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[143] Y. Huang, H. Zhang, and Z. Wang, "Multistability of complex-valued recurrent neural networks with real-imaginary-type activation functions," *Applied Mathematics and Computation*, vol. 229, pp. 187–200, 2014.

[144] J. Hu and J. Wang, "Global stability of complex-valued recurrent neural networks with time-delays," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 853–865, 2012.

[145] A. M. Sarroff, V. Shepardson, and M. A. Casey, "Learning Representations Using Complex-Valued Nets," Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.06351

[146] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2017, pp. 0588–0592.

[147] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Netw.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[148] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, 1989.

[149] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. Int. Symp. on Circuits and Syst.*, May 2010, pp. 253–256.

[150] E. K. Cole, J. Y. Cheng, J. M. Pauly, and S. S. Vasanawala, "Analysis of Deep Complex-Valued Convolutional Neural Networks for MRI Reconstruction," May 2020. [Online]. Available: http://arxiv.org/abs/2004.01738

[151] C.-A. Popa, "Complex-valued convolutional neural networks for real-valued image classification," in *International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 816–822.

[152] N. Guberman, "On Complex Valued Convolutional Neural Networks," Feb. 2016. [Online]. Available: http://arxiv.org/abs/1602.09046

[153] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[154] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image Transformer," in *Proc. Int. Conf. Machine Learning (ICML)*. PMLR, Jul. 2018, pp. 4055–4064.

[155] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI open*, vol. 3, pp. 111–132, 2022.

[156] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.

[157] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2020, pp. 5036–5040.

[158] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing Real-time Streaming Transformer Transducer for Speech Recognition on Large-scale Dataset," Feb. 2021.

[159] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 3, Tulsa, OK, USA, Apr. 1978, pp. 586–590.

[160] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 3013–3027, 2011.

[161] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[162] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2013, pp. 3697–3701.

[163] 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 *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[164] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Int. Telecommun. Union (ITU-T), Recommendation P.862, Nov. 2003.

[165] J. G. Beerends, A. Fellow, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, M. Keyhl, and A. Member, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II-Perceptual Model," *J. Audio Eng. Soc. (AES)*, vol. 61, no. 6, 2013.

[166] ITU_T_P863, "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," Int. Telecommun. Union (ITU-T), Standard, Jan. 2011.

[167] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[168] A. H. Andersen, "Speech intelligibility prediction for hearing aid systems," Ph.D. dissertation, Aalborg University, Aug. 2017.

[169] A. M. Gomez, B. Schwerin, and K. Paliwal, "Objective intelligibility prediction of speech by combining correlation and distortion based techniques," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2011.

[170] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[171] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[172] L. Lightburn and M. Brookes, "A Weighted STOI Intelligibility Metric based on Mutual Information," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2016.

[173] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.

[174] ——, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, Sep. 2018.

[175] P. Guiraud, A. H. Moore, R. R. Vos, P. A. Naylor, and M. Brookes, "Using a single-channel reference with the MBSTOI binaural intelligibility metric," *Speech Commun.*, vol. 149, pp. 74–83, Apr. 2023.

[176] "BS.1534-0 : Method for the subjective assessment of intermediate quality levels of coding systems," Int. TeleCommun. Union (ITU-R), Recommendation, Jun. 2001.

[177] "BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," Int. TeleCommun. Union (ITU-R), Recommendation, Oct. 1997.

[178] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in *Web Audio Conference*, 2015, pp. 1–6.

[179] "BS.1534 -2: Method for the subjective assessment of intermediate quality level of audio systems," Int. TeleCommun. Union (ITU-R), Recommendation, 2014. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534

[180] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "The peass toolkit-perceptual evaluation methods for audio source separation," in *Int. Conf. on Latent Variable Analysis and Signal Separation*, 2010.

[181] V. Tokala, M. Brookes, and P. A. Naylor, "Binaural Speech Enhancement Using STOI-optimal Masks," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Germany, Sep. 2022, pp. 1–5.

[182] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[183] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, 2004.

[184] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. on Applied Signal Process.*, vol. 2006, no. 1, pp. 1–14, 2006.

[185] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions," in *Proc. Int. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2006.

[186] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 461–465.

[187] T. Green, G. Hilkhuysen, M. Huckvale, S. Rosen, M. Brookes, A. Moore, P. Naylor, L. Lightburn, and W. Xue, "Speech recognition with a hearing-aid processing scheme combining beamforming with mask-informed speech enhancement," *Trends in Hearing*, vol. 26, Jan. 2022.

[188] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Florence, May 2014.

[189] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[190] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[191] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[192] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 531–539, 2012.

[193] L. Lightburn, E. De Sena, A. H. Moore, P. A. Naylor, and M. Brookes, "Improving the perceptual quality of ideal binary masked speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2017.

[194] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 145, no. 5, pp. 2971–2981, 2019.

[195] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, 2014.

[196] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[197] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," in *Proc. (IEEE) Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 1995, pp. 181–184 vol.1.

[198] L. Lightburn, "Mask-based enhancement of very noisy speech," Ph.D. dissertation, Imperial College London, 2020.

[199] "P. 56: Objective measurement of active speech level," Int. Telecommun. Union (ITU-T), Recommendation, Mar. 1993.

[200] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[201] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[202] L. Lightburn and M. Brookes, "SOBM - a binary mask for noisy speech that optimises an objective intelligibility metric," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 5078–5082.

[203] S. Gonzalez and D. M. Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, pp. 518–530, Feb. 2014.

[204] S. Gonzalez and M. Brookes, "Sibilant speech detection in noise," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Portland, Sep. 2012.

[205] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. on Applied Signal Process.*, vol. 7, pp. 668–675, 2003.

[206] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, p. 1053, 1994.

[207] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise database," TNO Institute for Perception, Tech. Rep., 1988.

[208] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Dallas, Texas, USA, Mar. 2010, pp. 4214–4217.

[209] 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," Linguistic Data Consortium (LDC), Philadelphia, USA, Corpus LDC93S1, 1993.

[210] A. Varga and H. J. M. 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 Commun.*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[211] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-Ear head-related and binaural room impulse responses," *EURASIP J. on Advances in Signal Process.*, vol. 2009, no. 1, p. 298605, Jul. 2009.

[212] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[213] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2006, pp. 1447–1450.

[214] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural Speech Enhancement using Deep Complex Convolutional Recurrent Networks," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, USA, 2023.

[215] ——, "Binaural Speech Enhancement using Deep Complex Convolutional Transformer Networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Seoul, South Korea, 2024.

[216] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Ray Liu, Eds. John Wiley & Sons, Inc., 2008.

[217] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[218] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Sep. 2018.

[219] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, 2018, pp. 3229–3233.

[220] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, 2020, pp. 9458–9465.

[221] V. Tokala, "Listening demo for Binaural Speech Enhancement Using Deep Complex Transformer Networks," 2023. [Online]. Available: https://vikastokala.github.io/bse_dcctn/

[222] P. Manuel, "Mpariente/pytorch_stoi," Feb. 2023. [Online]. Available: https://github.com/mpariente/pytorch_stoi

[223] V. Tokala, 2023. [Online]. Available: https://vikastokala.github.io/bccrn/

[224] J. Francombe, "IoSR Listening Room Multichannel BRIR Dataset - University of Surrey," https://doi.org/10.15126/surreydata.00813511, 2017.

[225] V. Tokala, E. d'Olne, M. Brookes, S. Doclo, Jensen, Jesper, and Naylor, Patrick A., "Multichannel binaural speech enhancement using deep complex convolutional transformer networks," *submitted to IEEE/ACM Trans. Audio, Speech and Language Proc. (TASLP)*, 2025.

[226] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2384–2397, Dec. 2015.

[227] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural Speech Processing with Application to Hearing Devices," in *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, Ltd, 2018, ch. 18, pp. 413–442.

[228] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for binaural noise reduction," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*. Germany: IEEE, 2022.

[229] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 5728–5732.

[230] Y. Yang, S.-F. Shih, H. Erdogan, J. Menjay Lin, C. Lee, Y. Li, G. Sung, and M. Grundmann, "Guided Speech Enhancement Network," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[231] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[232] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[233] A. Moore, P. Naylor, and M. Brookes, "Improving robustness of adaptive beamforming for hearing devices," in *Proc. Int. Symp. on Auditory and Audiological Research. (ISAAR)*, vol. 7, Nyborg, Denmark, Jul. 2019, pp. 305–316.

[234] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 397–401, 1988.

[235] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, ser. Digital Signal Processing, M. Brandstein and D. Ward, Eds. Berlin Heidelberg: Springer-Verlag, 2001, pp. 157–180.

[236] E. Grinstein, E. Tengan, B. Çakmak, T. Dietzen, L. Nunes, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Steered Response Power for Sound Source Localization: A Tutorial Review," *EURASIP J. on Audio, Speech, and Music Process.*, vol. submitted, May 2024.

[237] E. D'Olne, V. W. Neo, and P. A. Naylor, "Latency-Agnostic Speech Enhancement for Wireless Acoustic Sensor Networks Using Polynomial Eigenvalue Decomposition," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Sep. 2024, pp. 235–239.

[238] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[239] E. d'Olne, A. H. Moore, P. A. Naylor, J. Donley, V. Tourbabin, and T. Lunner, "Group Conversations in Noisy Environments (GiN) – Multimedia Recordings for Location-Aware Speech Enhancement," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 374–382, 2024.

[240] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural Localization of Speech in Noise," in *Accepted to Forum Acusticum*, Spain, 2025.

[241] M. D. Good and R. H. Gilkey, "Sound localization in noise: The effect of signal-to-noise ratio," *J Acoust Soc Am*, vol. 99, no. 2, pp. 1108–1117, Feb. 1996.

[242] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in normal-hearing listeners," *J Acoust Soc Am*, vol. 105, no. 3, pp. 1810–1820, Mar. 1999.

[243] M. L. Folkerts, E. M. Picou, and G. C. Stecker, "Spectral weighting functions for localization of complex sound. II. The effect of competing noise," *J Acoust Soc Am*, vol. 154, no. 1, pp. 494–501, Jul. 2023.

[244] N. Kopčo, V. Best, and S. Carlile, "Speech localization in a multitalker mixture," *J Acoust Soc Am*, vol. 127, no. 3, pp. 1450–1457, Mar. 2010.

[245] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[246] T. May, S. van de Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[247] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2122–2131, Nov. 2018.

[248] J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[249] C. V. Pavlovic, "Derivation of primary parameters and procedures for use in speech intelligibility predictions," *J Acoust Soc Am*, vol. 82, no. 2, pp. 413–422, Aug. 1987.

[250] ANSI, "Methods for the calculation of the speech intelligibility index," American National Standards Institute (ANSI), ANSI Standard S3.5-1997 (R2007), 1997.

[251] E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP Method for Universal Robust Multi-Source Tracking," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 19–28, 2024.

[252] V. Tokala, "BIL - Implementation Code," 2025. [Online]. Available: https://github.com/VikasTokala/BiL

[253] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, Sep. 2018.