INVESTIGATIONS ON THE DEEP LEARNING BASED SPEECH ENHANCEMENT ALGORITHMS FOR HEARING-IMPAIRED POPULATION

Zhuohuang Zhang

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Speech, Language, and Hearing Sciences,

and the Luddy School of Informatics, Computing, and Engineering,

Indiana University

June 2022

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Donald Williamson, Ph.D.

Yi Shen, Ph.D.

William Shofner, Ph.D.

David Crandall, Ph.D.

Steven Lulich, Ph.D.

Date of Defense: 05/11/2022

Dedicated to my family and my dear motherland.

ACKNOWLEDGEMENTS

To begin with, I would like to express my sincere gratitude to my advisors, Prof. Yi Shen and Prof. Donald Williamson. It is my great honor to have them as my advisors and I have benefited a lot from them on how to become a strong researcher. This dissertation would not be possible without their persistent support, guidance, and encouragement during my graduate studies. I feel very fortunate and honored to have the opportunity working with them. Prof. Shen's expertise in hearing sciences, statistics and Prof. Williamson's knowledge in signal processing and machine learning have helped me to overcome the many challenges I faced during my research studies. Their dedication, energy and hard-working will continue to inspire me.

I must also thank Prof. William Shofner, Prof. David Crandall, and Prof. Steven Lulich for their time and efforts to serve on my research committee. I learned the fundamentals of auditory anatomy and physiology from Prof. Shofner's class, which cleared many of my questions on human perception of sounds. I also learned many techniques from Prof. Crandall's class on artificial intelligence that broadened my knowledge in machine learning. Prof. Lulich's course on instrumentation has greatly enriched my skills in instrument calibration, data collection and analysis.

I am fortunate enough to spend three summers outside IU for internships in industry. I thank Dr. Hui Song for giving me the opportunity to work on monaural speech enhancement with his signal processing team in Didi Chuxing in 2019. I also thank Dr. Yong Xu for hosting me in Tencent AI Lab in 2020, I really enjoyed our discussions on neural beamforming. I must also thank Dr. Takuya Yoshioka for hosting me in Microsoft in 2021, I gained a lot of knowledge from him on speech separation and front-end speech processing. I deeply appreciate these internship experiences, which have broadened my view on the applications of deep learning speech processing techniques.

I would also like to sincerely thank my colleagues and friends that I met during my

doctoral study. I need to thank Anna Hopkins, Audrey Hiner, Donghyeon Yun, Jillian Bassett, Bailey Henderlong, Yi Liu, Annie Main and Lauren Langley for their help and assistance on subject recruitment for my research projects. I also thank Dr. Xuan Dong for his help and support on building the deep learning based speech assessment systems. I really enjoyed the spare time chats and discussions with my friends and lab members at IU.

Finally, I must express my heartfelt gratitude to my parents, my father Xinguo Zhang and my mother Qi Huang, nothing would be possible without their love and encouragement. I must also thank my wife Shuo Wang for sharing my frustration as well as happiness along this journey. I would never reach this far without her support.

Zhuohuang Zhang

INVESTIGATIONS ON THE DEEP LEARNING BASED SPEECH ENHANCEMENT ALGORITHMS FOR HEARING-IMPAIRED POPULATION

Speech signals are often contaminated by unwanted background noise or interfering speakers that makes daily communication more difficult, especially for hearing-impaired (HI) listeners. Speech enhancement algorithms are thus proposed to alleviate this problem, by removing unwanted sounds. A plethora of these algorithms have been proposed over the years, however, most of them have been optimized and evaluated using only objective measures due to the efficiency of these measures and the cost of alternative methods. These objective measures may not generalize well with subjective responses. Moreover, the underlying mechanisms of the human auditory system for speech quality judgment is not well understood, further investigations are needed.

It is important to investigate the factors that contribute to speech quality perceived by human listeners, to facilitate future development of speech enhancement systems with improved efficiency and user experience. There are several research questions that are worth investigating on the design of speech enhancement algorithms, including (1) How do existing speech enhancement algorithms perform on HI listeners? (2) Could HI listeners perceive the phase distortion contained in a speech signal? And (3) Does phase estimation at different frequency bands contribute equally to perceived speech quality?

Contributions of this dissertation involve the following three aspects. First, we thoroughly investigated the performance of certain speech enhancement algorithms for both simulated HI

and normal hearing (NH) listeners. Second, we conducted a series of listening studies to investigate how human perception is impacted by phase distortions in speech signals. Third, we inspected the band importance of estimating phase on speech quality through several pairwise comparison experiments. We further proposed a novel hybrid speech enhancement framework to efficiently perform phase-aware noise reduction.

Donald Williamson, Ph.D.

Yi Shen, Ph.D.

William Shofner, Ph.D.

David Crandall, Ph.D.

Steven Lulich, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

Speech perception is an important component in daily communication for people. The real world environments, however, usually consist of many unwanted sound sources, including different background noises and sometimes interfering speakers. The human auditory system demonstrates a remarkable ability to separate a target speech signal from interfering sound sources. Unfortunately, the perception of speech in noise can be challenging for people living with hearing impairments as these listeners often struggle with attending to temporal and spectral cues that are important for speech segregation and recognition [1, 2, 3].

It is estimated that 466 million people suffer from some degree of hearing loss, and the World Health Organization estimates that this number will increase to over 900 million over the next 30 years. Although hearing-aid fittings can provide prescribed amplifications in different spectral bands to ensure audibility, fewer than 30% of people who need these devices actually use them. This is mainly because hearing-aids often amplify both speech and noise, which produces low quality and even unintelligible sounds to the users. Therefore, it is crucial to first remove the noise and interfering speech to ensure the functionality of hearing-aid devices.

The goal of speech enhancement is to eliminate the unwanted interference that is contained in a noisy speech signal, to improve speech intelligibility and quality for better human perception. Speech enhancement also serves as an important front-end for many speech communication systems, such as automatic speech recognition (ASR) systems [4, 5, 6, 7, 8, 9, 10], digital hearing-aid devices [11, 12, 13, 14, 15] and voice over Internet pro-

tocol (VoIP) systems [16, 17, 18]. Classic speech enhancement approaches can be divided into two categories (1) Unsupervised approaches that include statistical-based modeling and Wiener filtering [19, 20, 21, 22, 23, 24], and (2) Supervised approaches, such as non-negative matrix factorization (NMF) [25, 26, 27, 28, 29, 30, 31]. A multi-layer perceptron (MLP) was first used in the 1980s [32, 33] for clean speech estimation, but the potential was not fully discovered due to the simple structure of the MLP and a lack of speech materials. Other approaches, such as Gaussian mixture models (GMMs) [34] and support vector machines (SVMs) [35] were also investigated for speech enhancement, however, it has been reported that GMM-based methods cannot generalize well to unseen noises and kernel SVMs cannot scale up for larger datasets [36].

Traditional hearing-aids simply amplify sounds at frequency points that are best for the end user, which is not effective when interfering noise simultaneously occurs. Therefore, modern hearing-aids usually come with noise reduction features, where the system first classifies the environmental noise type, then it estimates the noise power to enhance the target speech signal. Normally, SVMs [37, 38] or GMMs [34, 39] are often adopted for environmental noise classification. Later, noise reduction algorithms are applied, which are usually based on statistical models (e.g., log minimum mean square error (logMMSE) [40, 41]) and Wiener filters [23]. Despite having noise reduction features in modern hearing-aids, the performance is still limited, which often leads to a lack of hearing-aid usage. It presents a challenge to develop better algorithms that produce high quality and more intelligible speech, especially according to HI listeners.

Computational auditory scene analysis (CASA) [42] is a research field that models the process of the human auditory system during speech recognition in noise, where speech extraction is performed by applying a time-frequency (T-F) mask on the noisy speech spectrogram. The short-time Fourier transform (STFT)[1] is used to convert a time-domain signal into the T-F domain, whereas the inverse STFT (iSTFT) can be applied to inverse the

[1]Time-domain speech signals can be transformed into the T-F domain via the STFT, where a short sliding window (e.g., a hamming window) is often used with a specified step size.

transformation. The STFT results in a complex-valued spectrogram that consists of magnitude and phase components. The ideal binary mask (IBM) serves as the primary goal for CASA [43] and is constructed with the premixed speech and noise signals, details will be introduced in Chapter 2. It has been reported that the IBM can greatly improve speech intelligibility in extreme conditions for both normal-hearing (NH) and hearing-impaired (HI) listeners [44]. The IBM-based speech enhancement algorithms have been widely adopted for hearing-aid devices [45, 46, 47], partly because of their efficiency and simplicity, which makes them more suitable for real-time applications. However, it is noteworthy that although the IBM represents the optimal binary mask, it may not necessarily be the optimal target for improving the speech quality, which is an important evaluation criterion for speech enhancement systems.

With the renaissance of deep learning, many data-driven speech enhancement algorithms have been proposed in the past decade [48, 49, 50, 51, 52, 53, 54], and many of them have demonstrated superior performance over conventional signal processing and model based methods. Early data-driven speech enhancement algorithms use fully-connected (FC) feed-forward deep neural networks (DNNs) for T-F mask estimation [49, 50, 55, 56, 57]. More recently, researchers have incorporated recurrent structures into the speech enhancement algorithms to better encode the temporal information of the speech signals. These systems are primarily based on long short-term memory (LSTM) and gated recurrent unit (GRU) networks [4, 58, 59, 60, 61]. Once the estimated speech mask is generated, it is later applied to the noisy speech magnitude spectrogram to extract the magnitude spectrogram of the enhanced speech. Note the original noisy phase is often used to resynthesize the T-F spectrogram back to time-domain enhanced speech, where the IBM and ideal ratio mask (IRM) are often adopted as the learning targets in the early deep learning speech enhancement methods [43, 62, 63, 64, 50, 65]. However, this is not an optimal training strategy as phase contributes significantly to the speech quality and intelligibility. At low frequencies, the phase is associated with the temporal fine structure (TFS) of the speech

signal, which further influences the human perception of talker gender, pitch and intonation [66]. Many recent studies have shown that speech enhancement systems with better phase estimation can lead to improved speech quality and intelligibility for NH listeners [67, 55]. These systems propose to use phase-aware T-F masks, such as a phase-sensitive mask, the complex ideal ratio mask (cIRM) [55, 68, 69, 70], or directly performing spectral mapping in the complex domain [71, 72, 73, 74, 75, 76]. Recent advancements in deep learning based speech enhancement algorithms have provided new opportunities to tackle the speech-in-noise problem for the HI population. For example, in [77], Park et al. proposed a convolutional neural network (CNN) and a deep neural network (DNN) based algorithm for noise classification and reduction in digital hearing-aid devices, respectively. Experimental results demonstrated significant improvements over conventional speech enhancement algorithms. However, it remains unclear how the neural networks should be configured (e.g., network architectures, learning targets) for the best performance for hearing-aid users. Therefore, more efforts should be spent on deep learning based speech enhancement algorithms, especially on its applications for HI listeners.

Besides these speech enhancement algorithms that operate in the T-F domain (i.e., CASA approaches), some other speech enhancement algorithms have been proposed that receive and estimate the time-domain signals in an end-to-end fashion (i.e., time-domain systems) [78, 51, 52], to avoid estimating the magnitude and phase components directly. However, it is difficult to inject auditory system design into these time-domain systems, and the loss of auditory patterns could be potentially harmful to the system's performance [79]. It has been reported that compared to T-F domain systems, time-domain speech enhancement algorithms show an inferior generalization ability [80]. Furthermore, time-domain systems often need larger receptive fields to directly encode the waveform with high temporal resolution, and the end-to-end training scheme with time domain loss functions does not capture the perceptual quality well the and is very sensitive to temporal alignment errors [81].

Meanwhile, beamforming is often adopted in multi-channel speech enhancement systems [82, 83, 84, 85, 86]. Improvements can be observed from multi-channel speech enhancement compared to single-channel systems with spatial information available. Yet single-channel speech enhancement algorithms are more flexible and can be extended or adapted to multi-channel scenarios, whereas multi-channel speech enhancement algorithms often are trained for a specific microphone-array configuration and hence do not generalize well to unseen microphone geometries. This dissertation focuses on T-F domain single-channel speech enhancement algorithms as (1) Compared to time-domain speech enhancement systems, T-F domain algorithms are more flexible for adjustments to meet perceptual needs of HI listeners, and more suitable for perceptually inspired algorithm optimization, (2) Besides serving as a starting point for algorithm development, single-channel speech enhancement algorithms are also more flexible and can be extended to multi-channel scenarios, usually using beamforming approaches.

1.2 Motivations

Approximately one third of the older adult population in the United States live with hearing impairments. Many speech enhancement algorithms have been proposed, however, individuals with hearing impairments are often left out of the development process. Moreover, most of the existing speech enhancement systems have only been evaluated using objective measures designed for healthy young listeners with NH, such as the perceptual evaluation of speech quality (PESQ) [87] for speech quality and the short-time objective intelligibility (STOI) [88] for speech intelligibility. Although these objective metrics provide good correlation results on the systems' performance for NH listeners, it is not clear if the findings from these metrics hold for HI listeners.

Meanwhile, many speech enhancement systems have adopted various frequency scales for data processing, including the conventional linear frequency scale and a Mel frequency scale that simulates the non-linearity of the human auditory system. There lacks a direct

comparison on the performance between speech enhancement systems with these two frequency scales. Hence, its impact on human listeners is not fully understood.

Recent advancements in phase-aware speech enhancement algorithms have demonstrated superior speech quality in the enhanced signals as compared to phase-insensitive peers. Nevertheless, it has only been validated on NH listeners, and it remains unclear whether HI listeners would actually benefit from a phase-sensitive speech enhancement algorithm. HI listeners have poorer sensitivity to the TFS cues [89, 90] and thus benefit less from these cues in speech understanding tasks [91, 92]. Therefore, one might expect that preserving the phase information in speech enhancement algorithms may not lead to the same degree of benefit for HI listeners compared to the NH group. It would be necessary to first evaluate their ability to perceive phase distortion before incorporating any phase-aware speech enhancement algorithm into hearing-aids.

Finally, current phase-aware speech enhancement algorithms try to estimate the phase component for all frequency regions, which makes the system more complicated and require more computational power than phase-insensitive versions. The TFS is strongly correlated with the phase and is less important for perception of high-frequency stimuli. Therefore, estimating phase at higher frequency regions could be redundant and a waste of computational power. It is not clear if the estimated phase contributes equally across frequency regions and what are the most important frequency bands for phase-aware speech enhancement. If such frequency bands exist, using this as prior knowledge could further help design more efficient speech enhancement algorithms in the future.

To summarize, there are still many research questions remain unsolved that how speech enhancement algorithms should be adopted and designed for the HI community, including but not limited to the following aspects:

- It is not clear how existing speech enhancement algorithms perform for HI listeners, as they are often designed for NH listeners with a general purpose. Additionally, results obtained from objective metrics designed for NH listeners may not generalize

to HI listeners.

- Existing T-F domain speech enhancement algorithms often adopt different frequency scales to process the speech signals, including the commonly used linear frequency scale and Mel frequency scale that simulates the non-linearity in the human auditory system. It is unclear how different frequency scales effect of the speech quality perceived by human listeners.

- Phase-aware speech enhancement algorithms have been proposed recently that showed better performance (e.g., speech quality) according to NH listeners, however, it is not clear whether phase-aware speech enhancement algorithms would benefit the HI population with poorer TFS cues.

- Degraded sensitivity to high-frequency components is common for human listeners (including HI listeners), therefore it is not clear if estimating phase components at higher frequencies is beneficial. Understanding the frequency regions where phase-aware speech enhancement is most important could further help better design future speech enhancement algorithms, especially for implementations on low-resource platforms that require small and compatible models.

1.3 Objectives

As discussed in the previous sections, this dissertation primarily aims to address the aforementioned challenges of how speech enhancement algorithms can be better implemented especially for the HI population. The main objectives of this dissertation include the following:

- **Performance of speech enhancement algorithms on HI population:** To better understand how do newly emerged speech enhancement algorithms perform on HI listeners, we conduct systematic examinations on the performance of various speech

enhancement algorithms for both simulated NH and HI listeners. PESQ is used to measure the speech quality received by NH listeners, while the hearing-aid speech quality index (HASQI) [93] is adopted to estimate speech quality perceived by HI listeners.

- **Preference of frequency scales (i.e., linear vs. Mel):** We investigate the influence of different frequency scales on the human perceived speech quality for several data-driven speech enhancement algorithms. A human listening study is conducted for NH listeners on the quality of enhanced speech generated by algorithms operating in linear and Mel scales.

- **Perception of phase distortion in HI listeners:** It is important to know if HI listeners could benefit from phase-aware speech enhancement before incorporating these approaches into hearing-aids. To verify, we conduct several listening studies to investigate whether HI listeners can perceive the phase distortions introduced to the speech signals under different background conditions.

- **Importance of phase estimation at different spectral bands:** It is not well understood whether phase information from different spectral bands contribute equally to the perceived quality. A listening study is conducted to investigate the importance of estimating phase components at different spectral regions. We further develop a novel speech enhancement framework that adopts different strategies handling phase estimation at different spectral regions.

1.4 Thesis Organization

The remaining of this dissertation consists of five chapters. Chapter 2 provides a detailed review of important speech enhancement techniques, including conventional signal processing and model-based speech enhancement algorithms, as well as supervised speech enhancement systems.

In Chapter 3, we implement and compare the performance of several speech enhancement algorithms for both simulated NH and HI listeners. These algorithms include an NMF-based speech separation algorithm [30, 31], two DNN-based speech enhancement systems [50, 55] and another two systems based on recurrent neural networks (RNNs) [94, 58]. The average hearing loss profiles for different listener groups [95] are collected and used as input to HASQI to simulate the speech quality perceived by HI listeners. Furthermore, we conduct a listening study following the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) procedure (recommended in ITU-R BS.1534) [96] to investigate the human preference on different frequency scales (i.e., linear and Mel) for speech processing.

Chapter 4 examines whether HI listeners are able to perceive the phase distortions in speech stimuli, to verify if it is necessary to implement phase-aware speech enhancement algorithms into hearing-aids. Several MUSHRA listening tests are conducted on both NH and HI listeners under different conditions, including noisy and reverberant conditions. Various SNRs and reverberation times are considered, artificial phase distortions are further applied to the speech spectrograms at different degrees. For the above mentioned conditions, we repeat the human evaluations with speech processing from a phase-insensitive speech enhancement algorithm based on ideal ratio mask (IRM) in the T-F domain. This allows us to investigate whether the perceived speech quality by human listeners would be adversely affected if phase distortion remained in the enhanced speech following traditional magnitude-based enhancement.

In Chapter 5, we investigate the influence of estimating phase at different spectral regions on the quality of enhanced speech. Two speech enhancement systems based on variants of convolutional recurrent network [97, 71] are adopted to produce phase-aware and phase-insensitive enhanced speech. Next, we apply low-pass and high-pass filtering on the two versions of enhanced speech to generate speech samples with phase-aware and phase-insensitive components at different cutoff frequencies. A pairwise comparison listening

study is then conducted to determine the frequency boundary on awareness of estimated phase. Based on the experimental findings, we further propose a novel speech enhancement framework with reduced model size that applies different strategies handling phase components at different spectral regions.

Finally, we summarize the contributions of this dissertation and discuss future work in Chapter 6.

CHAPTER 2

PRIOR WORKS ON SPEECH ENHANCEMENT

In this chapter, we provide an overview of some existing speech enhancement methods. Note that speech enhancement and speech separation are two similar concepts, where speech enhancement is commonly defined as the task of removing interfering non-speech noise, while speech separation is used when the interfering sources are speech signals. We first briefly discuss some classical speech enhancement methods, including signal processing, statistical model-based and NMF-based speech enhancement algorithms in Section 2.1. Then, we introduce some newly emerged speech enhancement algorithms in the field of computational auditory scene analysis (CASA) in Section 2.2. We also describe some existing works on time-domain speech enhancement in Section 2.3. Finally, we give a brief discussion on perceptually-motivated algorithms in Section 2.4.

2.1 Classical Speech Enhancement Methods

A plethora of speech enhancement algorithms have been proposed over the last decades to address the long-standing speech in noise problem. Spectral subtraction [98, 99, 100] is perhaps one of the most widely used conventional speech enhancement approaches. It estimates the clean speech by subtracting an estimate of the noise spectrum from the mixture spectrum. In a typical spectral subtraction system, the subtraction is done in the magnitude or power domain, where the noisy phase is used to resynthesize the time-domain enhanced speech. The noise spectrum is often estimated from the initial few frames or during non-speech intervals of the noisy mixture under the assumption of stationary noise. One can easily surmise that spectral subtraction methods could fail when the noise is highly non-stationary. Moreover, the estimated speech spectrum after subtraction can contain negative values that need further post-processing [101, 102].

Wiener filtering is a popular speech enhancement method that operates in the complex domain. It assumes that the speech and noise components are uncorrelated and both have zero means. The optimal Wiener filter, which can be derived by minimizing the mean squared error (MSE) between the clean and estimated speech, is defined as

$$H(w) = \frac{P_s(w)}{P_s(w) + P_n(w)}, \tag{2.1}$$

where $H(w)$ denotes the Wiener filter, w is angular frequency, $P_s(w)$ and $P_n(w)$ represent the power spectrums of the clean speech and noise, respectively. The estimated speech is obtained by applying the Wiener filter to the noisy speech spectrum

$$\hat{S}(w) = H(w)Y(w), \tag{2.2}$$

where $Y(w)$ denotes the noisy speech. Estimating the *a priori* signal-to-noise ratio (SNR) is the key for Wiener filtering [103], where SNR is defined as $P_s(w)/P_n(w)$. Many algorithms have been proposed to calculate the *a priori* SNR from the speech and noise variances [104, 105, 106]. However, many of these noise estimation algorithms assume that the background noise is stationary, which fails for many cases when the assumption does not hold.

Another classical speech enhancement method is statistical model-based speech enhancement. Among them, the minimum MSE (MMSE) estimator [107, 108] is one of the representative approaches that minimizes the error between the estimated and clean speech spectra. It performs speech enhancement based on a speech distribution that is conditioned on a noise observation. It heavily relies on the accurate estimation of the speech and noise components, which is a nontrivial task for non-stationary noise.

Non-negative matrix factorization (NMF) has been widely used as a model based speech separation method [25, 26, 27, 109, 110, 111]. In essence, the NMF factors noisy speech

into the product of a basis matrix and a weight matrix, which can be formulated as

$$Y = BW^T, \tag{2.3}$$

where $B = [B_1, ..., B_n]$ and $W = [W_1^T, ..., W_n^T]$ denote the non-negative basis and weight matrices for n sources (e.g., different speakers and noises), respectively. The basis matrix can be learned from the training data and is later used during inference to re-estimate the non-negative weights that best reconstruct the observation Y. Many studies have been done to improve the performance of NMF-based algorithms by incorporating additional constraints and regulations, however, it has been reported that superior performance can be achieved by DNN-based methods especially when the noise is highly non-stationary [50, 112, 113]. Moreover, NMF is also known to have high computational cost, which limits its applicability for real-time processing.

2.2 Speech Enhancement in CASA

Inspired by the concept of T-F masking in CASA, many recent speech enhancement algorithms have been formulated as supervised approaches that are trained to estimate different oracle T-F masks. These approaches have substantially benefited from the renaissance of deep learning. Wang and Wang [114] first incorporated a DNN into the speech separation tasks and their experimental results demonstrated substantial improvements over conventional non deep-learning approaches. By training with a dataset that consists of diverse speakers and noise, the DNN-based speech enhancement system comes with much better generalizability compared to conventional SVM based methods. Improvements in speech intelligibility for both NH and HI listeners have also been reported in [44]. Since then, deep learning based approaches have become very popular with newly emerged speech enhancement systems and substantial progress has been made on further improving the enhancement performance and generalization ability. It has become the state-of-the-art for

noise reduction, and this has been demonstrated in several noisy environments [36, 115].

For a typical T-F domain deep learning based speech enhancement system, the speech waveform is first transformed into a T-F representation (i.e., spectrogram) as the input feature for the neural networks. Early works have used the IBM as the training target due to its simplicity, however, deep learning based speech separation is not limited to binary masking and many different types of T-F mask (or direct spectral mapping) have been extensively investigated over the past few years. This includes the ideal ratio mask (IRM), complex ideal ratio mask (cIRM), and complex spectral mapping, to name a few [50, 55, 94, 58, 72]. In the case of T-F masking, the estimated T-F mask is applied on the noisy speech spectrogram to generate the spectrogram of the enhanced speech. Whereas spectral mapping directly estimates the enhanced spectrogram without estimating the T-F mask. Eventually, the enhanced spectrogram is resynthesized into speech waveform using the iSTFT.

T-F domain speech enhancement systems can be further divided into two sub-categories: *phase-insensitive* and *phase-sensitive* methods, based on how they handle the phase component of the enhanced speech. Phase is a critical component of speech signal that contributes to speech intelligibility and perceived speech quality. Many recent studies have shown the benefits of estimating phase to speech enhancement [67, 55, 68]. In the following subsections, I will introduce some commonly used phase-insensitive and phase-aware speech enhancement methods.

2.2.1 Phase-insensitive algorithms

Figure 2.1 depicts a high-level framework of a T-F masking based phase-insensitive speech enhancement system, which involves a training stage (top) and a testing stage (bottom). In the training stage, the network first takes in input features (e.g., mixture magnitude spectrogram) and is trained to estimate a learning target, typically an oracle T-F mask. In essence, a T-F mask describes the dominance of the speech component at each T-F bin.

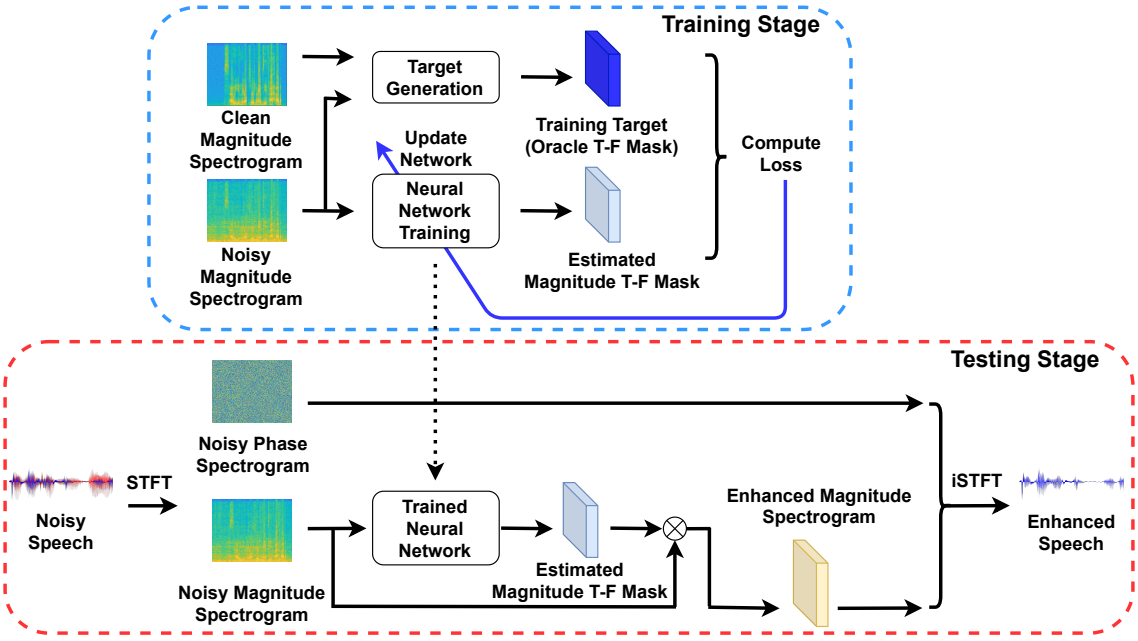

Figure 2.1: Illustration on the training and testing stages of a typical T-F masking based phase-insensitive speech enhancement system.

Given the input features, the neural network generates a prediction of the T-F mask, where a loss function will then compute the distance (e.g., MSE) between the estimate and the oracle target. The distance metric is used to adjust the weights in the neural network (e.g., through the backward propagation algorithm). By repeating the above process for a large number of stimuli and their corresponding oracle T-F masks, the weights are trained so that the network is able to generate close predictions of the target masks even for unseen stimuli. The weights are held frozen during the testing stage. During testing, the time-domain speech signal is first transformed into a spectrogram and fed into the network to generate an estimated T-F mask. The enhanced magnitude spectrogram is then derived by applying the estimated T-F mask to the noisy magnitude spectrogram. In the last step, the enhanced magnitude spectrogram together with the original noisy phase are used to resynthesize the time-domain enhanced speech using inverse short-time Fourier transform (iSTFT).

A visualization of magnitude spectrograms of the noisy and clean speech signals, as

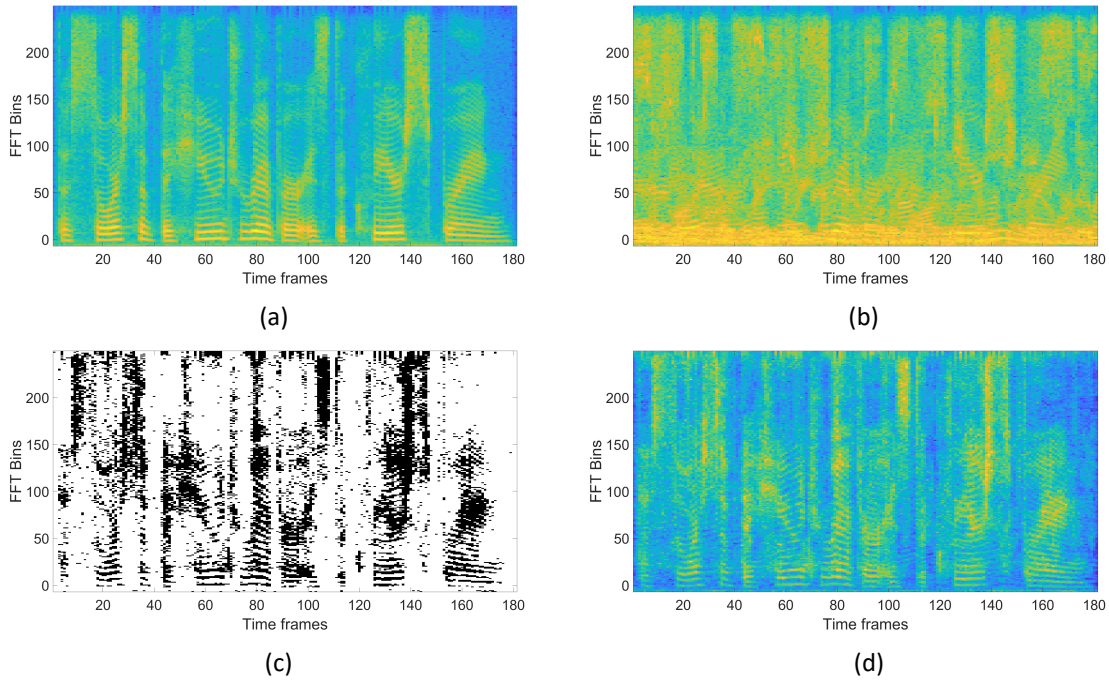

Figure 2.2: Example T-F representations of: (a) Magnitude spectrogram of a clean speech signal, (b) Magnitude spectrogram of a noisy speech signal (multi-talker babble noise at -5 dB SNR), (c) IBM and (d) IRM.

well as some commonly used phase-insensitive T-F mask targets, are shown in Figure 2.2. In a spectrogram, the temporal information is presented along the horizontal axis and the frequency information is displayed along the vertical axis (i.e., low frequencies at bottom and high frequencies at top). The magnitude spectrograms of a clean and a noisy speech signal (babble noise at -5 dB SNR) are provided in Figure 2.2 (a) and (b), respectively. Figure 2.2 (c) and (d) illustrate the corresponding IBM and IRM.

In the supervised learning scheme, the ground truth information of the noise and clean speech is available during training stage, where different oracle T-F masks can be constructed as learning targets. IBM estimation is used as a primary signal processing technique in CASA algorithms to extract the speech component. For each T-F bin, if the SNR is greater or equal to a predefined local criterion (LC), then it is considered as speech dominant and the bin is assigned a value of "1". Likewise, if it is noise dominant (i.e., the SNR is less than the specified LC value), a value of "0" is assigned to that T-F bin. The IBM is

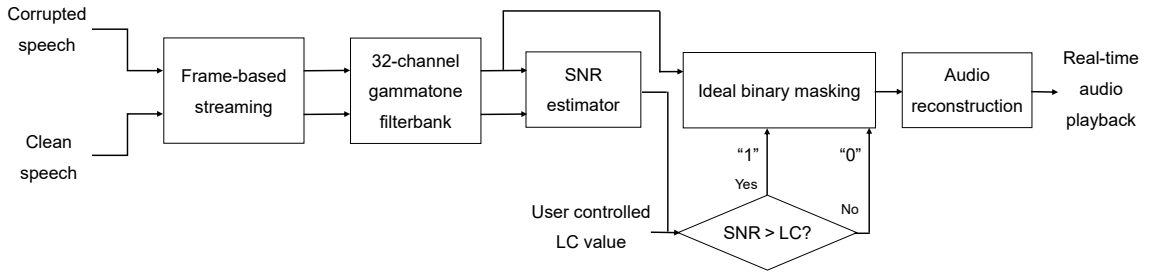

Figure 2.3: Block diagram of the procedure for real-time IBM processing described in [122]. This figure is borrowed from [122].

formulated as

$$IBM(t,f) = \begin{cases} 1, & SNR(t,f) \geq LC \\ 0, & otherwise \end{cases}, \tag{2.4}$$

where (t, f) represents the time frame and frequency indices, respectively. Resynthesizing a signal using only the T-F bins with favorable SNRs, the IBM-based algorithm hence performs speech glimpsing for the listener [116, 117]. It has been shown that IBM-based speech enhancement algorithms can achieve reasonably good intelligibility performance under both anechoic and reverberant conditions [118, 34, 119], where improved automatic speech recognition (ASR) accuracy is also reported in [120].

Because of its simplicity and efficiency, the IBM has been widely adopted in many hearing-aid devices for noise reduction. Previous studies have indicated that speech intelligibility is largely independent of LC for a wide range of LC values (typically between -12 and 0 dB) [121]. This lack of dependency on LC may result from the ceiling effect, since the intelligibility of IBM processed sentences often reaches 100-% correct recognition. The LC value is often arbitrarily chosen (e.g., 0 or -5 dB as in [43, 35]), since it has been believed that it will not significantly affect the intelligibility. Nevertheless, it is not yet clear whether human listeners have preference for certain LC values when listening to IBM-processed speech, even when the speech intelligibility is at the ceiling.

To address this question, our previous work [122] investigated the human preference on LC value for IBM-processed speech. In the listening study, everyday sentences were

mixed with different types of background noises and presented continuously to the listeners following IBM processing. Note the IBM algorithm was implemented so that the listeners were able to adjust the LC value in real-time using a programmable knob device. The procedure of real-time IBM processing is illustrated in Figure 2.3. The clean and corrupted speech signals in each frame are first passed through a gammatone filterbank. Then, the SNR for the n-th frame is estimated based on the corresponding root-mean-square (RMS) amplitudes of the clean and corrupted speech in each frequency channel. Given the LC value provided by the user, the IBM is constructed and applied to generate enhanced speech in real-time.

Experimental results suggest that the preferred LC value exhibits large individual differences and approximately half of the listeners show consistent patterns of results. The preferred LC value is higher for speech materials with less contextual cues and higher background SNRs. Results indicate that the effects of LC on listeners preference may depend on additional characteristics of the background noise besides the overall SNR. The results are also found to match previous findings that LC scales with overall SNR at a rate of 1 dB/dB and a LC value that is 5 dB below the overall SNR is recommended [123, 50].

Unlike the IBM, which performs binary classification on the mixture spectrogram, the IRM consists of continuous weights (i.e., a regression problem) for each T-F bin that could produce better speech quality [50]. As illustrated in Figure 2.2 (d), an IRM is typically derived as the ratio between the speech energy and noisy mixture energy, which can be formulated as

$$\begin{aligned} IRM(t,f) &= \left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\alpha} \\ &= \left(\frac{SNR(t,f)}{SNR(t,f)+1}\right)^{\alpha}, \end{aligned} \tag{2.5}$$

where α is a tunable scaling parameter that is typically set to 0.5 [50]. Although the IRM for CASA speech processing is different from Wiener filtering, the definition of the IRM is closely related to the frequency-domain Wiener filter as described in Eq. (2.1).

Various speech enhancement systems with different network architectures have been

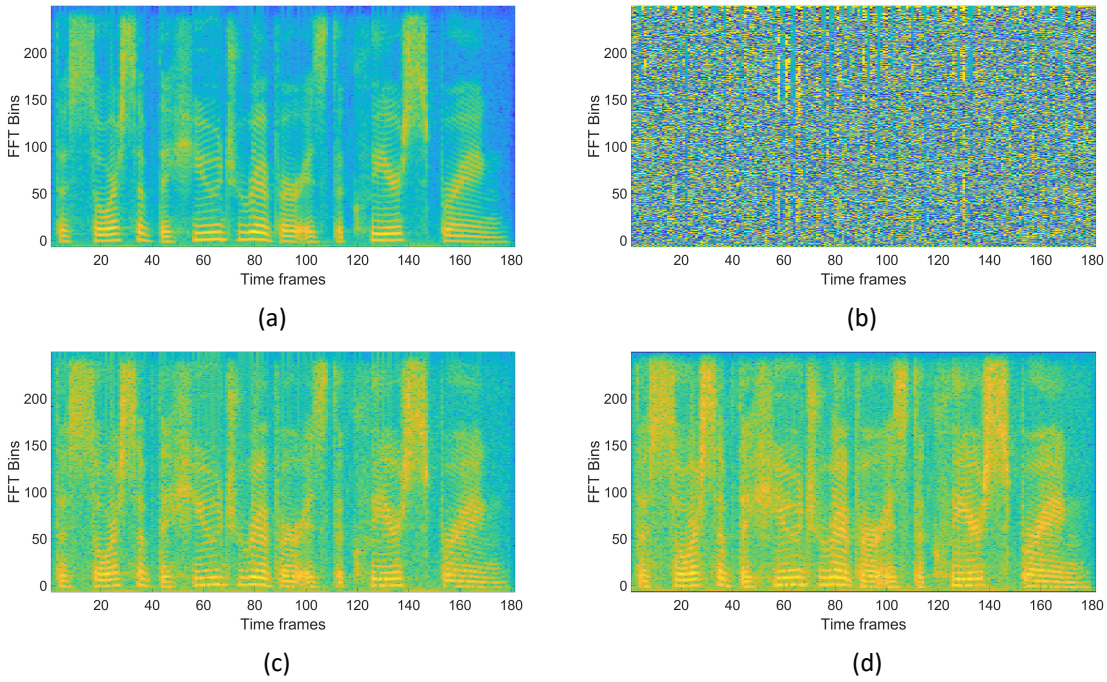

Figure 2.4: Visualization of: (a) Magnitude spectrogram, (b) Phase spectrogram, (c) Real part of the spectrogram and (d) Imaginary part of the spectrogram.

proposed to estimate the IRM, including early approaches based on DNNs [115, 49, 50, 55, 124, 56], and more recent approaches based on RNNs (e.g., LSTM) [4, 59, 125, 126, 60, 127] and CNNs [128, 129, 130, 131]. These methods often adopt the MSE between the network estimate and the oracle training target as the loss function to update the systems. It is also worth pointing out that some early work in CASA used hand-crafted features as system inputs [50, 55, 68]. These features mainly include Mel frequency cepstral coefficient (MFCC) [132], gammatone frequency cepstral coefficient (GFCC) [133], perceptual linear prediction (PLP) [132], and relative spectral transform PLP (RASTA-PLP) [134]. However, due to the powerful learning and representation abilities of data-driven deep learning models, more recent speech enhancement approaches often let the neural networks learn the feature representation from the raw input data (e.g., a spectrogram) during training.

2.2.2 Phase-sensitive algorithms

Phase-aware speech enhancement algorithms not only estimate the magnitude component of speech but also the phase component. Example complex-domain spectrograms (i.e., magnitude/phase and real/imaginary pairs) are shown in Figure 2.4. In the phase spectrogram (i.e., (b) in Figure 2.4), little structure is observed and it could be intractable for a neural network to directly estimate the phase component. The real and imaginary spectrograms, on the other hand, contain similar temporal and spectral structures with the magnitude spectrogram (i.e., (c) and (d) with (a) in Figure 2.4). In practice therefore, many phase-aware speech enhancement algorithms estimate the real and imaginary part of the spectrogram to implicitly handle the phase component [55, 68, 69, 72, 73, 74].

As described in [55, 68], the relationship between the complex representation of the T-F domain noisy speech, Y, clean speech, S, and cIRM, M, can be formulated as

$$\begin{aligned} S &= M \times Y \\ S_r + iS_i &= (M_r + iM_i) \times (Y_r + iY_i), \end{aligned} \tag{2.6}$$

where the subscripts r and i denote the real and imaginary parts, respectively. i is the complex number and $\times$ denotes the element-wise complex multiplication. For display purposes, (t, f) are not shown, but it is implied that the operations are performed at each T-F point. The real and imaginary parts of the cIRM are defined as

$$\begin{aligned} M_r &= \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2}, \\ M_i &= \frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2}. \end{aligned} \tag{2.7}$$

In addition to cIRM for phase-aware speech enhancement, direct spectral mapping in the complex domain has also been explored [135, 71, 72]. In such case, the network directly estimates the real and imaginary parts of the speech spectrogram instead of performing T-F masking. Over recent years, many phase-aware T-F domain speech enhancement systems

have been proposed, including deep complex U-Net [69], deep complex convolution recurrent network (DCCRN) [70], and phase-and-harmonics-aware speech enhancement network (PHASEN) [73]. Better performance can be achieved compared to phase-insensitive ones in terms of speech quality and intelligibility [68, 135, 71].

However, in order to perform speech enhancement in the complex T-F domain, the system is often required to deal with both real and imaginary parts which leads to increased network complexity compared to phase-insensitive methods. An example of a complex network structure for phase-aware speech enhancement is provided in Figure 2.5.

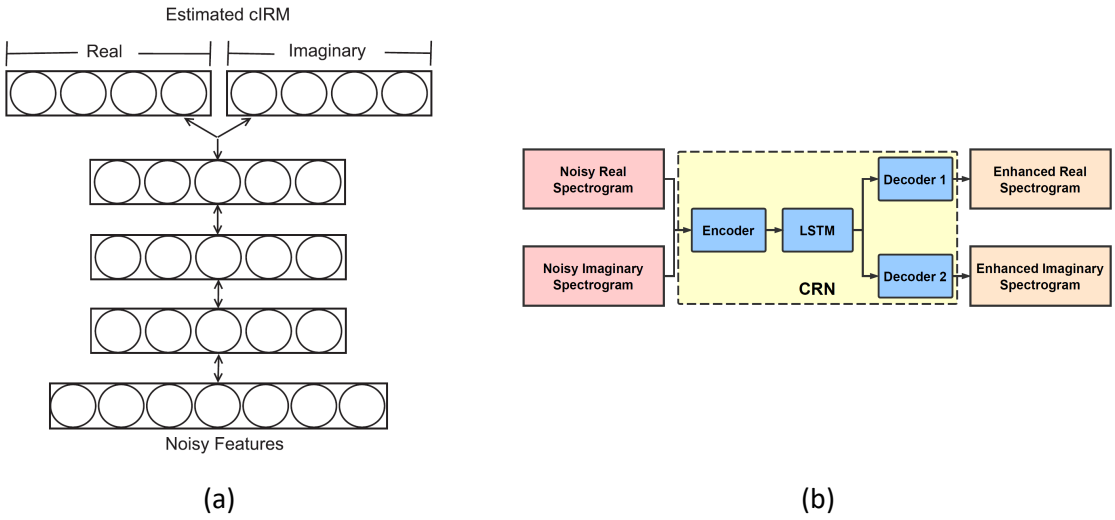

Figure 2.5: Examples of T-F domain phase-aware speech enhancement networks: (a) DNN for cIRM estimation, figure borrowed from [55], (b) Complex convolutional recurrent network (CCRN) for complex spectral mapping, figure borrowed from [71].

As one may notice here, the phase-aware speech enhancement network often consists of additional sub-networks (or sub-layers) to deal with implicit phase estimation (i.e., real and imaginary parts). Therefore, it may hinder the applicability of these phase-aware speech enhancement systems for mobile devices such as hearing-aids despite their better performance. Previous studies have investigated using perceptually weighted loss or metrics to update the neural networks to achieve better performance [136, 137, 138, 139]. However, not much effort has been spent on the physiologically inspired model for improved effi-

ciency (e.g., reduced network size). In Chapter 5, we will provide some promising future directions to reduce the network complexity for phase-aware speech enhancement algorithms.

2.3 Time-domain speech enhancement

Many time-domain (i.e., end-to-end) speech enhancement systems have been proposed recently, for example, the Wave-U-Net [78], TasNet [51] and Conv-TasNet [52]. These end-to-end systems replace the conventional STFT and iSTFT signal processing procedures with a learnable neural network based encoder-decoder-like structure (e.g., 1-D CNN), to implicitly perform the feature extraction. In [51], the encoded latent features are then adjusted by a learned-latent mask and are later fed to the decoder for time-domain signal reconstruction. These systems have been reported to achieve good performance in terms of speech quality on different speech separation tasks. The schematic diagram of a typical time-domain speech enhancement system (i.e., TasNet [51] illustrated here) is provided in Figure 2.6.

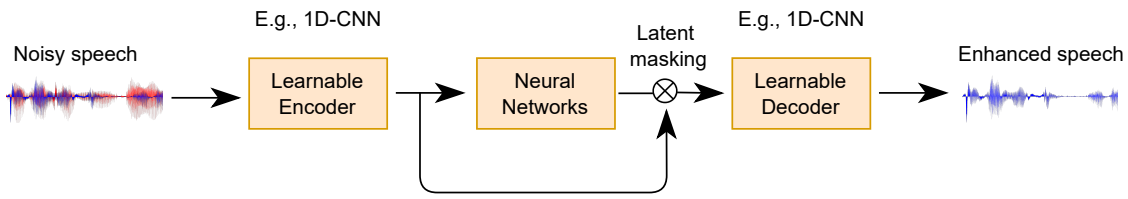

Figure 2.6: An illustration of TasNet, a time-domain speech enhancement network.

As discussed in the previous chapter, although time-domain speech enhancement systems feature a more simplified pipeline by replacing some conventional operations such as STFT/iSTFT and T-F masking with neural networks and latent feature operations, they often feature inferior generalization ability compared to CASA approaches [80]. To directly encode the waveform with much higher temporal solution compared to a spectrogram, time-domain approaches generally require larger receptive field compared to CASA

approaches, and that time-domain loss (e.g., L1 and L2 loss) is sensitive to alignment errors and often fails to capture the contextual information [81]. Moreover, the implicit nature of time-domain speech enhancement systems makes it intractable to incorporate prior knowledge on auditory systems [79]. Thus, this dissertation will be focusing on CASA approaches.

2.4 Perceptually-motivated speech enhancement

Most of the existing deep learning speech enhancement methods use the MSE distance (either in the T-F or time-domain) between the estimated term and oracle target to optimize the model parameters. This design, however, rarely considers the perceptual factors that contribute to human speech perception and leaves future space for improvements. Some newly proposed speech enhancement algorithms feature a perceptually-motivated design that demonstrate improvements over conventional approaches [140, 137, 141, 142]. Usually, it is done by incorporating a perceptually weighted loss or enforcing a learning scheme to optimize an objective quality measure, as it is believed they are more closely correlated with human perception of sounds. Furthermore, Valin et al. [142] suggests that a model focusing on the spectral envelope and periodicity is able to generate high quality enhanced speech in real time.

It is worth pointing out that most of the existing speech enhancement algorithms serve the interests of the large NH population (e.g., front-end processing in online meetings), whereas the HI individuals who are in more urgent need are often ignored when designing such algorithms. Designing a specific or specially tuned speech enhancement algorithm for HI population is an important future direction of algorithm development. Besides taking perceptual factors from NH population into account during algorithm development, factors on how HI population differs from the NH population in speech perception and quality judgment should also be considered when designing HI-specific algorithms.

It is important to understand how speech enhancement algorithms could be better im-

plemented for human listeners including HI populations. As discussed before, this dissertation will investigate this topic from different aspects, including the performance of different speech enhancement algorithms, frequency scales for speech processing, phase awareness in HI listeners and the importance of phase estimation across spectral regions to the perceived speech quality.

CHAPTER 3

PERFORMANCE OF SPEECH ENHANCEMENT ALGORITHMS FOR HEARING-IMPAIRED LISTENERS

The goal of speech enhancement is to remove unwanted interfering noises and speakers from the noisy speech to improve the speech quality and intelligibility. It can be applied as a front-end for many different communication systems such as ASR, mobile phones and digital hearing-aid devices. However, most of them have not been evaluated for HI listeners, therefore, it is not clear how do they perform on this population that is most affected by the speech-in-noise problem.

In this chapter, we first systematically compare the performance of different speech enhancement algorithms for simulated NH and HI listeners, by using two objective speech evaluation metrics (i.e., PESQ [87] and HASQI [93]). The speech enhancement algorithms are trained on a diverse dataset with a broad range of SNRs and noises. We also investigate the impact of the data's processing frequency scale (Mel vs. linear), since there is a trend that many recent studies were using audio features in a non-linear frequency scale (i.e., Mel scale) rather than in the original linear frequency scale. The simulation results reveal that a LSTM-based speech enhancement system with data in the Mel frequency scale achieves the best performance in PESQ, while another bidirectional LSTM-based network achieves the best performance in HI settings with linear frequency scale for data processing. In general, the Mel frequency scale leads to improved PESQ scores, but reduced HASQI scores.

The human preference on the frequency scale for data processing is investigated next. We conduct a human listening study on NH listeners to collect the perceived speech quality on enhanced speech samples produced by speech enhancement systems implemented in two frequency scales, following a MUSHRA procedure (recommended in ITU-R BS.1534) [96]. Results show that human listener prefers the enhanced speech in linear frequency

scale, which contradicts PESQ predictions and is consistent with the results obtained from HASQI index.

The objective comparison between speech enhancement algorithms for both simulated NH and HI listeners is presented in Section 3.1. Section 3.2 describes the human listening study and reports the results we collect on the human preference on speech processing frequency scale. The partial work presented in this chapter has been published in the Proceedings of 2019 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP) [143] and Proceedings of 23rd International Congress on Acoustics (ICA) [144].

3.1 Objective Comparison of Speech Enhancement Algorithms

3.1.1 Introduction and motivation

Speech degradation in the presence of noise is a common problem for individuals, especially for people with hearing impairments [145]. A plethora of speech enhancement algorithms have been proposed for improving speech intelligibility and perceived speech quality in noisy environments, including those based on statistical-model [40], non-negative matrix factorization (NMF) [146, 30, 31, 147], deep neural networks (DNNs), and recurrent neural networks (RNNs) [50, 55, 94, 58]. Recent DNN- and RNN-based speech enhancement algorithms have shown superior performance over traditional approaches, but there has been a lack of parallel comparisons amongst these algorithms. A previous study by Hu and Loizou [41] provides an overview of the relative performance of different speech enhancement algorithms, but many of the algorithms are no longer considered, largely due to the rise of deep learning approaches. In many recent studies, speech materials and background noises are limited, and only a narrow range of signal-to-noise ratios (SNRs) are used. One goal of this section is to compare some traditional and newly-emerged deep learning speech enhancement algorithms, using a large database that contains diverse mixtures of speech and background noise under a broad range of SNRs.

A second goal of this section is to evaluate the performance of these algorithms for

people with hearing impairments. Hearing loss affects tens of millions of individuals in the United States [148], and it is highly prevalent among older adults [95]. An example of the prevalence of hearing loss by age is shown in Figure 3.1. It can be observed that as age increases (e.g., 50-59 to 80+), the percentage of population affected by hearing impairment gets significantly higher (e.g., 16% to 88%).

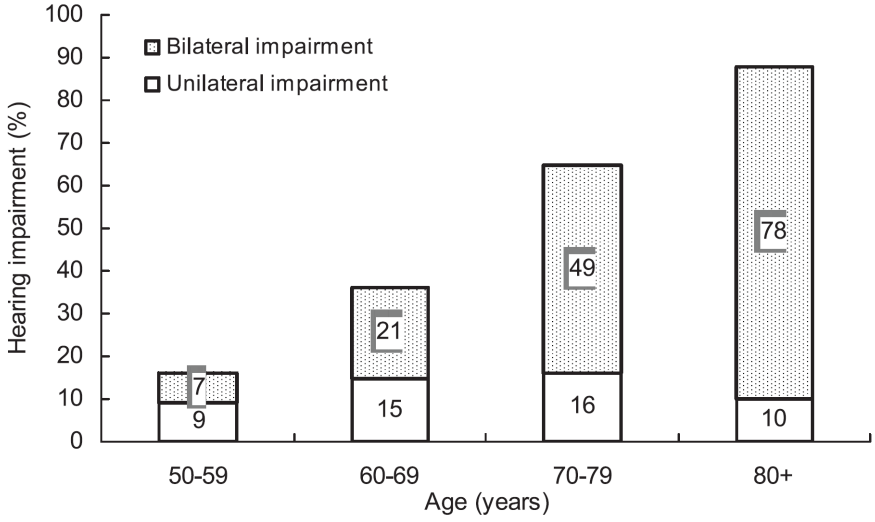

Figure 3.1: Prevalence of hearing loss (bilateral and unilateral) by age. This figure is borrowed from [149].

Approximately one in three people in the United States between 65 and 74 years of age live with a hearing impairment. In contrast, most previous studies evaluated speech enhancement outcomes using metrics developed for healthy young adults, such as the widely-adopted perceptual evaluation of speech quality (PESQ) [87]. It is not clear whether the findings using these metrics hold for the HI population. It is crucial to evaluate speech enhancement algorithms for HI listeners in order to generalize laboratory results to real-world applications involving listeners of all ages. This work includes evaluations using the hearing-aid speech quality index (HASQI) [93], to quantify how speech enhancement algorithms impact those with hearing loss. It is important to note that HASQI's computations resemble the processing performed by normal and impaired auditory systems. Compared to PESQ, HASQI is more adept at predicting the perceived speech quality ratings that are

provided by hearing-impaired listeners that use hearing aids [150, 151].

In this section, the performance of various speech enhancement algorithms are compared using both PESQ and HASQI. These algorithms include one NMF-based approach [30] that serves as a strong baseline model, two DNN-based approaches [50, 55] that perform well for normal-hearing listeners, and two RNN-based approaches [94, 58], which determines the impact of recurrent structures on speech enhancement for the hearing impaired listeners. Meanwhile, the impact of the frequency scale (Mel versus linear) on the input and output data is also investigated in our study, since previous studies have used different frequency scales without performing direct comparisons. Its impact on normal-hearing and hearing-impaired listeners is not fully understood. The evaluations of speech quality using HASQI are separately conducted for different genders and age groups according to the typical trajectories of age-related hearing loss for female and male listeners.

This section is organized as follows. Section 3.1.2 introduces the speech enhancement algorithms that are investigated. A detailed experimental setup including speech materials and hearing loss profiles is given in Section 3.1.3. Simulation results for NH and HI listeners are provided in Section 3.1.4. Section 3.1.5. provides a summary on the findings for this section.

3.1.2 Speech enhancement algorithms

Active-set Newton algorithm

NMF is an efficient method for extracting target signals from mixtures and it is widely used for speech enhancement and other applications [147]. As an extension of NMF, the active-set Newton algorithm (ASNA) [30, 31] can be expressed as $\hat{s} = BW^T$, where $\hat{s}$ is the estimated target speech signal, B here denotes the trained speech dictionary and W represents the activation weights. ASNA applies the Newton method to update the weights more efficiently than other NMF approaches, and it has been shown to outperform them in different environments. It will serve as the baseline model for this work. The number of

speech and noise exemplars are set to 1000 and 300, respectively.

DNN-based ideal ratio mask estimation

A DNN-based method that estimates the ideal ratio mask (IRM) in the time-frequency (T-F) domain is included for this work, since it shows performance advantages over other DNN-based training targets [50]. The IRM is defined as

$$M^{rm}(t,f) = (\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2})^{\frac{1}{2}} \tag{3.1}$$

where $|S|$ represents the magnitude spectrogram of the clean speech signal and $|N|$ is the magnitude spectrogram of the noise signal at time frame t and frequency index f.

As mentioned in [50], this DNN-IRM network has three fully-connected layers with 1024 units each. The rectified linear (ReLU) [152] activation function is applied to the hidden layers and a sigmoid activation function is used for the output layer. The ReLU and sigmoid activation functions are defined as

$$\begin{aligned} \sigma_{ReLU}(x) &= max(0,x), \\ \sigma_s(x) &= \frac{1}{1+e^{-x}}. \end{aligned} \tag{3.2}$$

A set of complementary features [50] are used as the input to the network, including amplitude modulation spectrogram (AMS), relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP), Mel frequency cepstral coefficients (MFCC) and power spectra derived from a 64-channel Gammatone filterbank. An auto-regressive moving average filter [153] is further applied to smooth the extracted features temporally

$$\overline{C}(t) = \frac{\overline{C}(t-m) + \cdots + C(t) + \cdots + C(t+m)}{2m+1}, \tag{3.3}$$

where $\overline{C}(t)$ denotes the filtered feature vector at time frame t, C is the feature vector, m

is the filter order, which is set to 2. Additionally, the window size is set to 40 ms with a step size of 20 ms. We use adaptive gradient descent (AdaGrad) as the optimizer, with a mini-batch size of 512, and maximum training epoch number of 80. The mean square error between the estimated IRM and oracle IRM is used as the loss function.

Note that the estimated IRM is in the linear frequency scale, to investigate the impact of frequency scales (linear or Mel), a Mel frequency (with 100-bin) domain implementation is also investigated in this work. A signal can be converted between the linear and Mel frequency scales using the following transformations

$$|S^{Mel}(t,f)| = B|s(t,f)|, \quad |S^{iMel}(t,f)| = B^T|S^{Mel}(t,f)| \tag{3.4}$$

where $|S^{Mel}(t,f)|$ is the Mel-scale T-F domain signal's magnitude and $|S^{iMel}(t,f)|$ is the linear-scale T-F domain signal's magnitude after an inverse-Mel transformation. B here represents a matrix of weights to combine short-time Fourier transform (STFT) bins into Mel bins, and B^T represents the transpose of B. Note that Mel-transformation is a lossy process, so some information from the original spectrogram will be lost during reconstruction.

DNN-based complex ideal ratio mask estimation

The authors in [55] propose a network that estimates the complex ideal ratio mask (cIRM) in the T-F domain. This enables the DNN to predict the phase response in addition to the magnitude response. We include this method, since it is shown to outperform other training targets in objective and subjective evaluations, but the importance of phase for the hearing impaired is not well understood. The cIRM is defined as

$$\begin{aligned} M^{crm}(t,f) &= \frac{|S(t,f)|}{|Y(t,f)|}\cos(\theta(t,f)) + i\frac{|S(t,f)|}{|Y(t,f)|}\sin(\theta(t,f)), \\ \theta(t,f) &= \theta^S(t,f) - \theta^Y(t,f), \end{aligned} \tag{3.5}$$

where $|Y(t,f)|$ represents the magnitude spectrogram of the noisy speech, i indicates the imaginary number, and $\theta(t,f)$ is the phase difference between the clean speech and noisy speech. The cIRM is predicted with a network that consists of three hidden layers with 1024 units each. All hidden layers use ReLU activation functions. The two output layers use linear activation functions. Other parameters for this network are the same as the ones for the DNN-IRM. A Mel frequency domain implementation is also included, which has not been previously done for the cIRM.

LSTM-based ideal ratio mask estimation

Long short-term memory (LSTM) is a special type of RNN, which solves the problem of exploding and vanishing gradient of traditional RNNs [154]. The recurrent structure within LSTM networks makes it powerful in time series prediction, such as problems dealing with speech recognition, and speech enhancement. An LSTM unit consists of a memory cell, an input gate, an output gate and a forget gate, which can be formulated as [154]

$$\begin{aligned} f(t) &= \sigma_s(W_f x(t) + U_f h(t-1) + b_f), \\ i(t) &= \sigma_s(W_i x(t) + U_i h(t-1) + b_i), \\ o(t) &= \sigma_s(W_o x(t) + U_o h(t-1) + b_o), \\ \tilde{c}(t) &= \sigma_g(W_c x(t) + U_c h(t-1) + b_c), \\ c(t) &= f(t) \circ c(t-1) + i(t) \circ \tilde{c}(t), \\ h(t) &= o(t) \circ \sigma_g(c(t)), \end{aligned} \tag{3.6}$$

where $x(t)$, $f(t)$, $i(t)$, $o(t)$, $\tilde{c}(t)$, $c(t)$ and $h(t)$ represent the input vector, forget gate's activation vector, input gate's activation vector, output gate's activation vector, memory cell input activation vector, memory cell state vector and hidden state vector, respectively. $\circ$ is the Hadamard product, W and U denote the weights of the input and recurrent connections.

b is the bias term. σ_g is the tangent activation function, which is defined as

$$\sigma_g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{3.7}$$

We implement the LSTM network architecture described in [94], since it shows impressive performance on speech enhancement tasks. The network has two LSTM layers with 256 nodes in each layer, followed by a third fully-connected layer with sigmoid activation function. It takes 100-bin log-Mel magnitude spectrograms as input and predicts an IRM for the clean speech signal in the Mel scale. During training, the window size is set to 25 ms with a hop size of 10 ms. Mask approximation (MA) [94] is used as the loss function and it is defined as

$$E^{MA}\left(M_{pred}\right) = \sum_{t,f} \left(M_{true} - M_{pred}\right)^2 \tag{3.8}$$

where M_{pred} is the predicted mask and M_{true} is the IRM. Note that the previously mentioned DNN approaches also use a mask approximation loss function. The network is trained with time steps of 100, a mini-batch size of 25 sequences, and a maximum training epoch number of 100. RMSprop is applied as the optimizer, since it has been proven as a good choice for RNNs [155]. We further trained and tested this LSTM structure using inputs and outputs in the linear frequency scale for comparison.

Bidirectional LSTM-based Phase-sensitive Mask estimation

A bidirectional-LSTM (BiLSTM) is an extended version of LSTM network that considers 'memory' in both directions (i.e., past series and future series). An illustration of BiLSTM layers is provided in Figure 3.2, where the BiLSTM units compute both the forward and backward sequences to update the output layer.

To investigate the influence of this memory difference, a BiLSTM architecture developed by Erdogan et al. [58] is investigated in this work. The BiLSTM estimates a Mel-scale

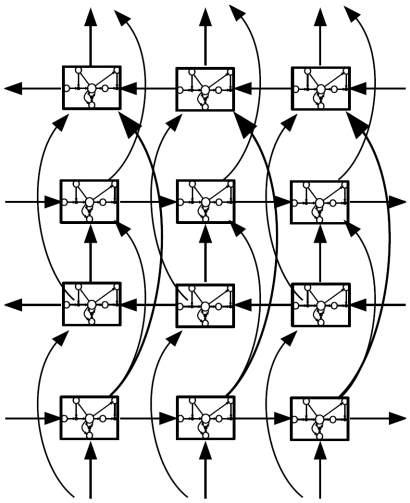

Figure 3.2: An illustration on the structure of BiLSTM layers. This figure is borrowed from [156].

phase-sensitive mask (PSM) that is defined as

$$M^{psm}(t,f) = \frac{|s(t,f)|}{|y(t,f)|} cos(\theta(t,f)). \tag{3.9}$$

We further truncate the PSM between 0 and 1 as suggested in [58]. A phase-sensitive spectrum approximation (PSA) is used as the loss function, since it leads to significant improvements over the mask-based loss function [58]. This is defined as

$$E^{PSA}\left(M_{pred}\right) = \sum_{t,f} \left(M_{true}|Y(t,f)| - M_{pred}|Y(t,f)|\right)^2, \tag{3.10}$$

where M_{true} is the ideal PSM and M_{pred} is the estimated one. The BiLSTM-based network contains two BiLSTM layers with 256 nodes in each layer. Other settings are identical to the LSTM method. A linear frequency domain implementation of the BiLSTM is also included for comparison.

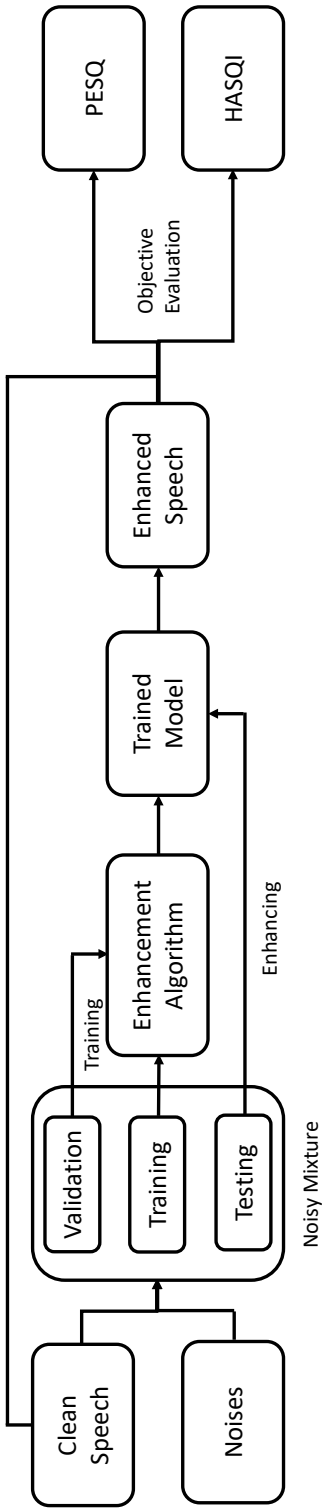

Figure 3.3: Experimental workflow for Section 3.1.

3.1.3 Experimental setup

The general workflow for the experiment in this section is depicted in Figure 3.3. Once the speech dataset is generated, we split them into 3 subsets including development, training and testing sets. The training set is used to train different speech enhancement algorithms and the development set is used to monitor the cross validation loss in order to determine the best trained model. The testing set is then used to generate enhanced speech for objective evaluation of speech quality. We also provide a summary of the speech enhancement systems in Table 3.1.

Table 3.1: Summary on the configurations of evaluated speech enhancement systems.

Methods	Deep learning approach	Learning target	Objective function	Optimizer
ASNA	✗	N/A	N/A	N/A
DNN-IRM	✓	IRM	MSE on T-F masks	AdaGrad
DNN-cIRM	✓	cIRM	MSE on T-F masks	AdaGrad
LSTM-IRM	✓	IRM	MSE on T-F masks	RMSprop
BiLSTM-PSM	✓	PSM	MSE on spectrograms	RMSprop

Speech Materials

Utterances from three different speech corpora are combined, in order to investigate the performance of the above-described speech enhancement algorithms on diverse speech materials. The speech data includes 1440 IEEE utterances [157] for both male and female speakers (720 utterances each), 250 male-speech utterances from the Hearing in Noise Test (HINT) corpus [158] and 2342 male and female utterances from the TIMIT database [159]. This results in a total number of 4032 clean speech utterances, where 2822 (70%) of them are used for constructing the training set and 605 (15%) are used for generating both the testing and development sets. The clean utterances are further corrupted by four types of noise at different levels, including airplane, babble, dog barking, and train noises. Noises are extracted from the AzBio [160] and ESC-50 datasets [161]. The clean speech and noise are mixed at several SNRs ranging from -5 dB to 20 dB with a step of 5 dB. All speech and

noise signals are resampled to a 16 kHz sampling rate before mixing. In total, the training set contains 16932 mixtures for each noise type. The development and testing sets consist of 3630 mixtures. We train and evaluate the systems' performance on each of the noise types separately.

Objective metric: PESQ

We use PESQ [87] as the objective speech quality metric for simulating evaluations by normal-hearing listeners, which has been widely used in the speech processing field, also known as the ITU-T P. 862. It was originally designed for evaluating speech signals that are transmitted over telephone lines, but it has been shown to have correlations with subjective evaluations by individuals with normal hearing [162]. The structure of the PESQ model is revealed in Figure 3.4.

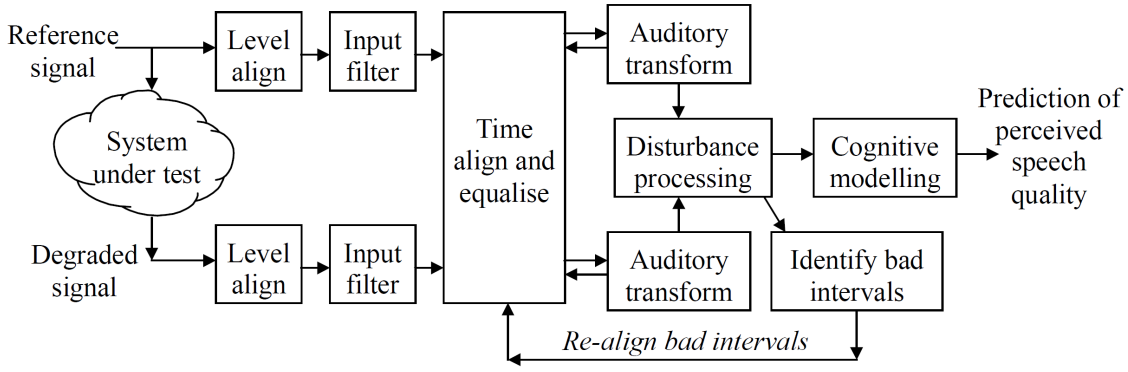

Figure 3.4: Structure of PESQ model. This figure is borrowed from [87].

As illustrated above, the PESQ model mainly consists of 6 parts, including signal pre-processing, time alignment, auditory transform, disturbance processing, cognitive modelling, identify and realign bad intervals. In the first step, the reference and degraded (i.e., enhanced signals in our case) signals are rescaled to a standard listening level. Next, the time alignment stage involves the following steps: filtering both signals with a narrow band filter (emphasize perceptually important parts), estimating delay based on envelope, dividing reference signal into utterances, estimating delay for each utterance, identifying

delay for each utterance based on fine correlation histogram. After these steps, the frame-by-frame delay can be derived, which is later used in the auditory transform. The auditory transform part later maps the signals into a representation of perceived loudness in time and frequency, which includes stages of bark spectrum, frequency equalization, equalization of gain variation and loudness mapping. PESQ considers the absolute difference between the degraded and the reference signals as a measure of audible error and the final score involves a linear combination of two (averaged) disturbances, i.e., symmetric disturbance d_{SYM} and asymmetric disturbance d_{ASYM}. The regression coefficients are derived based on 30 human listening tests, the testing conditions include mobile network, fixed network, VOIP and multi-type network. Eventually, the PESQ score can be computed as

$$PESQ = 4.5 - 0.1 d_{\text{SYM}} - 0.0309 d_{\text{ASYM}}. \tag{3.11}$$

It predicts a mean opinion score (MOS) that ranges from -0.5 (bad) to 4.5 (excellent).

Objective metric: HASQI

HASQI, a more recent objective evaluation metric, captures the noise effects, nonlinear distortions, linear filtering and spectral changes between two signals in order to resemble the processing that is performed by normal and impaired auditory systems. The HASQI score is calculated as a product of two terms, namely non-linear $Q_{\text{NON-LIN}}$ and linear Q_{LIN} indexes, which can be formulated as

$$HASQI = Q_{\text{NON-LIN}} \times Q_{\text{LIN}}, \tag{3.12}$$

where the non-linear index measures changes in the signal temporal fine structure (TFS) while not considering any long-term spectral changes. It can be calculated from a combi-

nation of the cepstral correlation index and vibration correlation value as

$$Q_{\text{NON-LIN}} = c^2 v, \tag{3.13}$$

where c and v represent the cepstral correlation and vibration correlation, respectively. The linear index, on the other hand, compares the long-term spectral representations of the reference and degraded signals while not considering the short-term changes in signal modulation and TFS

$$Q_{\text{LIN}} = 1 - 0.579\sigma_1 - 0.421\sigma_2, \tag{3.14}$$

where σ_1 and σ_2 denote the standard deviations of the differences in the spectral shape and spectral slope, respectively. The HASQI algorithm returns a quality measure between 0 (poorest) and 1 (perfect).

Not only does HASQI correlate well with speech quality perceived by HI listeners, it has also been reported that HASQI achieves comparable performance to PESQ for normal-hearing based evaluations [150]. HASQI requires hearing thresholds at six audiometric frequencies (i.e., 250, 500, 1000, 2000, 4000 and 6000 Hz) to model the hearing loss of hearing-impaired individuals. We use the hearing profiles of various age groups for typical females (based on 936 listeners) and males (based on 756 listeners) reported in [95], also shown in Figure 3.5.

Table 3.2 summarizes the average hearing thresholds used for HASQI simulation of different age groups and gender conditions. Higher hearing thresholds (in dB HL) indicate a greater degree of hearing loss. We observe that high-frequency hearing loss is typical among older adults, and the severity of hearing loss grows with age and is greater for male listeners.

Note the signals are spectrally shaped to compensate for hearing loss before they are evaluated by HASQI. A standard formula (i.e., NAL-R) for hearing-aid fitting [163] is used to determine the amount of amplification in each frequency region. The amplification is

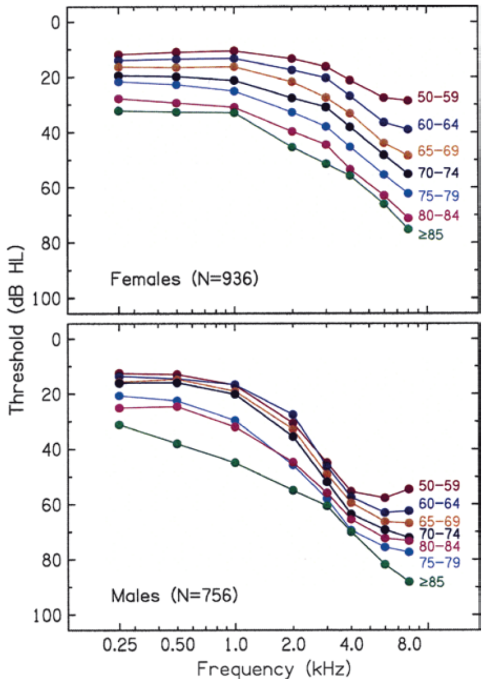

Figure 3.5: Hearing profiles of different age groups and genders. This figure is borrowed from [95].

implemented to ensure that the speech signals are equally audible under all age and gender conditions. Therefore, the predicted speech quality is not driven solely by audibility.

3.1.4 Simulation results

Normal-hearing simulation results

Figure 3.6 provides the PESQ scores for the original mixtures without enhancement, as well as the enhanced signals that are produced by the speech enhancement algorithms. We show the PESQ scores averaged across the four noise types for brevity. The simulation results for deep learning speech enhancement algorithms on Mel frequency scale are also included.

All these algorithms improve the quality of the noisy mixtures according to the PESQ scores. Deep learning algorithms also significantly outperform the NMF-based ASNA approach (e.g., comparing LSTM and ASNA), which is consistent with results from [50, 55, 94]. At the lower SNRs (i.e., -5 dB and 0 dB), the LSTM Mel-scale method performs the

Table 3.2: Hearing thresholds (dB HL) of male (M) and female (F) subjects across different age groups.

Age Group	Frequency (Hz)					
	250	500	1000	2000	4000	6000
50-59 M	12.3	12.6	16.4	30.4	55.1	57.5
50-59 F	11.6	10.9	10.4	13.2	21.1	27.4
60-69 M	14.8	14.8	17.7	29.9	58.3	64.5
60-69 F	15.1	14.9	14.7	19.5	29.8	40.0
70-79 M	18.3	19.1	24.7	40.4	66.1	72.1
70-79 F	20.7	21.3	23.1	30.1	41.5	51.4
80+ M	28.0	31.2	38.3	49.6	67.5	76.7
80+ F	29.9	30.9	31.7	42.4	54.3	64.1

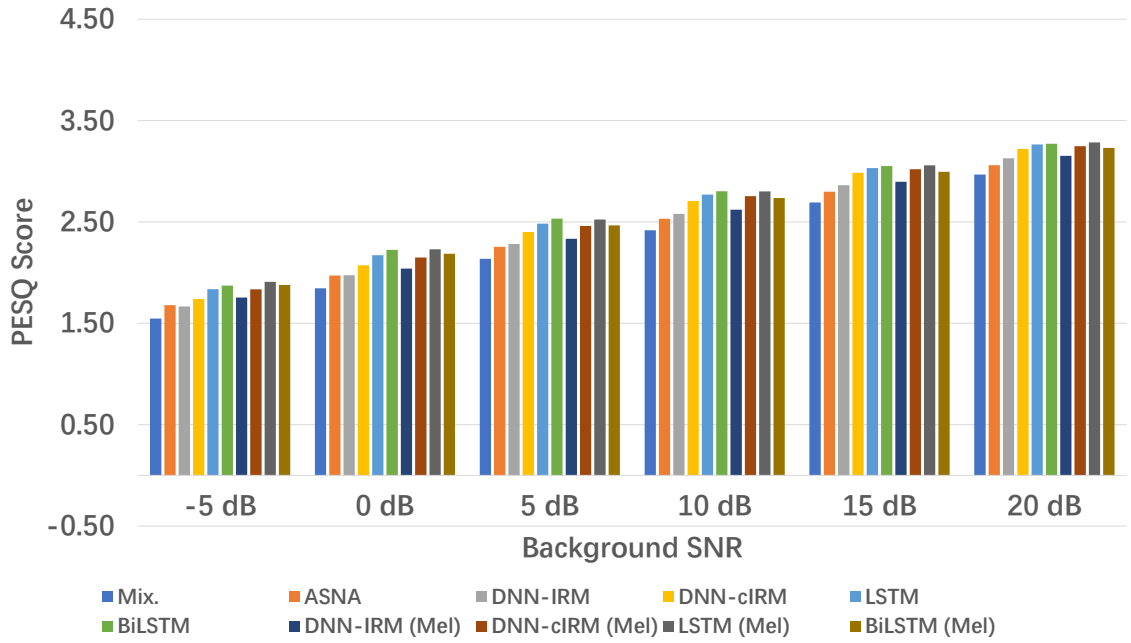

Figure 3.6: PESQ scores for different speech enhancement algorithms at different SNRs.

best (i.e., PESQ scores: 1.91 and 2.23). In SNRs from 5 to 10 dB, the BiLSTM linear-scale approach performs better than other deep learning methods (e.g., 2.80 at 10 dB). While at higher SNRs (i.e., 15 and 20 dB), the LSTM Mel-scale method performs the best (i.e., PESQ scores: 3.06 and 3.28). Averaging across the SNRs, the LSTM Mel-scale method slightly outperforms the other algorithms (i.e., PESQ score: 2.63 on average). Mel frequency domain processing often leads to improved performance for both DNN- and RNN-

based structures. We also find that RNN-based structures perform better than conventional DNN-based methods, and we attribute this to LSTM's advantages in terms of dealing with time series data and solving the problem for vanishing and exploding gradients [154] in RNNs.

Hearing-impaired simulation results

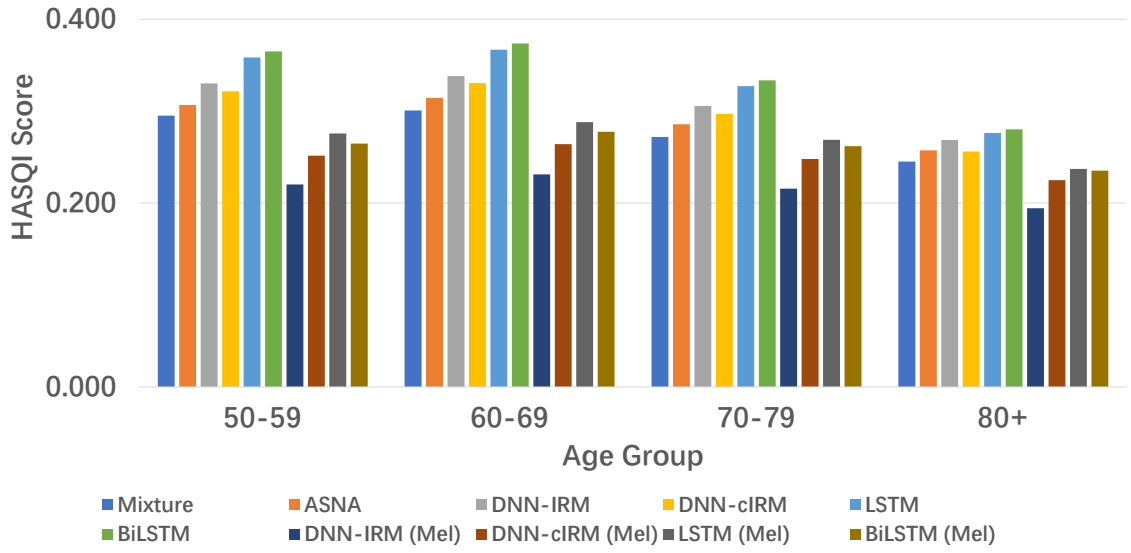

Figure 3.7: HASQI scores (male group) for speech enhancement algorithms across noise conditions and background SNRs.

Figure 3.7 and 3.8 present the HASQI scores for male and female listeners with various hearing loss conditions, averaged over noise types and background SNR levels for brevity. In general, most speech enhancement algorithms show improvement in speech quality for hearing-impaired listeners. Moreover, the amount of improvement decreases with increasing age, which is expected since these are the more challenging cases. Among the speech enhancement algorithms, the BiLSTM system with PSM in linear-scale performs the best across all age groups (i.e., HASQI score: 0.378 on average across genders), but its results are almost identical to the LSTM system with IRM. Within the DNN-based speech

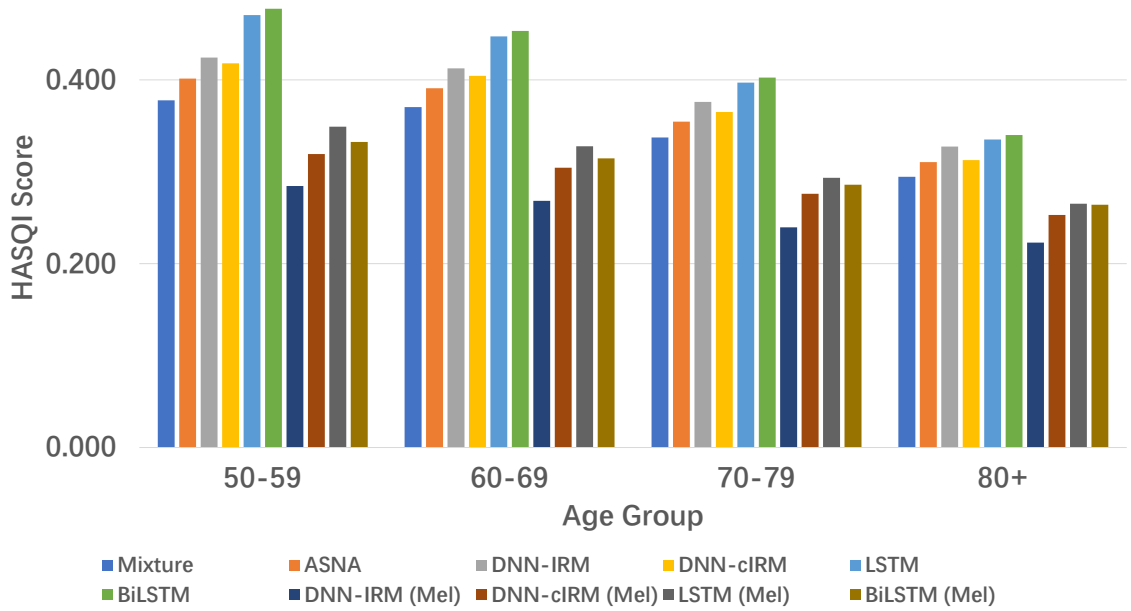

Figure 3.8: HASQI scores (female group) for speech enhancement algorithms across noise conditions and background SNRs.

enhancement approaches, the IRM linear-scale approach and cIRM approach perform similarly (linear and Mel scale).

Surprisingly, we notice that for the DNN- and RNN-based speech enhancement systems, the Mel-domain processing results in reduced HASQI scores as compared to the linear-frequency domain approaches (e.g., average HASQI scores for LSTM-based method: 0.372 vs. 0.288 between linear and Mel frequency scales). This is contrary to the PESQ results, where the Mel-domain processing usually improves speech quality. We infer that this may result from the deteriorated frequency resolution of the enhanced signals following the Mel-domain transformation, especially at higher frequencies. PESQ is unaffected by this, since it assesses the speech quality on a narrower frequency range (3.1 kHz) [87] than HASQI (12 kHz) does [93].

3.1.5 Summary

We investigate the performance of several speech enhancement algorithms on a diverse speech dataset, with a particular interest in simulated hearing loss environments. The RNN-based methods result in significantly higher PESQ and HASQI scores for normal-hearing listeners. For hearing-impaired listeners, the BiLSTM method achieves the best performance in all age groups for both genders. We also found that for both DNN- and RNN-based methods, Mel frequency domain processing can often lead to improved PESQ scores, but reduced HASQI scores. In the following section (Section 3.2), we conduct a MUSHRA listening study to investigate the human preference on the processing frequency scale for speech enhancement algorithms.

3.2 Human Preference on Frequency Scale for Data Processing

3.2.1 Motivation and organization

Speech enhancement has been a popular research topic over the years, most studies are using objective speech quality metrics for performance evaluation due to the complication of conducting listening studies. PESQ and HASQI are two widely adopted metrics, where PESQ is purely based on acoustic features and HASQI simulates the human auditory system. Recent deep learning approaches are found powerful in dealing with speech enhancement problems since it can utilize tons of training data. There is a trend that many recent studies were using audio features in a non-linear frequency scale (i.e., Mel scale) rather than in the original linear frequency scale. Experimental results from the previous section showed a mismatched pattern in terms of objective scores obtained from PESQ and HASQI, on the performance of the enhancement systems implemented in linear and Mel frequency scales. PESQ scores suggest Mel frequency processing could further improve the quality of enhanced speech, whereas HASQI scores indicate an opposite pattern.

To better understand how does different frequency scales influence human perception,

we conducted a human listening study on the preference of frequency scales for speech enhancement following the MUSHRA procedure (ITU-R BS.1534-1) [96]. The findings on human preference of frequency scales could serve as a pilot study for future implementation of speech enhancement algorithms. We further analyze the correlations between human responses and objective scores predicted by PESQ and HASQI.

The remainder of this section is organized as follows. Section 3.2.2 describes the details of the human listening study. We present and analyze experimental results in Section 3.2.3. A summary is provided in Section 3.2.4.

3.2.2 Listening study

Subject recruitment

A total of 10 participants from the undergraduate population at Indiana University were recruited, all participants were native speakers of American English and self-reported to be NH. The informed consents, approved by the Institutional Review Board (IRB) at Indiana University, were obtained from all participants before data collection. The listening study was conducted following the Declaration of Helsinki. The experiment was completed in a single test session, which took about two hours for each participant. A cash compensation is provided after completing all the experimental tasks.

Speech Materials

Details on the dataset generation is provided in Section 3.1.3. For listening experiments, we use the top 50 enhanced speech utterances (from testing set) that are most different in terms of their HASQI scores for Mel and linear frequency scales. As a result, enhanced speech utterances from IEEE, HINT and TIMIT sentences with SNRs ranging from -5 to 20 dB were collected for the listening experiment. Babble noise is considered. We included enhanced speech generated by four deep learning based speech enhancement algorithms: DNN-based IRM estimation, DNN-based cIRM estimation, LSTM-based IRM estimation,

and BiLSTM-based PSM estimation.

Experiment procedure

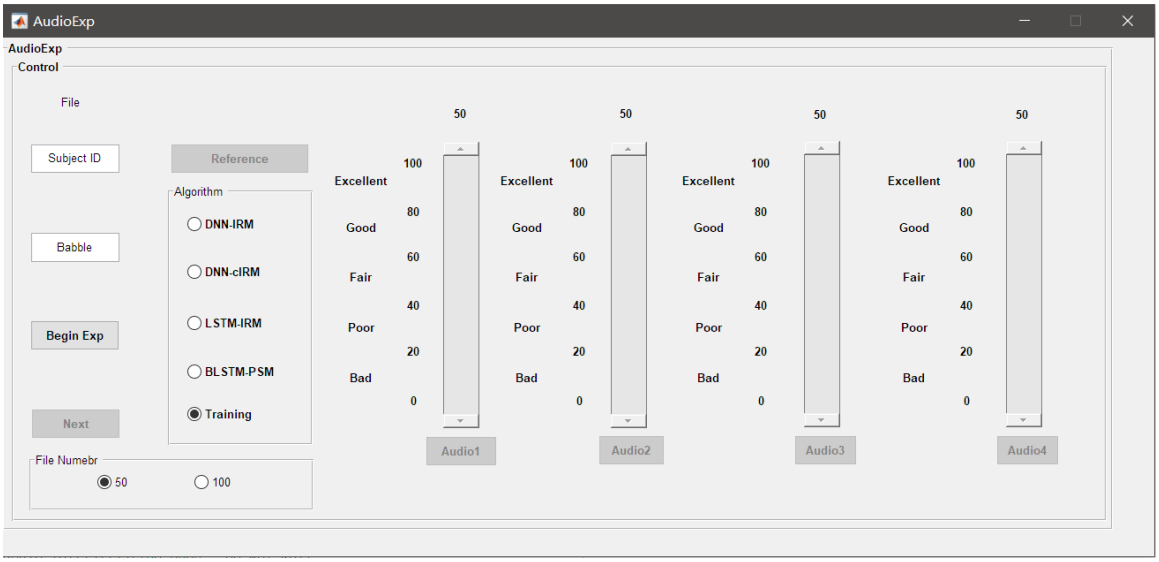

Figure 3.9: The MATLAB GUI for the MUSHRA listening study.

The participants are first asked to fill out a data sheet indicating if they are NH or not. Then, an informed consent statement is provided, all the participants are required to read through it then provide a signature if they agree to participate in the listening study. The consent statement describes the purpose and requirements of this study, including:

- Study purpose, risks and benefits of taking part in the study.
- Voluntary nature of study, and option of not participating.
- General procedure of the listening study, time costs, and payment amount.
- Data confidentiality, contact for questions or problems.

The experimenter (i.e., a graduate student in the lab) can provide help if the participant has any questions regarding the experimental procedure. The subjects have the option to keep a copy of the consent sheet. During the listening experiment, the subjects are encouraged to take breaks as often as needed to reduce the possible fatigue or distraction.

The listening study is built with MATLAB, an illustration on the graphical user interface (GUI) for this experiment is provided in Figure 3.9. It follows the MUSHRA procedure [96], where in each trial, the clean reference speech will be presented at the beginning. There are in total four hidden audios with their corresponding "Audio1–4" buttons to playback the audios and sliders to give quality ratings. All the audios (including the reference speech) are allowed to be played for unlimited number of times. Among the four hidden audios, there is one hidden clean reference speech, a hidden low-pass filtered (designed using "fdesign.lowpass" function in MATLAB, filter order set to 50, cutoff frequency at 3.5 kHz) version of the clean reference speech, a hidden enhanced speech (linear frequency scale) and another hidden enhanced speech (Mel frequency scale). The order of the four hidden audio signals is randomly generated. The participants were asked to provide a rating for each audio before moving to the next trial, they were also instructed to provide ratings based on the 0–100 scale (i.e., 0–20: 'bad', 20–40: 'poor', 40–60: 'fair', 60–80: 'good', 80–100: 'excellent'). Subjects were informed that the clean reference speech corresponds to a rating of '100' and other ratings should be given in a sense that they are all relative to each other.

The listening experiment consists of the following steps: a training phase to make the subject familiarize with the task, followed by four listening tests presented in random order that correspond to four types of speech enhancement algorithms. Each listening test takes approximately 20 minutes to finish.

All stimuli were presented diotically to the participants via a 24-bit soundcard (Microbook II, Mark of the Unicorn, Inc.) and a pair of headphones (HD280 Pro, Sennheiser electronic GmbH and Co. KG). The participants were seated in a sound-attenuating booth during the experiment.

3.2.3 Experimental results

Human preference on frequency scale

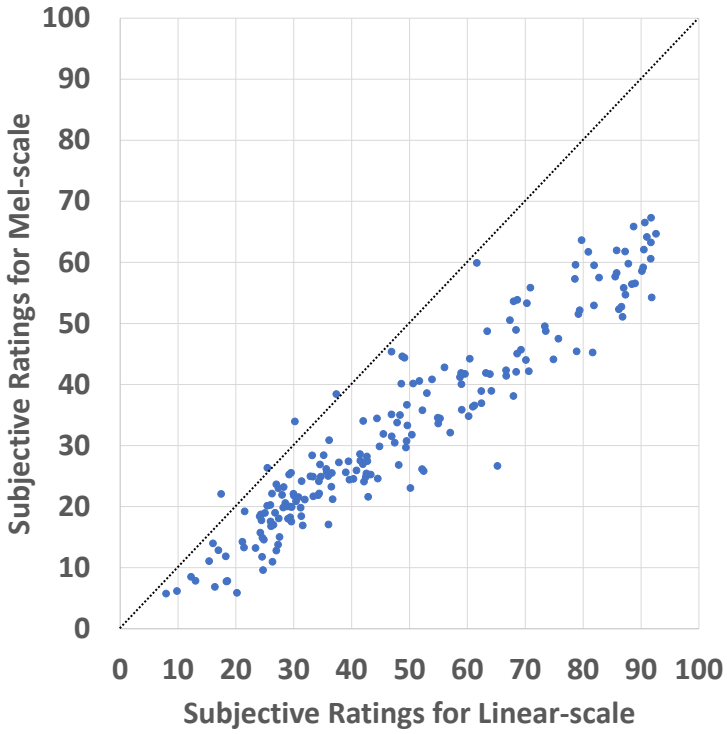

Figure 3.10: Scatter plots of human ratings on enhanced speech samples processed in linear vs. Mel frequency scales. The diagonal (i.e., equal ratings) is represented by the dashed line.

We first provide the quality ratings on the speech signals processed with different frequency scales in Figure 3.10, where the responses for four speech enhancement algorithms are included. We can observe that there is a clear human preference over speech signals processed in linear frequency scale than the ones processed in Mel frequency scale, as most of the data points fall below the dashed diagonal line (i.e., ratings for linear scale are higher). In total, there are 196 out of 200 pairs (i.e., 98%) that demonstrate speech processed in the linear frequency scale has better quality than the one processed in the Mel frequency scale.

Results here indicate a strong human preference on speech signals processed in linear frequency scale rather than Mel frequency scale. The average score for enhanced speech processed in linear frequency scale is about 47% higher (i.e., linear vs. Mel: 48.8 vs. 33.1) than speech signals processed in Mel frequency scale. For speech signals with lower qual-

ity (e.g., ratings between 0 to 40), the ratings given to Mel and linear are similar. However, the preference over linear frequency scale becomes stronger (i.e., increased distance from data points to diagonal line) for higher quality speech.

Correlation results for DNN-based systems

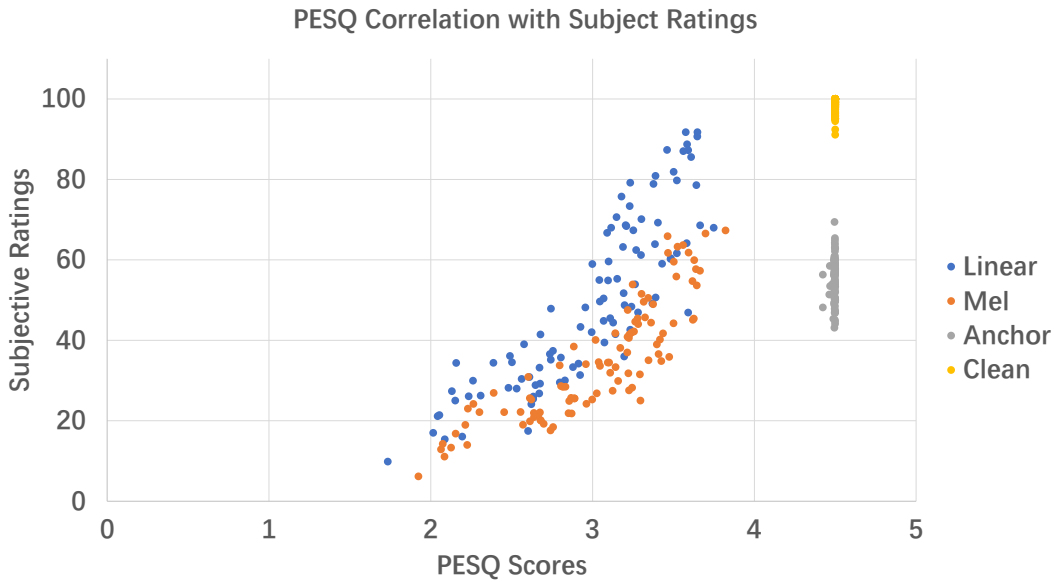

Figure 3.11: PESQ correlation with subjective ratings for DNN-based speech enhancement systems.

Figures 3.11 and 3.12 show the correlations between objective scores (PESQ and HASQI) and the human quality ratings for the DNN-based approaches (i.e., DNN-IRM and DNN-cIRM).

It can be observed that HASQI scores demonstrate a better correlation with subjective ratings (i.e., Figure 3.12), as compared to the correlation for PESQ scores in Figure 3.11. Especially for ratings on the anchor signals (i.e., low-pass filtered clean speech), PESQ returns a score that is near perfect (e.g., 4.5) where HASQI is more conservative (i.e., scores around 0.6). We believe this is mainly caused by the difference in the bandwidth for speech assessment, PESQ focuses on the lower-frequency part (i.e., below 3.1 kHz) whereas HASQI assesses the speech signals at a much broader frequency band (i.e., to 12

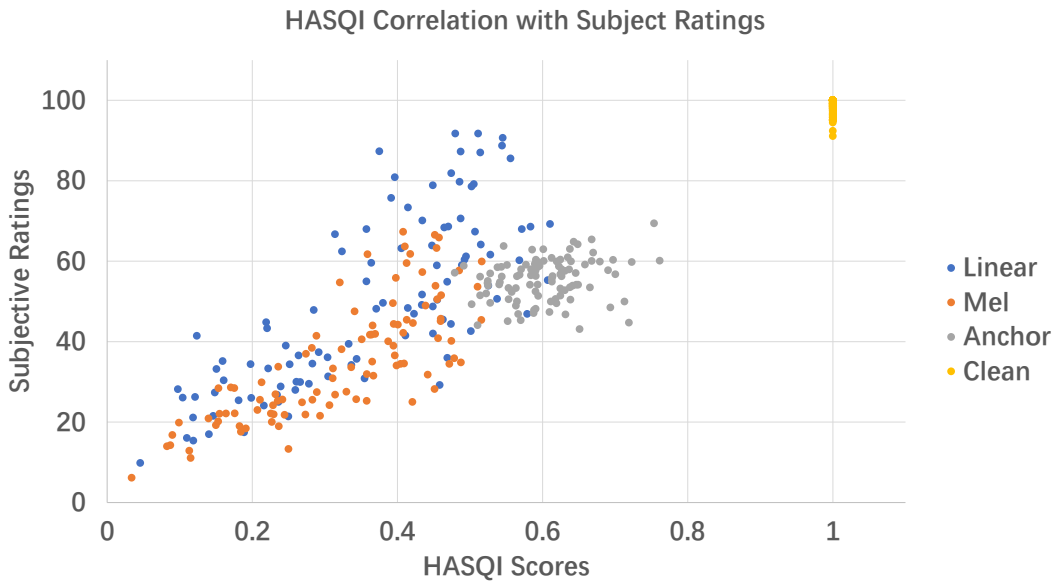

Figure 3.12: HASQI correlation with subjective ratings for DNN-based speech enhancement systems.

kHz). The effects from low-pass filtering are captured by HASQI and also NH human listeners.

On average, the Pearson correlation coefficients for PESQ and HASQI between objective and subjective ratings for DNN-based speech enhancement systems are 0.747 and 0.921, respectively. Their corresponding r^2 values are 0.558 and 0.848.

Correlation results for RNN-based systems

Likewise, we present the correlation results between objective and subjective scores for RNN-based speech enhancement systems (i.e., LSTM-IRM and BiLSTM-PSM) in Figures 3.13 and 3.14. We observe a similar pattern as the correlation results obtained for DNN-based systems. The HASQI provides a better correlation with subjective ratings especially when considering the anchor signals. Again, the PESQ does not correlate well with human ratings when low-pass filtering is introduced, because of its limited bandwidth for speech assessment.

The Pearson correlation coefficients for PESQ and HASQI on RNN-based speech en-

hancement systems are 0.784 and 0.956, respectively. Their r^2 values are 0.614 and 0.914.

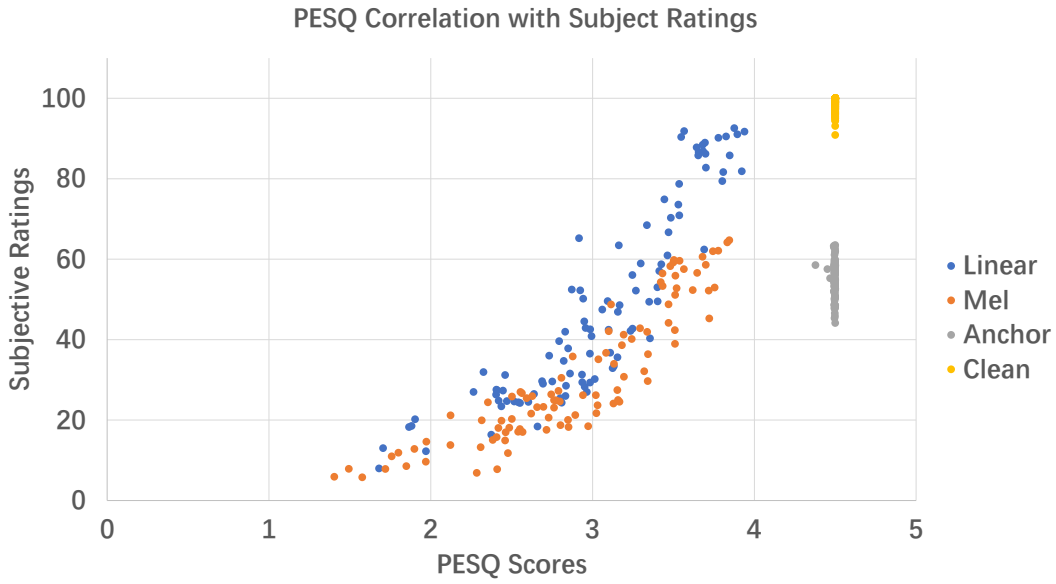

Figure 3.13: PESQ correlation with subjective ratings for RNN-based speech enhancement systems.

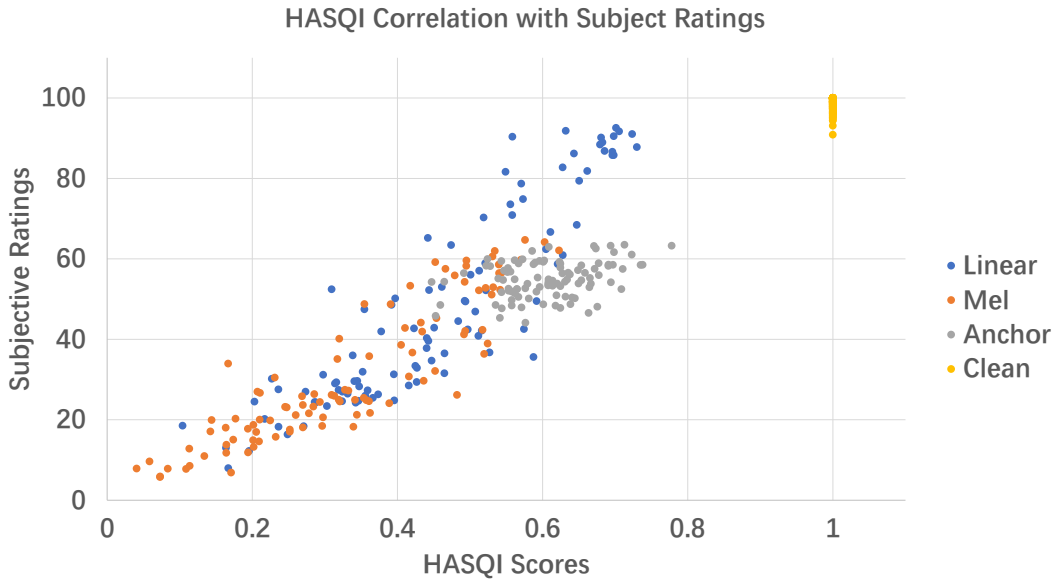

Figure 3.14: HASQI correlation with subjective ratings for RNN-based speech enhancement systems.

3.2.4 Summary

In this section, we investigated human preference on frequency scales for speech enhancement. Several MUSHRA listening tests were conducted on NH listeners and experimental results indicate a strong human preference for speech signals processed in linear frequency scale. Results are consistent with HASQI predictions and contradict the findings from PESQ index.

We further analyzed the correlations between objective (i.e., PESQ and HASQI) and subjective ratings, where HASQI demonstrates a stronger correlation with subjective ratings for both DNN- and RNN-based speech enhancement systems. Due to a narrower bandwidth for speech assessment, PESQ fails on estimating the speech quality for anchor signals that are low-pass filtered.

In conclusion, speech enhancement algorithms that operate in linear frequency scale are preferable to the ones that operate in Mel frequency scale for NH listeners. Possibly due to the lower frequency resolution (i.e., 100 Mel-bins) of our implementation for the Mel-frequency processing. Future implementation on speech enhancement algorithms should consider this factor. Moreover, considering the fact that most of the current speech signals are sampled at 16 kHz or above, it is recommended to use wide-band evaluation metrics such as HASQI for quality assessment instead of PESQ, which could not capture high-frequency factors that may negatively affect the speech quality perceived by human listeners.

CHAPTER 4

PERCEPTION OF PHASE DISTORTION FOR HEARING-IMPAIRED LISTENERS

Phase serves as a critical component of speech that influences quality and intelligibility. Current speech enhancement algorithms are beginning to address phase distortions, but the algorithms focus on normal-hearing (NH) listeners. It is not clear whether phase enhancement is beneficial for hearing-impaired (HI) listeners.

In this chapter, we investigated the influence of phase distortion on speech quality through a listening study, in which NH and HI listeners provided speech quality ratings following the MUSHRA procedure. In one set of conditions, the speech was mixed with babble noise at 4 different signal-to-noise ratios (SNRs) from -5 to 10 dB. In another set of conditions, the SNR was fixed at 10 dB (with babble noise) and the noisy speech was presented in a simulated reverberant room with reverberation time (T60s) ranging from 100 to 1000 ms. The speech level was kept at 65 dB SPL for NH listeners and amplification was applied for HI listeners to ensure audibility.

Ideal ratio masking (IRM) was used to simulate speech enhancement. Two objective metrics (i.e., PESQ and HASQI) were utilized to compare subjective and objective ratings. Results obtained from the listening study indicate that phase distortion has a negative impact on perceived quality for both NH and HI listeners. Both objective metrics demonstrate good correlations with human ratings.

The work in this chapter has been published in the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH) [164].

4.1 Introduction

Phase serves as a critical component of a speech signal, and it makes important contributions to speech intelligibility and perceived quality. When analyzing a speech signal in the time-frequency (T-F) domain by performing short-time Fourier transform (STFT) analysis, each time-by-frequency location in the resulting spectrogram contains not only magnitude but also phase information. At each time frame, the phase spectrum represents how the various frequency components in the speech signal are temporally aligned. At each frequency, the progression of phase across consecutive time frames represents the temporal fine structure (TFS) of the speech signal, which is important to the perception of talker gender, voicing, and intonation [66]. Replacing or distorting the phase information when reconstructing speech from the spectrogram would lead to degraded speech intelligibility [165].

Many of the existing speech enhancement systems only operate based on the magnitude spectrogram and keep the noisy phase unchanged when converting the enhanced speech to the time domain [50, 94, 166]. Recently, a number of studies have shown that better phase estimation of the original speech improves both subjective and objective speech quality [67, 55]. However, these studies addressing the importance of phase in speech enhancement have only focused on normal-hearing (NH) listeners. Approximately 30% of older adults above the age of 65 years in the United States suffer from hearing loss [148]. Modern digital hearing aids, besides amplifying the acoustic signals, also consist of built-in speech enhancement algorithms to remove unwanted background noise that corrupts the speech [167, 143]. The noise reduction is performed before amplification. Before the phase-preserving speech enhancement algorithms can be implemented into hearing aids, it is necessary to evaluate whether hearing-impaired (HI) individuals would actually benefit from them.

It is known that HI listeners have poorer sensitivity to the TFS [89, 90] and benefit

less from TFS cues for speech understanding [91, 92]. Therefore, it may be expected that preserving the phase information in speech enhancement may not lead to the same degree of benefit for HI listeners compared to NH listeners. In order to optimize speech enhancement algorithms for the HI population, it is crucial to quantify their sensitivity to phase distortions, especially for the phase distortions remained in the enhanced speech following phase-insensitive enhancement. If HI listeners consistently rate the phase-distorted speech as having lower quality, even after phase-insensitive enhancement, then it would suggest the potential benefit for a phase-sensitive speech enhancement system. Furthermore, objective speech quality metrics developed for NH listeners may not directly generalize to HI listeners [93]. Therefore, it is not clear whether existing speech quality metrics could be adequately used to capture the effect of phase distortion for HI listeners.

To address these open questions, we collected human quality ratings on noisy speech signals with different degrees of phase distortion (or equivalently distortion to the TFS) for both NH and HI listeners using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) procedure [96]. The quality ratings were repeated on speech signals with and without processing from a common phase-insensitive algorithm based on an ideal ratio mask (IRM) in the T-F domain [50]. This allowed us to investigate whether the perceived quality by NH and HI listeners would be adversely affected if phase distortion remained in the enhanced speech following traditional, magnitude-based speech enhancement. Comparing the ratings from these two conditions also reveals the expected benefits from a phase-insensitive speech enhancement system. To assess the agreement between subjective and objective speech quality, the subjective quality ratings were compared against two objective metrics, namely perceptual evaluation of speech quality (PESQ) [87] and hearing-aid speech quality index (HASQI) [93].

The rest of this chapter is organized as follows, the implementation of phase distortion and the procedure for the subjective listening test are described in Section 4.2. In Section 4.3, the results obtained from the listening test and their correlations to the objective metrics

are presented. Finally, we provide the summary in Section 4.4.

4.2 Methods

4.2.1 Phase distortion

Phase distortion was artificially applied to the speech materials by introducing random perturbations to the phase spectrogram using the following steps. First, both the magnitude spectrogram [$|s(t,f)|$] and the phase spectrogram [$\angle s(t,f)$] in the T-F domain were extracted, where $\angle$ extracts the angle for a complex value. A (hamming) window size of 25 ms with a step size of 10 ms was used during STFT analysis. Second, four degrees (i.e., 25%, 50%, 75%, and 100%) of phase distortion were applied to the phase spectrogram according to

$$\angle s(t,f)_{\text{distorted}} = \angle s(t,f) + \alpha \cdot \phi(t,f), \tag{4.1}$$

where $\angle s(t,f)_{\text{distorted}}$ denotes the distorted phase in the T-F domain; α denotes the amount of phase distortion that ranges from 25% to 100%; and $\phi(t,f)$ represents random phase perturbations drawn from a uniform distribution between 0 and 2π, independently for each T-F location. Similar distortion amounts were used in earlier studies involving NH listeners [168, 169, 170]. These amounts could also reflect potential errors due to inaccurate phase estimation from a phase-aware speech enhancement system. Finally, the phase-distorted speech was resynthesized with the inverse STFT that combines the original magnitude spectrogram and the distorted phase spectrogram.

4.2.2 Stimuli

Speech utterances from the IEEE corpus [157] produced by a female talker were adopted for the listening test. To ensure speech audibility for HI listeners, a standard hearing-aid prescription formula (i.e., NAL-R) [163] was used to amplify the speech stimuli. This formula prescribes linear gains for various frequency regions according to the degrees of

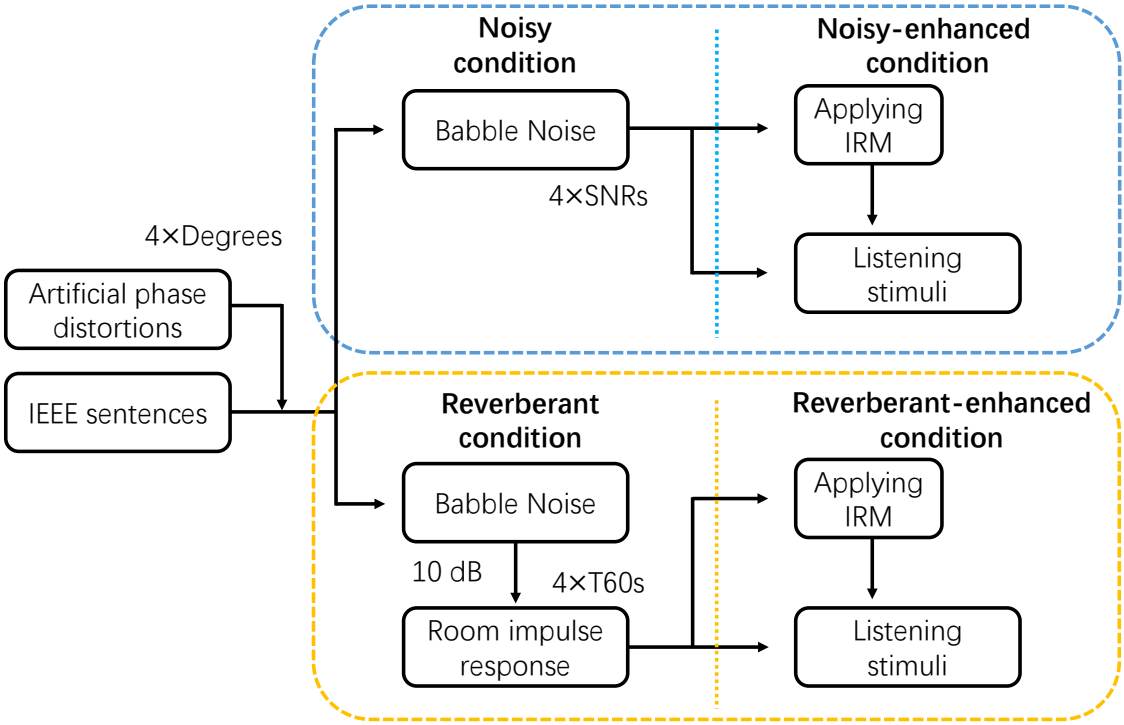

Figure 4.1: A depiction on the generation of speech stimuli for the four testing conditions.

hearing loss in these regions. The speech level was calibrated to 65 dB SPL before amplification. All signals were resampled to 16 kHz before further processing.

There were in total four test conditions in the listening test as illustrated in Figure 4.1. In the Noisy condition, the speech stimuli was presented with a simultaneous 10-talker babble from the AzBio database [160] at 4 different signal-to-noise ratios (SNRs), from -5 dB to 10 dB with a 5 dB step. In the Noisy-Enhanced condition, the stimuli were the same as those in the Noisy condition except that they were further masked by the IRM [50] before presented to the listeners. The IRM is defined as

$$\text{IRM}(t,f) = \left(\frac{|s(t,f)|^2}{|s(t,f)|^2 + |n(t,f)|^2}\right)^{\frac{1}{2}}, \tag{4.2}$$

where $|s(t,f)|$ and $|n(t,f)|$ represent the magnitude spectrograms of the clean speech signal and the interfering noise, respectively. Note that IRM performs phase-insensitive

speech enhancement since it only recovers magnitude of the speech, where the original noisy phase is used during iSTFT.

In the Reverberant condition, the SNR between the speech and babble noise was fixed at 10 dB and the stimuli were presented with simulated reverberation using the image-source method [171]. The reverberation algorithm simulated a room with a dimension of 4 m$\times$4 m$\times$3 m (length$\times$width$\times$height), the sound source was located at (2 m, 3.5 m, 2 m), and the listener was located at (2 m, 1.5 m, 2 m). The sound velocity was assumed to be 340 m/s. The reverberation times (T60s) were set to 100, 200, 500, and 1000 ms. In the Reverberant-Enhanced condition, the stimuli were the same as those in the Reverbrant condition except that they were further masked by the IRM before presentation. Note that the IRM in this condition was applied only on the noise without removing reverberation, which resembles a system that is not trained on reverberant data. The noisy and reverberant conditions (with and without IRM-based speech enhancement) were chosen since they follow but extend a similar approach that studied the importance of phase for NH listeners [67].

Since most HI listeners have high-frequency hearing loss, all stimuli were low-pass filtered at 4 kHz. This ensures that the perceived speech quality is not dominated by hearing loss at high-frequency bands. All stimuli were presented monotonically in the participants' better ear based on the hearing screening results. A 24-bit soundcard (Microbook II, Mark of the Unicorn, Inc.) and a pair of headphones (HD280 Pro, Sennheiser electronic GmbH and Co. KG) were used. The participants were seated in a sound-attenuating booth during the study.

4.2.3 Subject recruitment

A total of 18 participants were recruited, including 10 NH listeners (4 males, 6 females, recruited from the undergraduate population at Indiana University) and 8 HI listeners (3 males, 5 females, average age: 68 ($SD = 5.53$)). The inclusion criteria include (1) No ongoing outer- or middle-ear disorders and no history of ear-related surgeries, (2) No sign

Table 4.1: Average auditory thresholds of participants from NH and HI groups with standard deviations shown in parentheses.

Average Auditory Thresholds (dB HL)						
	Frequency (kHz)					
	.25	.5	1	2	4	6
NH	10.0	8.5	6.0	5.0	5.0	11.0
	(4.7)	(5.3)	(3.9)	(5.3)	(6.7)	(6.2)
HI	23.1	20.6	26.9	36.3	44.4	46.3
	(8.0)	(9.0)	(14.1)	(18.5)	(19.7)	(16.6)

of dementia, as confirmed by the Min Mental Status Exam (MMSE-2), and (3) Native speaker of American English.

Audiometric evaluations were performed on all NH listeners, including otoscopy and air-conduction pure-tone audiometry. All NH listeners had audiometric thresholds below 20 dB HL from 250 to 6000 Hz. For the HI listeners, the hearing evaluation additionally included bone-conduction audiometry, tympanometry, and hearing-related case history. All HI listeners had at least mild symmetric hearing loss of a sensorineural origin. The average audiometric thresholds for the two groups are listed in Table 4.1. The current study was conducted following the Declaration of Helsinki and approved by the Institutional Review Board (IRB) at Indiana University. Informed consents were obtained from all participants before the data collection, containing information about (1) Study purpose, potential risks and benefits, (2) General procedure of the listening study and task duty, (3) Time costs and payment amount and (4) Data confidentiality.

During the listening experiment, all subjects are encouraged to take breaks to reduce the possible fatigue and alleviate boredom. A monetary reward is provided to all participants after completing the experiment. Each experiment requires two 2-hour sessions to complete.

4.2.4 Procedure

Listeners provided subjective ratings on the stimuli following the MUSHRA procedure, recommended in ITU-R BS.1534 [96]. During the experiment, a graphic user interface (GUI) developed using MATLAB was shown on a computer screen in front of the listener, which is also illustrated in Figure 4.2.

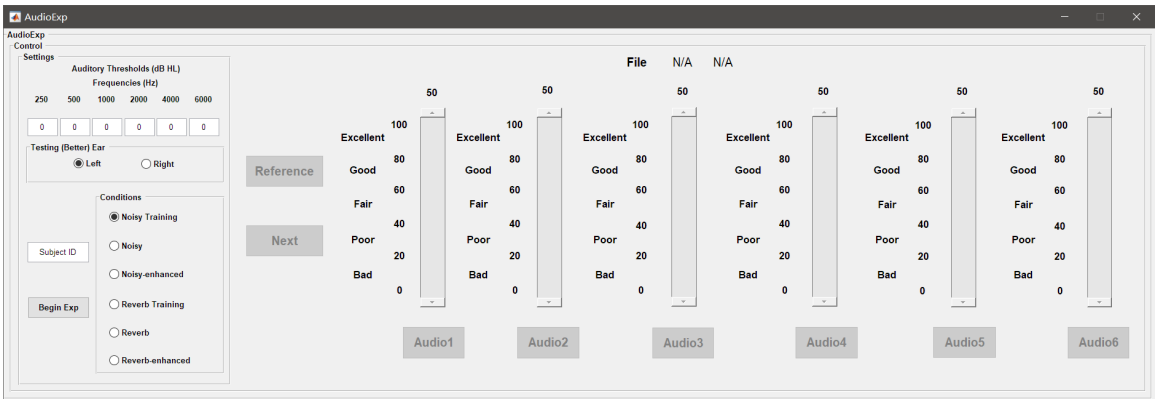

Figure 4.2: MATLAB GUI for the MUSHRA listening study on perception of phase distortion.

The experimenter will first put the audiometric thresholds at 250, 500, 1000, 2000, 4000 and 6000 Hz into the corresponding boxes (for sound amplification) and choose the better ear before beginning the experiment. Practice stages (i.e., "Noisy and Reverb Training") are included for both noisy and reverberant conditions in order to familiarize participants with the listening study.

The user interface contained a "Reference" button that corresponded to a reference stimulus, which was the original clean speech low-pass filtered at 4 kHz. The six additional buttons (i.e., "Audio1–6") representing the six test stimuli, which were six versions of the same sentence stimulus as the reference. One of these buttons corresponded to the hidden reference stimulus; one button corresponded to a hidden anchor stimulus, which was the original clean speech low-pass filtered at 2 kHz; the remaining four buttons corresponded to the phase-distorted speech with four degrees of phase distortion. The correspondence between the buttons and the stimuli was randomized from trial to trial. On each trial, the

clean reference will be automatically played at the start, then the listener clicked on each of the buttons to hear the corresponding speech stimulus and rate the quality of the stimulus on a scale from 1 to 100 using a slider next to the button. A score within $[1, 20]$ is indicated as bad, $[21, 40]$ as poor, $[41, 60]$ as fair, $[61, 80]$ as good and $[81, 100]$ as excellent. The listener was instructed that the quality of the reference stimulus corresponded to a rating of "100" and other ratings should be given in a manner that they are all relative to each other. The listener was able to playback the reference and test stimuli more than once.

For half of the listeners in each of the two listener groups, the Noisy and Noisy-Enhanced conditions were tested in the first session while the Reverberant and Reverberant-Enhanced conditions were tested in the second session. For the other half of the listeners, the test sequence for the two sessions was reversed. At the beginning of each session, eight practice trials were run to familiarize the listener with the stimuli and the GUI.

If the Noisy and Noisy-Enhanced conditions were tested in the session, the practice trials included stimuli at the four different SNRs (from -5 to 15 dB), with and without the IRM-based speech enhancement. After the practice trials, the two experimental conditions were tested using two blocks of 40 trials. The order in which the two conditions were tested was counterbalanced across listeners. Within each block, 10 trials were run at each of the SNRs, in random order.

If the Reverberant and Reverberant-Enhanced conditions were tested in the session, the practice trials included stimuli at the four different values of T60 (i.e., 100, 200, 500 and 1000 ms), with and without the IRM-based speech enhancement. Following the practice trials, the two experimental conditions were tested in blocks, with each block containing 10 trials at each of the T60 values in random order.

No sentence was repeated in more than one trial, leading to a total of 176 unique sentences used in the current experiment. Due to the limited availability, one HI listener did not finish the Reverberant and Reverberant-Enhanced conditions and another HI listener did not finish the Noisy and Noisy-Enhanced conditions, resulting in seven listeners in the

HI group for each condition. For data collected from each session, a mixed-effect analysis of variance (ANOVA) was conducted to identify any significant effects of listener group, speech enhancement, SNR, T60, phase distortion and any significant interactions among them.

Two objective quality metrics, PESQ [87] and HASQI [93], were adopted to further investigate the correlations between objective measures and actual human ratings, especially on speech signals with distorted phase under noisy and reverberant conditions. PESQ is a widely adopted metric for speech quality assessment that gives outputs ranging from -0.5 to 4.5, while HASQI is a more recently proposed speech quality metric that includes a physiologically inspired model of human auditory system with predicted scores ranging from 0 to 1. The inclusion of this model allows HASQI to predict the perceived quality by both NH and HI listeners. The same stimuli (with NAL-R linear amplification and low-pass filtering) presented to each subject were given as inputs to both evaluation metrics.

4.3 Results and Discussion

Subjective quality ratings

The average ratings for the anchor and reference speech obtained from NH listeners are 73.8 ($SD = 4.13$) and 96.6 ($SD = 2.96$), respectively; and for HI listeners, the average ratings are 87.7 ($SD = 4.41$) and 97.2 ($SD = 1.82$) for anchor and reference speech, respectively.

Figures 4.3 and 4.4 show the subjective ratings for the four different degrees of phase distortion and the four SNRs in the Noisy and Noisy-Enhanced conditions. The error bars indicate $\pm$ one standard deviation. A mixed-design ANOVA shows significant main effects of listener group [$F(1, 15) = 6.35, p = .024$], enhancement [$F(1, 15) = 68.14, p < .001$], SNR [$F(1.7, 25.1) = 90.67, p < .001$, Greenhouse-Geisser corrected], and phase distortion [$F(1.1, 16.6) = 77.33, p < .001$, Greenhouse-Geisser corrected]. There are significant interactions between listener group and SNR [$F(1.7, 25.1) = 3.21, p = .032$, Greenhouse-

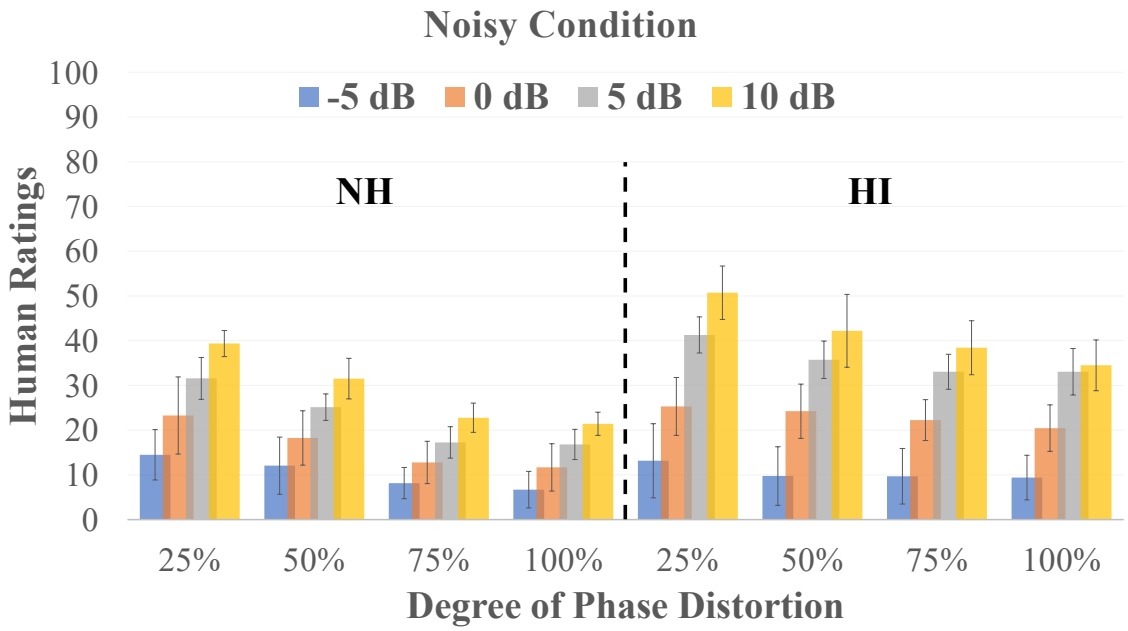

Figure 4.3: Human ratings under Noisy condition, NH listeners are represented in the **left** block and the HI listeners are shown in the **right** block.

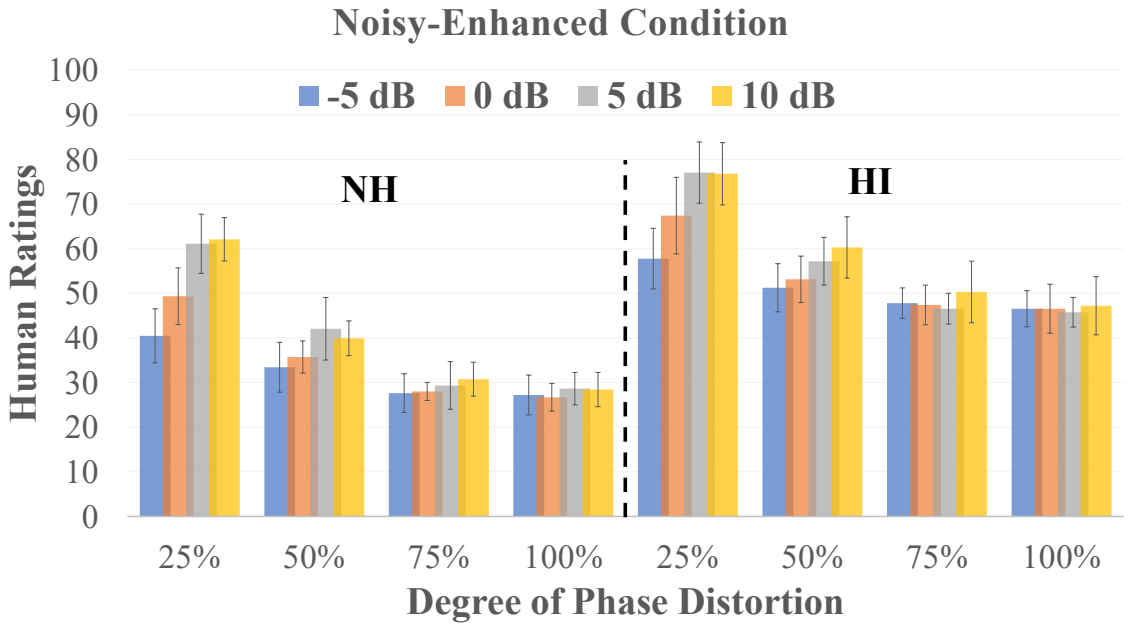

Figure 4.4: Human ratings under Noisy-Enhanced condition, NH listeners are represented in the **left** block and the HI listeners are shown in the **right** block.

Geisser corrected], between enhancement and SNR [$F(1.5, 22.7) = 24.54, p < .001$, Greenhouse-Geisser corrected], between enhancement and phase distortion [$F(1.3, 19.9) =$

$47.86, p < .001$, Greenhouse-Geisser corrected], and between SNR and phase distortion [$F(9, 135) = 25.08, p < .001$], as well as significant three-way interactions among listener group, enhancement, and SNR [$F(1.5, 22.65) = 5.59, p = .016$, Greenhouse-Geisser corrected] and among enhancement, SNR, and phase distortion [$F(3.8, 56.9) = 4.85, p < .001$, Greenhouse-Geisser corrected].

For the Noisy condition (see Figure 4.3), higher SNRs lead to higher quality ratings and greater degrees of phase distortions lead to lower ratings for both listener groups. The effects of SNR and phase distortion also interact with each other, with stronger effects of phase distortion observed at higher SNRs. When the noise level is high (i.e., at low SNRs), the phase distortion may be masked by the noise and become less noticeable to the listeners. The HI listeners tend to give higher ratings than the NH listeners, suggesting that they have higher tolerance for phase distortion and background noise.

For the Noisy-Enhanced condition (see Figure 4.4), the quality ratings are generally higher than those in the Noisy condition, indicating that the IRM-based speech enhancement improved perceived quality. Contrary to the strong effect of SNR in the Noisy condition, the effect of SNR is not reliably observed in the Noisy-Enhanced condition across all phase distortions. This suggests that following phase-insensitive speech enhancement the contribution from the background noise to the quality ratings is much reduced. Instead the quality ratings become dominated by phase distortion especially when the distortion amount is above 25% for the enhanced speech. For both listener groups, the quality rating decreases as the degree of phase distortion increases. In particular, the speech enhancement algorithm allows both the NH and HI listeners to better differentiate various degrees of phase distortion at low SNRs. Therefore, a speech enhancement algorithm that is capable of reducing phase distortions would likely lead to benefits in perceived speech quality for all listeners (with or without hearing loss), at both high and low SNRs.

Figures 4.5 and 4.6 show the subjective ratings for the four different degrees of phase distortion and the four T60 values in the Reverberant and Reverberant-Enhanced condi-

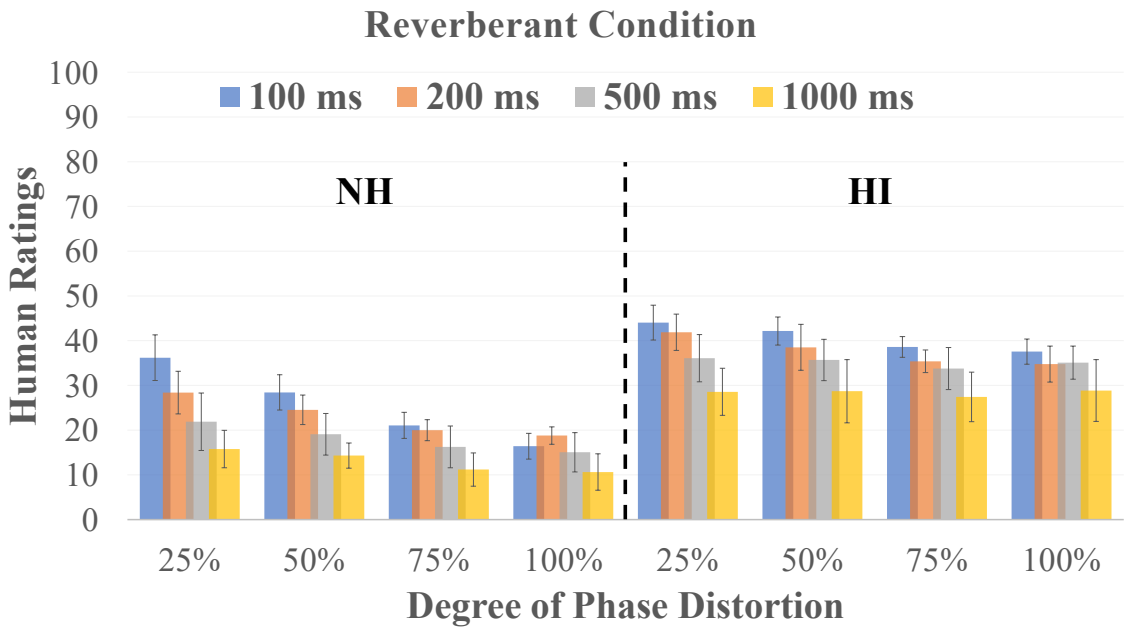

Figure 4.5: Human ratings under Reverberant condition, NH listeners are represented in the **left** block and the HI listeners are shown in the **right** block.

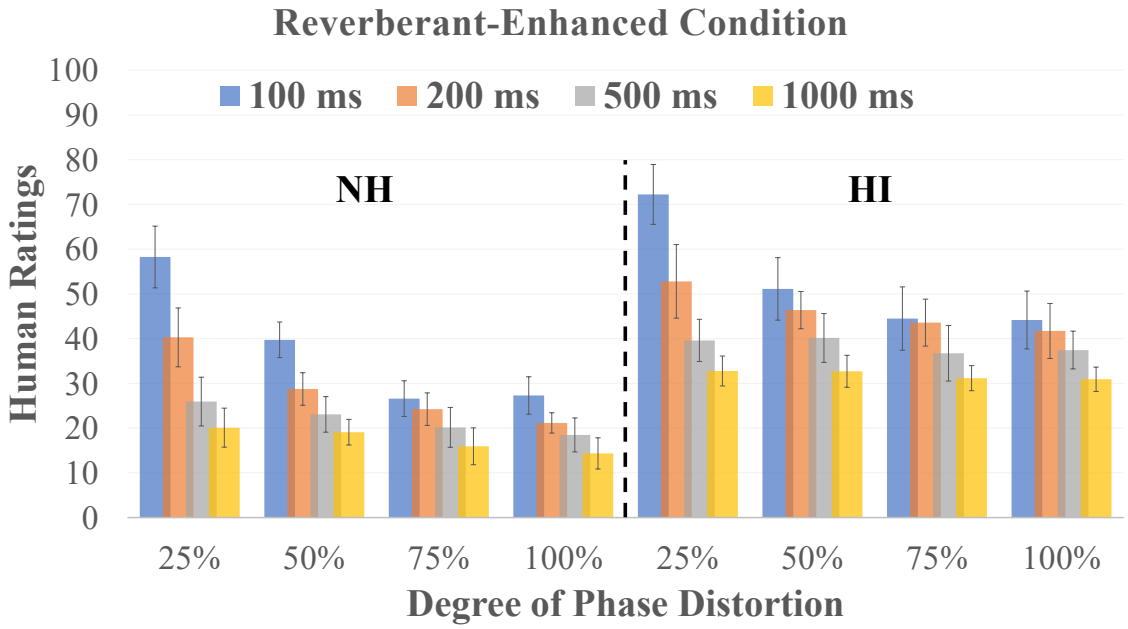

Figure 4.6: Human ratings under Reverberant-Enhanced condition, NH listeners are represented in the **left** block and the HI listeners are shown in the **right** block.

tions. The error bars indicate $\pm$ one standard deviation. A mixed-design ANOVA shows significant main effects of listener group [$F(1, 15) = 13.51, p = .002$], enhancement

[$F(1, 15) = 8.92, p = .009$], reverberation [$F(1.9, 27.8) = 48.60, p < .001$, Greenhouse-Geisser corrected], and phase distortion [$F(1.2, 18.1) = 51.74, p < .001$, Greenhouse-Geisser corrected]. There are significant interactions between enhancement and reverberation [$F(3, 45) = 12.12, p < .001$], between enhancement and phase distortion [$F(1.3, 19.0) = 9.87, p < .001$, Greenhouse-Geisser corrected], between reverberation and phase distortion [$F(2.8, 42.7) = 30.63, p < .001$, Greenhouse-Geisser corrected], as well as significant three-way interactions among enhancement, reverberation, and phase distortion [$F(3.6, 53.7) = 11.59, p < .001$, Greenhouse-Geisser corrected].

For the Reverberant condition (Figure 4.5), shorter reverberation times lead to higher quality ratings and greater degrees of phase distortions lead to lower ratings for both listener groups. The effects of reverberation time (T60) and phase distortion also interact with each other, with stronger effects of phase distortion observed for shorter reverberation times. It is possible that long reverberation times (e.g., 1000 ms) could mask the phase distortion introduced to the speech and make it less noticeable.

Similar to Noisy-enhanced condition, in the Reverberant-Enhanced condition (Figure 4.6), the quality ratings are generally higher than those in the Reverberant condition. This indicates that the IRM-based speech enhancement improved perceived speech quality for both listener groups. Following speech enhancement, the effect of T60 is still present for all degrees of phase distortion. This suggests that the IRM-based enhancement, when trained using non-reverberant noise and speech, is insufficient in removing the adverse effect of reverberation on speech quality. Moreover, following speech enhancement the effect of phase distortion becomes stronger for short reverberation times (100 and 200 ms). On the other hand, because of the remaining reverberation in the enhanced speech, the effect of phase distortion is absent for long reverberation times (1000 ms). The dependencies of the quality rating on phase distortion are similar between the NH and HI groups, suggesting that both NH and HI listeners could benefit from speech enhancement techniques that restore the phase spectrogram of the original clean speech.

Table 4.2: Pearson correlations between subjective and objective ratings at different conditions for NH and HI groups. The highest correlations are marked in **bold**. 'Orig.' indicates the original conditions before enhancement. 'Reverb.' stands for reverberant conditions.

Pearson Correlation Coefficients					
		PESQ		HASQI	
		Orig.	Enhanced	Orig.	Enhanced
Noisy	NH	**0.984**	**0.975**	0.956	0.755
	HI	**0.991**	**0.971**	0.976	0.871
Reverb.	NH	**0.993**	**0.992**	0.955	0.883
	HI	**0.990**	**0.986**	0.974	0.940

Correlation between objective and subjective scores

The subjective speech quality ratings obtained from the listening experiment were further compared to two objective metrics, PESQ and HASQI. The correlations between the subjective and objective speech quality are provided in Table 4.2.

PESQ yields the highest Pearson correlation coefficients for both NH and HI groups across all conditions (e.g., 0.984 and 0.975 for NH listeners under noisy and noisy-enhanced conditions). HASQI achieves similar but slightly decreased Pearson correlation coefficients for the Noisy and Reverberant conditions (e.g., 0.956 and 0.955 vs. 0.984 and 0.993 for NH listeners) compared to PESQ. This is contrary to the reports in [150, 172], where HASQI yields better correlation than PESQ for enhanced speech. We infer that the frequency range for evaluation in our experiments (note that all testing stimuli were low-pass filtered at 4 kHz to avoid influences from high-frequency hearing loss in the HI group) might be one of the factors that contribute to this phenomenon. Since PESQ is designed for measuring signal quality transmitted through a narrow bandwidth (3.1 kHz) [87], while HASQI is designed for a wider frequency band (12 kHz) [93]. The low-pass filtering applied in our experiment makes PESQ more fitted to this scenario and previous studies adopted wide-band stimuli that are more suitable for HASQI.

4.4 Summary

In this chapter, we investigated the influence of phase distortion on the perceived speech quality for both NH and HI listeners. Hearing-impaired listeners tend to provide higher ratings for the same speech stimulus, corrupted by either background noise or reverberation, than NH listeners. The quality rating depends on the degree of phase distortion in a similar way. Following phase-insensitive speech enhancement, HI and NH listeners can differentiate the degree of phase distortion that remained in the enhanced speech, indicating potential benefits from phase-sensitive enhancement techniques. We assume that these HI listeners may notice phase distortions because (1) They have good TFS sensitivity, or (2) TFS and phase cues are weighted higher for quality tasks as compared to recognition tasks. Between two objective speech quality metrics, PESQ provides closer correlations to the subjective ratings than HASQI, especially for the enhanced speech. This is likely caused by the low-pass filtering introduced in the experiments, which makes PESQ a better fit.

CHAPTER 5
REDUCING COMPLEXITY OF SPEECH ENHANCEMENT SYSTEM USING DIFFERENT PHASE ESTIMATION STRATEGIES ACROSS SPECTRAL REGIONS

Experimental results from Chapter 4 suggest potential benefits of phase-aware speech enhancement algorithm to HI (and NH) listeners. Many existing speech enhancement algorithms allocate the same computation resources to all frequency regions for phase estimation [173, 174, 55, 68, 71, 70, 75], however, it is not clear if the estimated phase is equally important to speech quality for human listeners. How various frequency regions contribute to speech understanding, i.e., speech intelligibility, has been extensively studied over the past decades [175, 176, 177, 178]. As a result, the band importance function (BIF) has been widely used to characterize the relative importance of different frequency bands to speech intelligibility. It is found that the most important spectral regions for speech understanding are at low frequencies, between 1600 and 2000 Hz [179, 180, 181]. On the contrary, little is known about the underlying mechanism of speech quality judgment and how does phase from different frequency bands contribute to the perceived quality. From a practical point of view, understanding the importance of frequency bands of phase on speech quality could help conserve computation resources for more perceptually important regions and may potentially reduce the complexity of speech enhancement models.

To begin, we hypothesized that phase information from different spectral regions contribute to speech-quality judgment unequally. To verify, we examined the band importance of phase estimation on speech quality by systematically removing the phase information of high-frequency regions in Section 5.1. Two versions of enhanced speech were generated from a phase-insensitive and a phase-aware speech enhancement model. The low-frequency portion of the phase-aware enhanced speech was merged with the high-frequency

portion of the phase-insensitive enhanced speech to generate the speech stimuli with different phase information across spectral regions. Then, we conducted a pairwise comparison listening study between full-band phase-aware and the filtered-merged speech, to investigate at which frequency the benefits from phase enhancement start to diminish. We further adopted HASQI as an objective measure to simulate responses from HI listeners. Experimental results indicated that estimating phase at lower-frequency regions are mostly important for speech quality in NH listeners. We infer that NH listeners are not sensitive to phase distortions at high-frequency regions and speech enhancement algorithms should conserve more computation resources for low-frequency phase-aware processing.

We further proposed a novel hybrid speech enhancement framework in Section 5.2. Specifically, the proposed hybrid-net adopts different strategies dealing with phase estimation in different frequency regions. A phase-aware sub-network is adopted for low-frequency regions and another phase-insensitive sub-network is applied for high-frequency regions, since we hypothesized that phase estimation is mostly important for speech quality at low frequency regions. Systems were evaluated on a simulated speech dataset using both objective (i.e., HASQI) and subjective measures (i.e., pairwise comparison). The proposed hybrid-net significantly improves the model compatibility for low-resource platforms (i.e., reduced network size and number of computation involved) while achieving comparable performance to the original full-band phase-aware speech enhancement model.

Finally, we provide hypotheses for the observed phenomena and explore the potential benefits of the proposed hybrid-net in HI listeners. We further discuss some limitations and future works of the current hybrid-net. Part of the work presented in this chapter has been submitted to INTERSPEECH, 2022.

5.1 Importance of frequency band for phase estimation

5.1.1 Introduction

Speech signals are often degraded by unwanted background noise or interfering speakers in real-world environments and thus pose a challenge especially for hearing-impaired (HI) listeners. Speech enhancement algorithms are often adopted to alleviate this speech-in-noise problem, and have been widely implemented as important front-ends for many speech communication systems including digital hearing-aid devices [11, 12, 13, 14, 15]. Early speech enhancement algorithms only aim to recover the magnitude of the speech while leaving the phase component unchanged, as it had been believed that phase is not as important as speech magnitude [182].

Paliwal et al. [67] showed, however, that estimating the phase component for speech enhancement could improve the perceived quality for NH listeners and this has led to many promising works on phase-aware speech enhancement methods that demonstrated improvements in both objective and subjective speech quality [183, 55, 68, 70, 71]. Additionally, results from [164] further indicated the potential benefits from phase-aware speech enhancement algorithms for HI listeners.

The importance of different frequency bands to speech intelligibility has been investigated by many studies [184, 185], where different bands were found to contribute unequally to speech intelligibility [178]. Specifically, the low-frequency regions (e.g., below 1600 to 2000 Hz as indicated in [179, 180, 181]) have been found to contribute more than high-frequency regions. On the contrary, less is known regarding the importance of phase estimation at different frequency bands to speech quality. Phase is strongly correlated with the temporal fine structure (TFS) of the speech signal as the TFS describes the progression of phase across consecutive time frames, which is important to human perception on voicing, intonation, and speaker gender [66]. Meanwhile, it is reported that phase locking to the TFS of stimuli is limited to frequencies below 1.5 kHz in human listeners [186, 187],

and similar cases were found in animal auditory nerve cells (3.5 kHz in guinea pigs [188] and 5 kHz in cats [189]). This may suggest that estimating phase in high frequency regions is not beneficial.

Current phase-aware speech enhancement algorithms treat the phase equally important across spectral regions. Although improvements were observed in both objective and subjective speech quality, it is not clear whether allocating equal computational resources across the spectrum is the most efficient approach. It might be better to reserve more computational power in low-frequency regions for phase-aware processing. Moreover, phase-aware speech enhancement algorithms usually consist of more complicated network structures in order to deal with phase estimation, which consumes more computational power. If phase estimation proves to be unimportant at high-frequency bands, then future developments on speech enhancement algorithms could assign more computational power to the lower frequencies for phase estimation and fewer resources at higher frequencies.

To investigate the importance of estimating phase at different frequency bands, we first generate speech samples by combining the enhanced speech produced from two speech enhancement algorithms (i.e., phase-aware and phase-insensitive), which are filtered with different cutoff frequencies. More specifically, the low-frequency components of the phase-aware enhanced speech are combined with the high-frequency components of phase-insensitive enhanced speech. If listeners do not notice a difference after replacing the high-frequency components with phase-insensitive speech, then we may infer that phase information above that certain frequency does not significantly contribute to speech quality. We conduct a pairwise comparison listening study between filtered-merged and full band phase-aware enhanced speech to find the cutoff frequency where the benefits from phase-aware speech enhancement start to diminish. In addition, we adopt the hearing-aid speech quality index (HASQI) [93] to simulate speech quality perceived by human listeners. Experimental results suggest that the estimated phase becomes less important at higher frequencies and the cutoff frequency varies across different listeners.

The rest of this section is organized as follows. Section 5.1.2 introduces the speech enhancement models we used to generate the enhanced speech. The listening study is described in Section 5.1.3 and experimental results are reported in Section 5.1.4. Lastly, we provide a summary in Section 5.1.5.

5.1.2 Speech enhancement systems

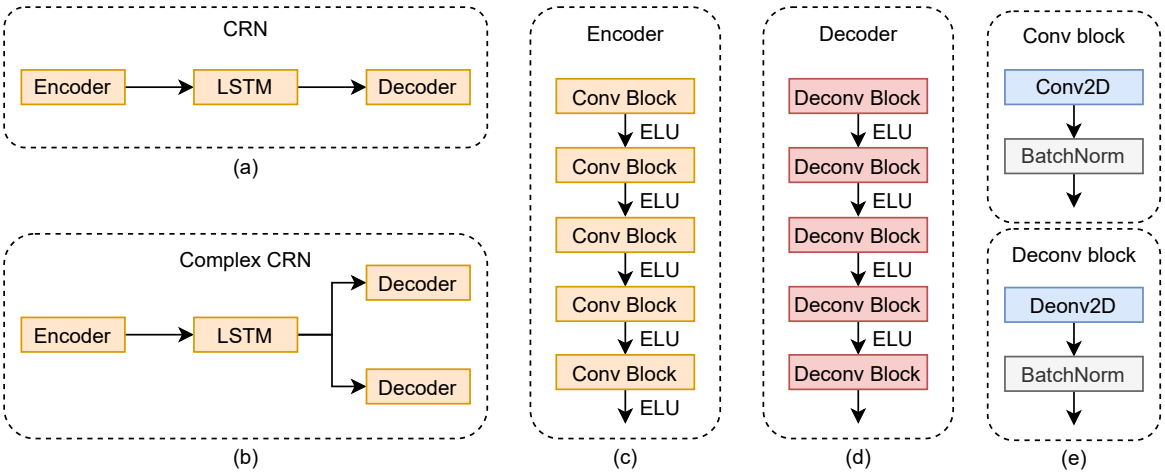

Figure 5.1: Network structures of CRNs for phase-insensitive and phase-aware speech enhancement. (a) Phase-insensitive CRN, (b) Complex (i.e., phase-aware) CRN, (c) Encoder of CRN, (d) Decoder of CRN, (e) Convolution and deconvolution blocks.

We adopt two speech enhancement models based on convolutional recurrent network (CRN) [97, 71] for phase-insensitive and phase-aware speech enhancement, respectively. As illustrated in Figure 5.1 (a), the phase-insensitive CRN features an encoder-decoder-like architecture, where an LSTM block (with two layers of unidirectional LSTM [154]) is employed between the encoder and decoder to help capture the temporal features. Furthermore, skip connections (not shown) are used to concatenate the output of each encoder layer to the input of the corresponding decoder layer [97], i.e.,

$$\text{Decoder}_i^l = \text{Concatenate}(\text{Encoder}_o^l, \text{input}_i^l), \tag{5.1}$$

where Decoder_i^l and input_i^l denote the input of decoder and input without skip connection at the l-th layer, respectively. Encoder_o^l denotes the output of the corresponding l-th encoder layer.

The phase-insensitive CRN takes in the magnitude spectrogram of a noisy speech and directly estimates the enhanced magnitude spectrogram (i.e., last deconvolution block with ReLU activation [152]). During reconstruction of the time-domain enhanced speech, the original noisy phase is used. This model is used to generate enhanced speech that does not include phase enhancement.

A complex version of CRN [71] is implemented as the phase-aware speech enhancement model, also illustrated in Figure 5.1 (b). It features the same network configuration as the phase-insensitive CRN except that it contains two separate decoders in order to estimate the real and imaginary parts of the speech spectrogram. Although better speech quality can be achieved, one drawback is that it requires additional computation resources. Note that linear activation is adopted for the last deconvolution blocks in the two decoders for complex CRN.

System setup

For both CRNs, the encoders and decoders consist of five convolution/deconvolution blocks that further contain 2D-convolutions and deconvolutions followed by batch normalization. An exponential linear unit (ELU) [190] is used as the activation function between successive convolution and deconvolution blocks. The detailed network configurations for the implemented CRNs are provided in Table 5.1. Both the kernel size and stride size have a dimension of $time \times frequency$. The input and output channels of the convolution layer are specified in the parentheses.

Both speech enhancement systems are trained with 60 epochs or until convergence, where ADAM [191] optimizer is used with a learning rate of $1e^{-3}$. A mini-batch size of 24 is used during training.

Table 5.1: Network configuration of CRNs, note the number of input channel(s) for the first convolution layer is dependent on the network type, one channel (i.e., magnitude spectrogram) is used for CRN and two channels (i.e., real and imaginary spectrograms) are used for complex CRN.

Layer	**Kernel size**	**Stride size**	**# of channels**	**# of units**
Conv2d_1	(1×3)	(1×2)	(1 or 2,16)	-
Conv2d_2	(1×3)	(1×2)	(16,32)	-
Conv2d_3	(1×3)	(1×2)	(32,64)	-
Conv2d_4	(1×3)	(1×2)	(64,128)	-
Conv2d_5	(1×3)	(1×2)	(128,256)	-
LSTM_1	-	-	-	1024
LSTM_2	-	-	-	1024
Deonv2d_1(1r/1i)	(1×3)	(1×2)	(512,128)	-
Deonv2d_2(2r/2i)	(1×3)	(1×2)	(256,64)	-
Deonv2d_3(3r/3i)	(1×3)	(1×2)	(128,32)	-
Deonv2d_4(4r/4i)	(1×3)	(1×2)	(64,16)	-
Deonv2d_5(5r/5i)	(1×3)	(1×2)	(32,1)	-

Materials

A total of 1440 clean speech utterances (i.e., 720 utterances for each gender) from IEEE corpus [157] are used. In the training set, 80% of them are mixed with eight different noises, including multi-talker babble, factory, cafeteria, thunderstorm, washing machine, vacuum, train and engine noises from AzBio [160], NOISEX-92 [192] and ESC-50 databases [161]. The signal-to-noise ratios (SNRs) in the training set range from -6 to 0 dB with a step size of 1 dB, resulting in 64512 mixtures.

Meanwhile, 10% of the speech signals are used for each of the development and testing sets. In the development set, we mix the speech with factory noise at -5 dB SNR. In the testing set, the speech utterances are mixed with babble and cafeteria noises at SNRs ranging from -5 to 5 dB with a step of 5 dB (864 mixtures in total), similar to the setup described in [71]. All signals are resampled to 16 kHz before further processing. We use a 320-point FFT together with a 20 ms hamming window (10 ms hop size) for STFT and iSTFT calculations.

5.1.3 Listening study

Stimuli

Listeners conducted paired comparisons in terms of perceived quality. In each trial, a pair of stimuli were presented, one of which was the enhanced speech signal produced by the phase-aware model. The other stimulus was generated by combining the estimates from the phase-aware model in low frequencies with the estimates from the phase-insensitive model in high frequencies. With varying cutoff frequencies, the lower frequency portion of the phase-insensitive enhanced speech is merged with the high-frequency portion of the phase-aware enhanced speech, in order to investigate where the benefits of phase estimation start to diminish.

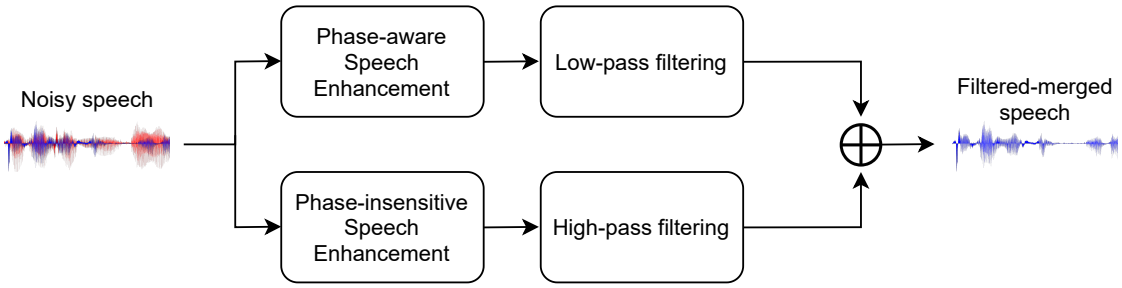

Figure 5.2: Illustration of the filter-and-merge process.

Specifically, for a noisy speech sample, two enhanced speech signals were first generated from the phase-insensitive and phase-aware models. The enhanced speech from the phase-aware model was low-pass filtered and the output from the phase-insensitive model was high-pass filtered using the same cutoff frequency (as illustrated in Figure 5.2). The filtered signals were then mixed to be used as the stimulus for the listening study. Several cutoff frequencies are used, including 0 Hz (i.e., using original noisy phase for reconstruction, the phase-insensitive model, denoted as 'Mag.'), 250 Hz, 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz. These stimuli will be referred to as filtered-merged speech.

We use 'fdesign.lowpass' and 'fdesign.highpass' functions in MATLAB to design the

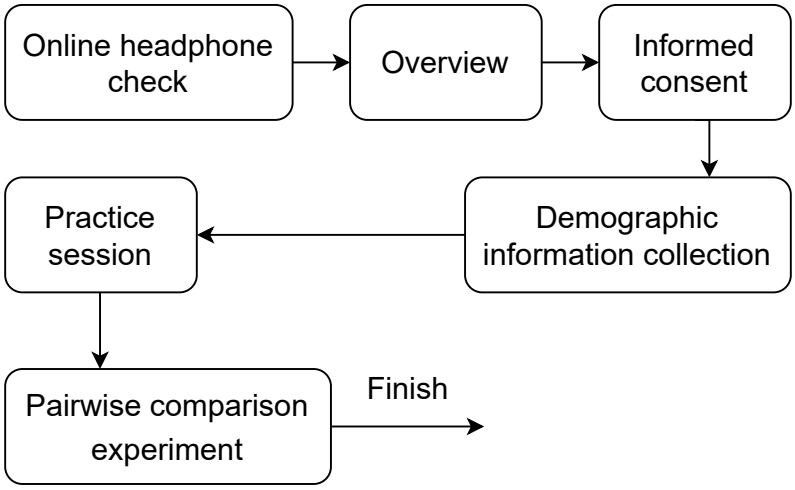

Figure 5.3: Workflow of the online listening study.

low-pass and high-pass filters, where the filter order is set to 50, with Butterworth type and a 70 dB attenuation in the stop band. For each cutoff frequency, there are 20 enhanced speech utterances randomly selected from the testing set. Half of them with original SNR at -5 dB and the other half at 5 dB.

Subject recruitment

A total of 20 participants were recruited (12 males and 8 females), aging from 23 to 58 years old (avg. 35.6 years) via Amazon Mechanical Turk. All participants were native American English speakers and self-reported to have normal hearing. The current study was conducted following the Declaration of Helsinki and approved by the Institutional Review Board (IRB) at Indiana University and the University of Washington. Informed consents were obtained from all participants before the data collection, where information on experiment description, compensation, potential risks, voluntary nature of study, and data privacy is included.

Procedure

The general protocol of the online listening study is depicted in Figure 5.3. An online headphone check [193] stage was first included to ensure headphones were worn by all participants, where the Huggins Pitch (HP) [194, 195] was utilized. Huggins Pitch is a perceptual phenomenon which could only be detected when stimuli (with interaural phase difference) are presented dichotically. During the headphone check stage, the order of HP was randomly hidden in the presented stimuli that consists of three noises. The headphone check would fail if speakers were used or the headphone was only worn on one ear.

In the next stage, the participant was provided with an overview of the listening study. Followed by the collection of informed consent and demographic information from the participant. Before starting the actual listening experiment, a practice stage was provided to familiarize the subjects with the experimental task. At the beginning of the practice stage, the participant was instructed to adjust the volume to a loud enough and comfortable level before proceeding. This volume was then used throughout the rest of the experiment to ensure all stimuli were presented at the participant's most comfortable listening level. Five practice trials were then run. On each practice trial, the participant compared the quality between a pair of full band phase-insensitive and phase-aware enhanced speech signals. The participant had the opportunity to replay each of the stimuli unlimited times. An example user interface is provided in Figure 5.4.

The main experimental stage of pairwise comparison consists of 120 pairs (i.e., 20 repetitions $\times$ 6 cutoff frequencies) of stimuli between full band phase-sensitive enhanced speech and filtered-merged speech. The order of these evaluation pairs are randomly generated per each subject. The order of the two stimuli within each pair is also permuted across different trials. All participants are asked to sit in a quiet environment during the listening study.

5.1.4 Results and analysis

Objective results

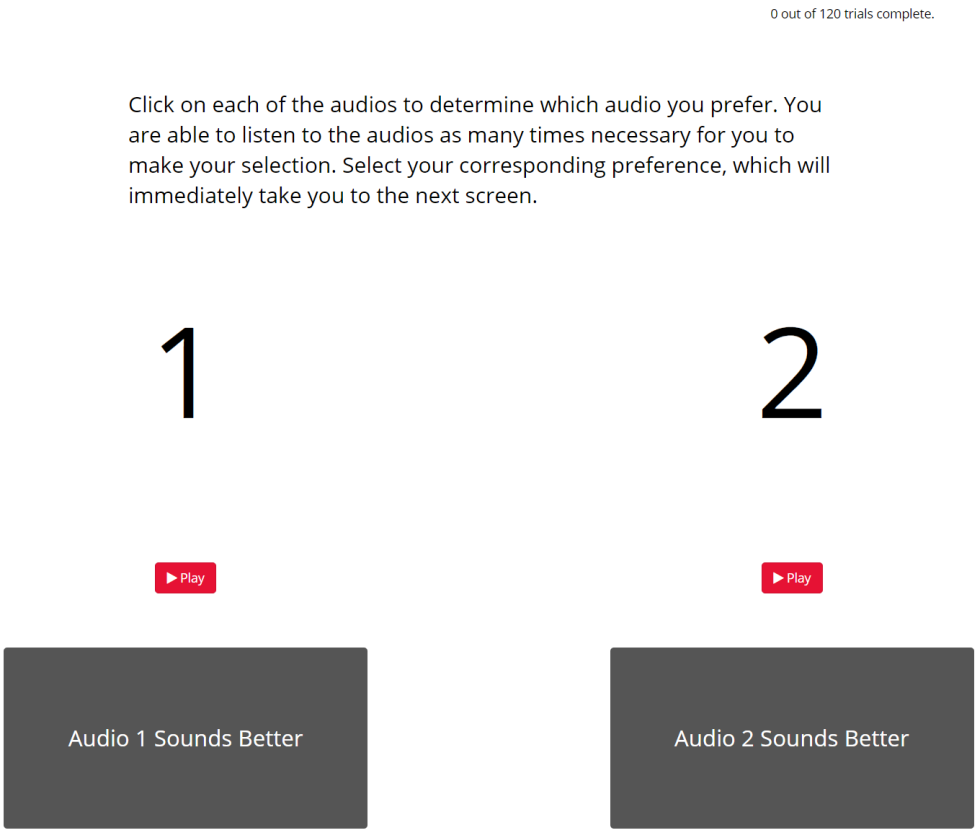

Figure 5.4: Example user interface for the online pairwise comparison experiment. The participant is asked to select the audio that has better perceived quality.

We first provide the HASQI results for simulated NH listeners (i.e., auditory thresholds set to 0 dB HL) in Table 5.2. We can observe that both speech enhancement systems (i.e., phase-insensitive and phase-aware) can significantly improve the speech quality, where the phase-insensitive model achieves relative 117% improvements (i.e., 0.163 vs. 0.075) while the phase-aware model leads to even better performance (i.e., 0.212 vs. 0.163). Among these two speech enhancement models, we observe additional benefits from estimating phase components (i.e., relative 30% improvement), which is consistent with previous findings [71, 72].

The HASQI scores for filtered-merged speech is presented in the bottom section of Table 5.2. A two-sided alternative hypothesis t-test showed significant differences between filtered-merged speech (average HASQI scores across SNRs) at 1000 Hz and 2000 Hz

Table 5.2: HASQI scores for NH listeners, performance from phase-insensitive and phase-aware systems is provided. The performance for the merged version of enhanced speech with different cutoff frequencies is also included. **Bold** font indicates the best performance.

	HASQI scores			
Type	-5 dB	0 dB	5 dB	Avg.
Mixture	0.026	0.064	0.136	0.075
Phase-insensitive	0.102	0.166	0.221	0.163
Phase-aware	**0.146**	**0.221**	**0.270**	**0.212**
250 Hz	0.106	0.171	0.227	0.168
500 Hz	0.109	0.175	0.232	0.172
1000 Hz	0.115	0.184	0.239	0.179
2000 Hz	0.133	0.204	0.254	0.197
4000 Hz	0.144	0.219	**0.270**	0.211

[$t(286) = 4.23$, $p < .001$], filtered-merged speech at 2000 Hz and 4000 Hz [$t(286) = 3.07$, $p = .002$]. No significant differences were found between phase-insensitive and filtered-merged speech at 250 Hz [$t(286) = 1.28$, $p = .203$], filtered-merged speech at 250 Hz and 500 Hz [$t(286) = 1.07$, $p = .286$], filtered-merged speech at 500 Hz and 1000 Hz [$t(286) = 1.68$, $p = .093$], filtered-merged speech at 4000 Hz and phase-aware enhanced speech [$t(286) = 0.32$, $p = .748$].

It is observed that the average score increases with increasing cutoff frequencies (e.g., 0.168 at 250 Hz vs. 0.172 at 500 Hz). This suggests that better speech quality can be obtained by estimating phase at a broader band. Meanwhile, we found that between the full-band phase-aware and filtered-merged speech at 4000 Hz, there is not much difference in speech quality (i.e., 0.212 vs. 0.211), which indicates estimating phase at higher frequencies is not that important compared to low-frequency phase estimation. We also notice that for cutoff frequency at 4000 Hz, the merged version achieves similar performance with original full band phase-aware enhanced speech (i.e., 0.211 vs. 0.212), which suggests that estimating phase components above 4000 Hz could be unnecessary.

Subjective results

In order to investigate the influence of different cutoff frequencies and background SNRs, we first compute the d-prime score, d', which is a measure of psychophysical performance [196]. It is defined as

$$d' = z(H) - z(F), \tag{5.2}$$

where z denotes the z-transform (i.e., inverse Gaussian distribution), H and F represent the hit rate and false alarm rate, respectively. A hit is when the first stimuli is the full-band phase-aware enhanced speech and it is chosen as the preferred stimuli. Similarly, a false alarm is when the second stimulus is the filtered-merged speech and it is chosen as the preferred one, as suggested in [197]. A large d' value indicates better discriminability and a positive d' here indicates preference towards full-band phase-aware enhanced speech for the presented data. The purpose of estimating d' is to derive the discriminability between the full-band phase-aware enhanced speech and filtered-merged speech. The estimated d' will be used as the dependent variable to study the effects of cutoff frequency and SNR.

Figure 5.5 provides the d' values for each condition ('Mag.' indicates the phase-insensitive model). Note that all conditions are compared against the stimuli generated by the full-band phase-aware speech enhancement model. A higher d' value would indicate the full-band phase-aware enhanced speech is more preferred. It is observed that the d' value decreases as the cutoff frequency increases.

A repeated measures analysis of variance (ANOVA) is conducted, and it shows significant main effects of cutoff frequency [$F(3.06, 58.10) = 48.58$, $p < .001$, $\eta_p^2 = .719$, Greenhouse-Geisser corrected], and background SNR [$F(1, 19) = 4.68$, $p = .043$, $\eta_p^2 = .198$]. Pairwise comparisons further indicate there is no significant difference between 0 Hz (denoted as 'Mag.') and 250 Hz [$t(19) = 2.40$, $p = .406$, Bonferroni corrected]. Significant differences in d' values are observed between Mag. and 500Hz [$t(19) = 6.40$,

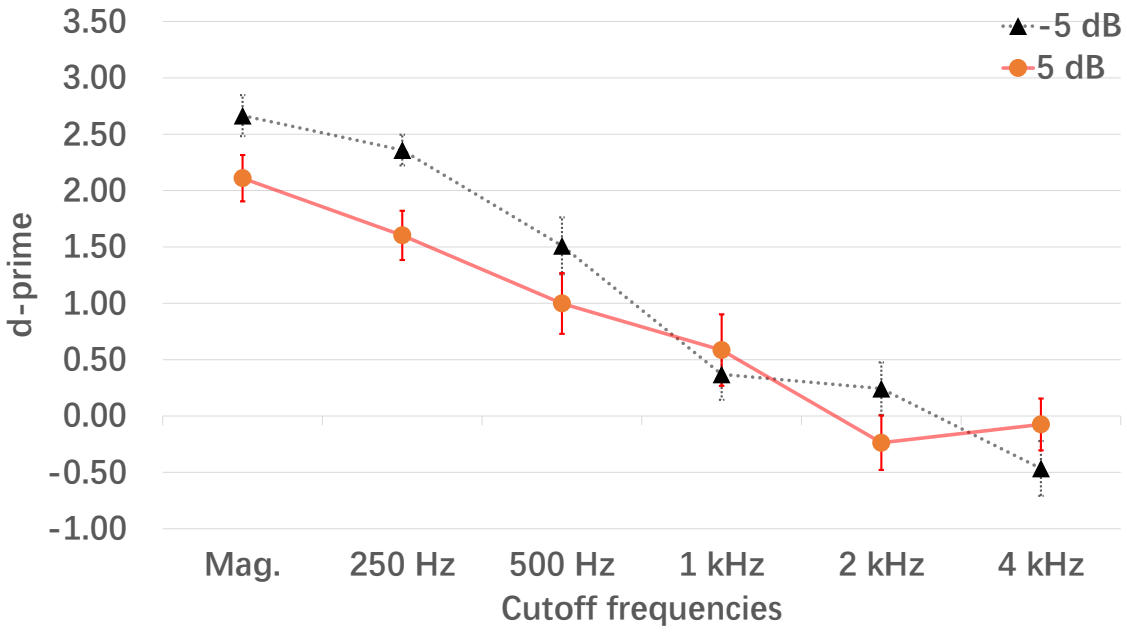

Figure 5.5: d' values across conditions. Error bars indicate the $\pm$ standard errors. Two background SNRs (before enhancement) are included.

$p < .001$, Bonferroni corrected], 1000 Hz [$t(19) = 12.17$, $p < .001$, Bonferroni corrected], 2000 Hz [$t(19) = 10.65$, $p < .001$, Bonferroni corrected], and 4000 Hz [$t(19)$ =10.07, $p < .001$, Bonferroni corrected]. Similarly, at 4000 Hz, significant differences are observed when compared to Mag. [$t(19) = -10.07$, $p < .001$, Bonferroni corrected], 250 Hz [$t(19) = -13.09$, $p < .001$, Bonferroni corrected], and 500 Hz [$t(19) = -6.87$, $p < .001$, Bonferroni corrected]. There are no significant differences between 4000 Hz and 1000 [$t(19) = -2.33$, $p = .463$, Bonferroni corrected], or 2000 Hz [$t(19) = -1.18$, $p = 1.0$, Bonferroni corrected].

We also found significant interactions between cutoff frequency and background SNR [$F(5, 95) = 3.73$, $p = .004$, $\eta_p^2 = .164$], where the effect of SNR is stronger at lower cutoff frequencies (as reflected in Figure 5.5). This suggests that estimating low-frequency phase information is more important at lower SNRs. On the other hand, estimating high-frequency phase information does not seem to improve speech quality compared to the phase-insensitive model, regardless of SNR. We also found that human listeners cannot re-

liably differentiate the full-band phase-aware enhanced speech and filtered-merged speech with high cutoff frequencies.

5.1.5 Summary

In this section, we investigated the importance of estimating phase at different frequency bands. Experimental results suggest that the benefits of phase estimation mostly come from low-frequency regions, and that phase estimation at high-frequency bands is less important for NH listeners.

5.2 Spectrally focused phase-aware enhancement: a hybrid speech enhancement framework

5.2.1 Introduction

In Section 5.1, we examined the influence of estimating phase at different frequency bands where both objective and subject results suggest that the benefits of a phase-aware speech enhancement algorithm are mostly coming from lower-frequency regions. Inspired by this finding, we introduce a novel hybrid speech enhancement framework in this section. The proposed framework consists of a phase-aware CRN that estimates low-frequency speech and another phase-insensitive CRN that processes the high-frequency components. Specifically, the low-frequency speech components are processed by a phase-aware CRN and then merged with the high-frequency speech components processed by another phase-insensitive CRN. This is the first time that we propose to use different strategies dealing with phase estimation across frequencies for speech enhancement.

The proposed hybrid framework is evaluated on a simulated dataset and compared with several baseline systems, including full-band phase-aware and phase-insensitive CRNs. HASQI simulation results suggest that for NH listeners, our proposed hybrid framework achieves comparable performance with original full-band phase-aware CRN. Moreover, the proposed hybrid speech enhancement system significantly reduces the model complexity

in terms of the model size and multiply–accumulate operations (MACs) involved. A pairwise comparison listening study is further conducted on human listeners, where subjective results showed consistent patterns with HASQI scores.

The rest of this section is organized as follows. We describe the experimental setup in Section 5.2.3. The simulation results are provided in Section 5.2.4 and the listening study together with subjective results are introduced in Section 5.2.5. Finally, we give a summary in Section 5.2.6.

5.2.2 Network architecture

The architecture of the proposed hybrid speech enhancement framework (denoted as hybrid-net) is shown in Figure 5.6. The noisy speech is first transformed into the T-F domain via STFT and split into separate halves (with cutoff frequency at 4 kHz). The real and imaginary parts of the lower frequency (i.e., 0 to 4 kHz) spectrogram are fed into a sub-network based on phase-aware CRN, on the other hand, the high frequency (i.e., 4 to 8 kHz) magnitude spectrogram is fed into another sub-network based on phase-insensitive CRN. Next, the two sub-networks encode the input features separately with additional residual connection (i.e., addition) to enable information shared across sub-networks at encoders, recurrent layers and decoders. Then, the phase-aware CRN estimates the real and imaginary parts for the lower-frequency half of the spectrogram, while the phase-insensitive CRN predicts the magnitude spectrogram for the higher-frequency half. Lastly, the estimated spectrograms (noisy phase used for the higher-frequency end) are resynthesized to time-domain enhanced speech using the iSTFT.

The configuration of the proposed hybrid-net is provided in Table 5.3. Where the 'Conv2d' and 'LSTM' denote the 2D convolution and LSTM layers in the encoder and recurrent block in both sub-networks, respectively. 'Deconv2d' represents the 2D transpose convolution layers in the decoder of both sub-networks, and the same configuration is used for the real and imaginary decoders of the phase-aware sub-network.

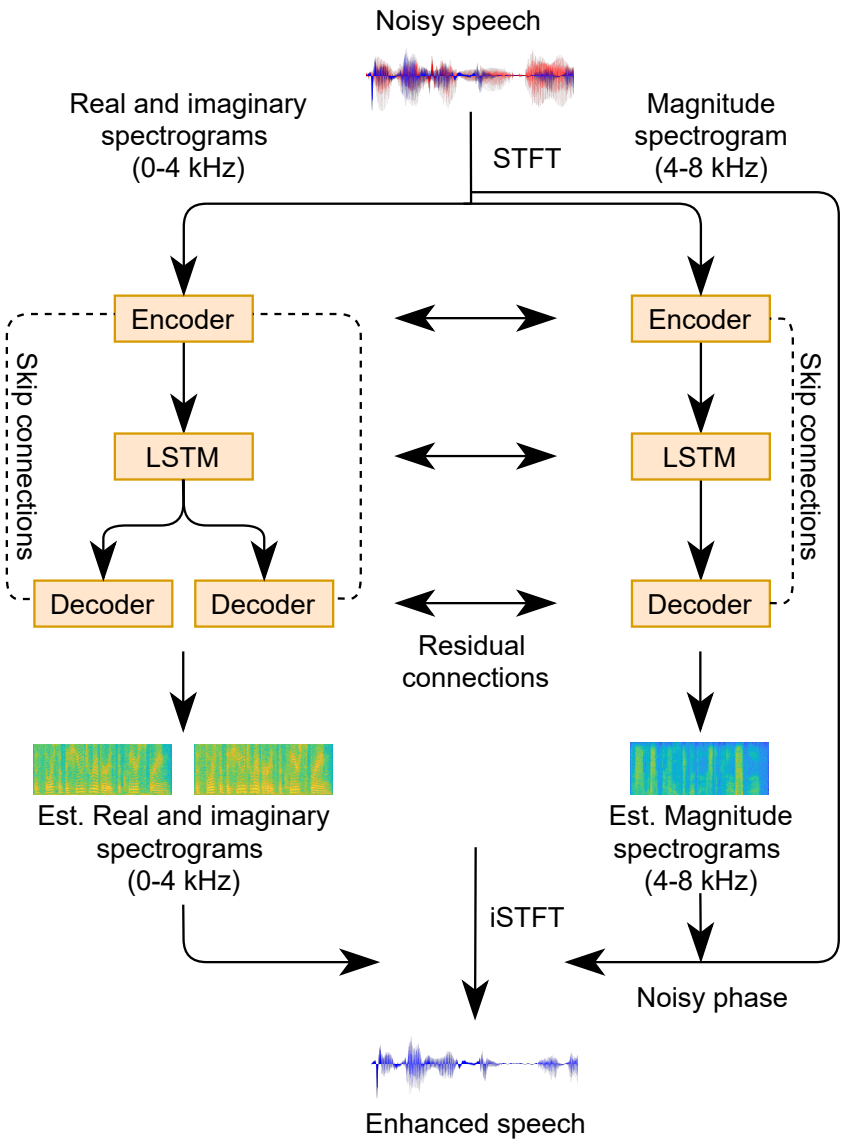

Figure 5.6: Network architecture of the proposed hybrid speech enhancement system.

Table 5.3: Network configuration of the proposed hybrid-net.

Layer	Kernel size	Stride size	# of channels	# of units
Conv2d_1	(1,3)	(1,2)	(1or 2,16)	-
Conv2d_2	(1,3)	(1,2)	(16,32)	-
Conv2d_3	(1,3)	(1,2)	(32,64)	-
Conv2d_4	(1,3)	(1,2)	(64,128)	-
Conv2d_5	(1,3)	(1,2)	(128,256)	-
LSTM_1	-	-	-	512
LSTM_2	-	-	-	512
Deconv2d_1	(1,3)	(1,2)	(512,128)	-
Deconv2d_2	(1,3)	(1,2)	(256,64)	-
Deconv2d_3	(1,3)	(1,2)	(128,32)	-
Deconv2d_4	(1,3)	(1,2)	(64,16)	-
Deconv2d_5	(1,3)	(1,2)	(32,1)	-

5.2.3 Experimental setup

Speech materials and network training

The same speech materials from Section 5.1.2 are used for these experiments, where the speech utterances are from IEEE sentences and the noises are extracted from AzBio, NOISEX-92 and ESC-50 databases. There are, in total, 64512 mixed utterances in the training set with SNRs ranging from -6 to 0 dB. The validation and testing sets contain 144 and 864 mixed utterances, respectively. The sampling rate is set to 16 kHz for all signals. The hybrid-net is trained with 60 epochs (or until convergence). ADAM optimizer is used with a learning rate of $1e^{-3}$. The mini-batch size is set to 24.

Objective evaluation

We use HASQI to simulate the speech quality perceived by NH listeners, in order to compare the performance of the proposed hybrid-net with the original phase-aware and phase-insensitive CRNs.

5.2.4 Simulation results

Computational efficiency

Table 5.4: Network statistics of different speech enhancement systems.

	Network Statistics		
System	# of Param. (M)	MACs (G)	Inference Speed (ms)
Phase-insensitive CRN	17.19	50.87	**8.7**
Phase-aware CRN	17.45	59.22	9.0
Hybrid-net	**9.46**	**36.28**	14.2

To evaluate whether the hybrid enhancement framework reduces the computational resources (e.g., model size, computations involved) compared to the original full-band phase-aware CRN. We present the model size (i.e., number of parameters), MACs[1] and inference speed (i.e., average running time for processing 1 s of audio input) as metrics for model complexity in Table 5.4. The inference speed is measured using a single Nvidia Tesla V100 GPU, where we set the batch size to 1.

The proposed hybrid-net is about half the size (54.2%) of the original phase-aware CRN (i.e., 9.46 M vs. 17.45 M) and achieves relative 38.7% reduction in MACs (i.e., 36.28 G vs. 59.22 G). The results suggest that the proposed hybrid-net has better compatibility for low-resource devices, such as digital hearing-aids. However, the inference speed for the proposed hybrid-net is slower than the other speech enhancement systems. This is likely caused by the two computation flows and their interactions for the low-frequency and high-frequency processing. This gap could be potentially alleviated by optimizing the parallel process between the two computation flows.

Objective speech quality predictions

The HASQI scores for simulated NH listeners are presented in Table 5.5. A two-sided alternative hypothesis t-test showed significant differences between phase-insensitive

[1]Measured using THOP, https://github.com/Lyken17/pytorch-OpCounter

Table 5.5: HASQI scores for NH listeners. The best performance is marked with **bold** font.

	HASQI scores			
Type	-5 dB	0 dB	5 dB	Avg.
Mixture	0.026	0.064	0.136	0.075
Phase-insensitive	0.102	0.166	0.221	0.163
Phase-aware	**0.146**	**0.221**	0.270	**0.212**
Filtered-merged (4000 Hz)	0.144	0.219	0.270	0.211
Hybrid-net	0.130	0.219	**0.284**	0.211

and hybrid-net enhanced speech (average HASQI scores across SNRs) [$t(286) = 11.72$, $p < .001$]. There are no significant differences between phase-sensitive and hybrid-net enhanced speech [$t(286) = 0.25$, $p = .802$], filtered-merged speech at 4000 Hz and hybrid-net enhanced speech [$t(286) = 0.07$, $p = .943$].

On average, the phase-aware CRN achieves the best performance (i.e., 0.212), we also observe that the proposed hybrid-net achieves comparable performance with phase-aware CRN on average (i.e., 0.211). On the other hand, the proposed hybrid-net significantly outperforms the phase-insensitive CRN by relative 29%. The proposed hybrid-net achieves the best performance (i.e., 0.284) at 5 dB background SNR, where slightly degraded performance is observed at lower background SNRs (e.g., 0.130 at -5 dB).

5.2.5 Listening study and subjective results

Methods

We further conducted a listening study, in which participants compared the perceived quality of the stimuli generated by the proposed hybrid-net against those generated by three other speech enhancement systems, including (1) The original phase-aware CRN, (2) Phase-insensitive CRN and (3) The filtered-merged speech with a 4-kHz cutoff frequency. We follow the same experimental procedure as described in Section 5.1.3, unless otherwise stated. After the practice stage, there are 60 trials (i.e., 20 repetitions $\times$ 3 conditions) randomly ordered in the main experimental stage and the entire online experiment took

less than 30 minutes for each participant to complete.

There were, in total, 20 subjects (12 males and 8 females) recruited from Amazon Mechanical Turk who participated in this study. The average age is 38.5 years old (from 23 to 46 years). All participants were native speakers of American English and self-reported to sit in a quiet environment during the listening experiment.

Results

Figure 5.7 provides the d' values for each pair of comparisons (i.e., phase-insensitive model, denoted as 'Mag.'; filtered-merged speech at 4000 Hz, denoted as '4 kHz'; and phase-aware model). Note that all models are compared against the stimuli generated by the proposed hybrid-net and a higher d' value (above 0) indicates that stimuli generated by the proposed hybrid-net are more preferred. A d' value near 0 indicates that the listeners cannot reliably differentiate the difference between the two stimuli. The average d' value are over 0 for all comparison models, suggesting that the proposed hybrid-net can produce enhanced speech with better perceived quality. The results also suggest that human listeners are able to differentiate the quality between enhanced speech with noisy phase and (partial) estimated phase. A two-sided alternative hypothesis t-test showed significant differences between d' value and the hypothesis of 0-mean for all conditions, i.e., hybrid-net and phase-insensitive enhanced speech [$t(19) = 18.80$, $p < .001$], hybrid-net and filtered-merged speech at 4000 Hz [$t(19) = 3.04$, $p = .007$], hybrid-net and full-band phase-sensitive enhanced speech [$t(19) = 5.41$, $p < .001$]. This indicates that listeners prefer the hybrid-net enhanced speech significantly over all other three enhancement approaches, even compared to the full-band phase-aware CRN.

Similar to Section 5.1, a repeated measures ANOVA was conducted on d' values and results showed a significant main effect of different models [$F(2, 38) = 101.67$, $p < .001$, $\eta_p^2 = .843$]. Post hoc paired comparisons showed that there were significant differences between the phase-insensitive condition and filtered-merged speech [$t(19) = 12.33$, $p < .001$,

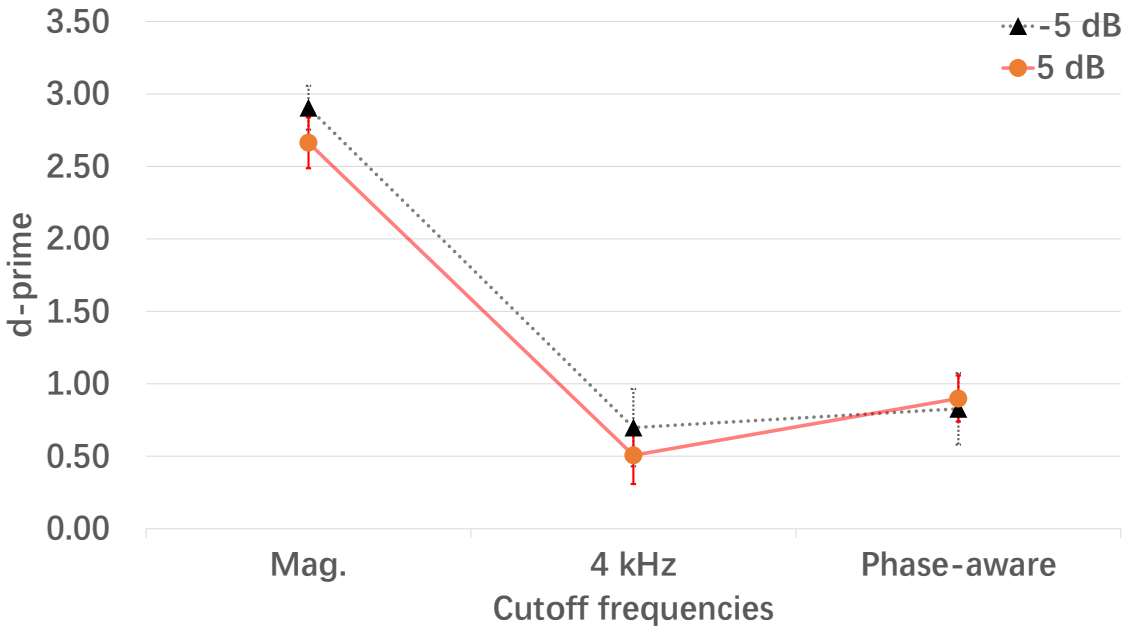

Figure 5.7: d' values across different models. Error bars indicate the $\pm$ standard errors. Two background SNRs (before enhancement) are included.

Bonferroni corrected], as well as between the phase-insensitive and phase-sensitive conditions [$t(19) = 11.64$, $p < .001$, Bonferroni corrected]. Moreover, there are no significant differences between the filtered-merged speech (at 4000 Hz) and full-band phase-aware enhanced speech [$t(19) = -1.63$, $p = .355$, Bonferroni corrected], which suggests listeners do not discriminate the hybrid enhanced speech and the filtered-merged speech better than discriminating the hybrid enhanced speech and the phase-aware enhanced speech. There was no significant effect of background SNR [$F(1, 19) = .691$, $p = .416$, $\eta_p^2 = .035$], neither was there a significant interaction between models and background SNR [$F(2, 38) = .61$, $p = .549$, $\eta_p^2 = .031$].

5.2.6 Summary

In this section, we proposed a novel hybrid framework for speech enhancement. The proposed system adopts two strategies handling phase at different frequencies (i.e., phase-aware below 4 KHz and phase-insensitive above). HASQI scores for simulated NH lis-

teners suggest that the proposed hybrid-net can achieve comparable performance to a full-band phase-aware speech enhancement system with much reduced network complexity. Subjective results also verify the advantages of the proposed hybrid-net over conventional phase-aware and phase-insensitive CRNs.

5.3 Discussions

5.3.1 Frequency importance for phase estimation

Both objective and subjective results demonstrate that estimating phase is mostly important at low frequency regions for human perception on speech quality in NH listeners. It is important to understand the underlying mechanisms why human listeners primarily use low-frequency information for quality judgment.

One possible explanation is that speech-quality judgment is, at least partially, related to pitch perception in human listeners where only frequencies between 30 to 4000 Hz (i.e., low-frequency region) are able to generate salient pitch sensation [198, 199]. This is also associated with the degradation of phase-locking ability (encoding the fine structure of the waveform) as the frequency of harmonics increases [200, 201]. It has been reported there are various cues that are associated with pitch coding, including spectral and temporal cues at low frequencies [202, 203], and only temporal envelope cues at high frequencies [204, 205, 206] as phase-locking ability degrades. Hence, high-frequency regions do not contribute to pitch perception as much as low frequencies. Meanwhile, studies have identified the existence of dominance region for pitch, where the spectral region of lower harmonics (500-1000 Hz as described in [207]) of a matched pitch is found to be the most important for perception [208]. Studies have reported that higher harmonics (about 5^{th} to 9^{th} in human listeners [209, 90]) of complex tones become unresolved. Future studies on the correlation between human pitch perception and quality judgment could provide a better understanding on frequency importance of phase estimation.

We also notice that the proposed hybrid-net achieves even better performance than the

full-band phase-aware model (i.e., as illustrated in Figure 5.7). We postulate that this could be caused by the more accurate estimation of phase components at low-frequency regions with hybrid-net compared to the full-band phase-aware model. As the proposed hybrid-net has a specific sub-network handling phase estimation at low-frequency regions, compared to the full-band phase-aware CRN that estimates the full-band phase components across entire spectral regions. It is possible that low-frequency components are weighted higher for speech quality judgment, therefore leading to better perceived quality for hybrid-net.

5.3.2 Potential benefits for HI listeners

Experimental results from Sections 5.1 and 5.2 demonstrate that estimating phase in low-frequency regions may be effective in improving speech quality in NH listeners. Moreover, as suggested in Chapter 4, HI listeners are negatively affected by phase distortions and may benefit from phase-aware enhancement approaches. Therefore, it is necessary to understand if the benefit from phase-aware enhancement algorithms, including our proposed hybrid-net, holds for the HI population.

In the following, we present the HASQI scores and analyses for simulated HI listeners on filtered-merged speech and hybrid-net enhanced speech.

Filtered-merged speech for HI listeners

We first provide HASQI scores for filtered-merged speech with different cutoff frequencies in Table 5.6, where we use the average auditory thresholds of HI listeners specified in Table 4.1. Note that a standard hearing-aid prescription formula (i.e., NAL-R) [163] is used to linearly amplify both the reference and degraded signals before HASQI evaluation (i.e., at the input end of the simulated hearing-aid).

Among the two speech enhancement algorithms (top of Table 5.6), the phase-aware speech enhancement system achieves slightly improved speech quality than the phase-

Table 5.6: HASQI scores for simulated HI listeners, performance of phase-insensitive and phase-aware systems is provided. The performance for the merged version of enhanced speech with different cutoff frequencies is also included. **Bold** font indicates the best performance.

	HASQI scores			
Type	-5 dB	0 dB	5 dB	Avg.
Mixture	0.052	0.128	0.261	0.147
Phase-insensitive	0.230	0.370	**0.472**	0.357
Phase-aware	**0.251**	0.373	0.457	0.360
250 Hz	0.234	0.372	**0.472**	0.359
500 Hz	0.237	0.373	**0.472**	**0.361**
1000 Hz	0.239	0.371	0.463	0.358
2000 Hz	0.249	**0.374**	0.459	**0.361**
4000 Hz	**0.251**	**0.374**	0.457	**0.361**

insensitive system on average (i.e., 0.360 vs. 0.357). Interestingly, although the phase-aware approach achieves a relative 9% improvement at low SNR levels (i.e., 0.251 vs. 0.230 at -5 dB), it underperforms the phase-insensitive system at higher SNR levels (i.e., 5 dB: 0.457 vs. 0.472). This could suggest that estimating phase is not necessary when the background noise is at a low level. Results here suggest that HI listeners are less sensitive to the enhanced phase components and therefore may benefit less from a phase-aware speech enhancement system compared to NH listeners.

Meanwhile, as the cutoff frequency increases for the filtered-merged speech, the average performance is getting closer to the full-band phase-aware enhanced speech. We also notice that for lower cutoff frequencies, performance at 5 dB is better than those with higher cutoff frequencies. This is likely caused by components from the phase-insensitive enhanced speech where it shows better performance at 5 dB. The gap between phase-aware and phase-insensitive speech enhancement algorithms is getting smeared especially at higher background SNRs for HI listeners. HI listeners show similar patterns with their NH peers (i.e., as shown in Table 5.2) except that they benefit less from a phase-aware speech enhancement system, possibly due to the degraded sensitivity to TFS cues.

Hybrid-net enhanced speech for HI listeners

We further provide the HASQI scores of simulated HI listeners (auditory profile specified in Table 4.1) for the proposed hybrid-net in Table 5.7.

Table 5.7: HASQI scores for simulated HI listeners. The best performance is marked with **bold** font.

	HASQI scores			
Type	-5 dB	0 dB	5 dB	Avg.
Mixture	0.052	0.128	0.261	0.147
Phase-insensitive	0.230	0.370	0.472	0.357
Phase-aware	**0.251**	0.373	0.457	0.360
Filtered-merged (4000 Hz)	**0.251**	0.374	0.457	0.361
Hybrid-net	0.249	**0.387**	**0.480**	**0.372**

Compared to results obtained for NH listeners, elevated speech quality is observed for HI listeners, which matches the previous findings in Chapter 4. Specifically, the proposed hybrid-net achieves the best performance on average in HI listeners (i.e., 0.372). We also notice that both hybrid-net and the phase-aware CRN yield comparable speech quality at low SNR condition (i.e., 0.249 and 0.251 at -5 dB), indicating that phase estimation at low-frequency regions is also important for HI listeners when background noise is strong.

5.3.3 Limitations and improvements of hybrid-net

Despite that the proposed hybrid-net achieves comparable (i.e., objective scores) and even better (i.e., subjective preference) performance in speech quality than the phase-aware CRN with significantly reduced network size and MACs, one can observe in Table 5.4 that it comes with increased inference time. This is likely caused by the two computation flows and their interactions (as illustrated in Figure 5.6) involved in the two sub-networks for low-frequency and high-frequency processing. One of the future directions could be optimizing the parallel processing among the two sub-networks to speed up the inference time, making it more suitable for efficient speech enhancement on low-resource platforms.

Another important future direction is refining the estimated components near the border cutoff frequency. The current hybrid-net estimates the low- and high-frequency components separately and this could introduce artifacts that may harm the signal integrity. Considering that the proposed hybrid-net adopts a cutoff frequency at 4 kHz, the artifacts could be hard for human listeners to detect but may still influence other speech communication systems. We could try using another neural network to 'smooth' the estimated components near the cutoff frequency for better performance.

Last but not least, it is important to evaluate the performance of hybrid-net on a more diverse speech dataset. As the current dataset only contains IEEE sentences with limited types of noise, which may not generalize well to real-world environments. It is necessary to investigate if the hybrid-net can still outperform or achieve comparable performance with conventional phase-aware speech enhancement systems in more realistic conditions (e.g., real-world recordings, reverberation effects, other languages besides American English, and speakers with diverse accents and dialects).

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Summary

Speech communications are often corrupted by unwanted noises, which poses a critical challenge for human listeners, especially for listeners with impaired hearing. Speech enhancement algorithms have been proposed to address this issue by removing the unwanted components from a mixture signal, however, it is not fully understood how these algorithms could be better implemented for human listeners, especially for hearing-impaired (HI) listeners. In this dissertation, we have addressed several aspects regarding how to implement speech enhancement algorithms for better performance on human listeners. Specifically, we have investigated (1) Performance of various types of speech enhancement algorithms, including conventional and deep learning based methods, (2) Influence of frequency scales on speech processing, (3) Human perception on phase distortion of speech signals, and (4) Importance of phase estimation across frequency bands. Furthermore, a novel speech enhancement framework is introduced in Chapter 5 that achieves comparable performance (i.e., speech quality) with a state-of-the-art phase-aware speech enhancement system, while with significantly reduced model complexity. We summarize the main findings and contributions of this dissertation below.

In Chapter 3, we first systematically examine and compare the performance of several speech enhancement algorithms for simulated normal-hearing (NH) and HI listeners, including an NMF-based method and other deep learning based methods (i.e., DNNs and RNNs). PESQ and HASQI are used to evaluate the objective speech quality, where results indicate that deep learning based speech enhancement methods significantly outperform a conventional speech enhancement algorithm that is based on NMF. Among the deep learn-

ing based methods, RNN-based systems are observed to achieve superior performance than DNN-based systems in terms of PESQ and HASQI scores. Similar trends are observed for both NH and HI listeners. Simulation results suggest a mismatch between HASQI and PESQ predictions on the speech quality following Mel-frequency domain processing. A human listening study is further conducted to investigate the influence of frequency scales on the speech quality. Subjective ratings suggest that human listeners prefer speech stimuli processed in the linear-frequency scale rather than in the Mel-frequency scale, which may be caused by the lower frequency resolution in our implementation of Mel-scale processing (i.e., 100 Mel-bins vs. 257 FFT-bins). We infer the difference between HASQI and PESQ predictions is a result of the deteriorated frequency resolution of the enhanced signals following the Mel-domain transformation, especially at high frequencies. PESQ is unaffected by this due to a narrower assessment bandwidth (3.1 kHz) than HASQI (12 kHz).

In Chapter 4, we investigate the influence of phase distortion on the perceived speech quality for both NH and HI listeners through several human listening studies. Experimental results indicate that HI listeners tend to provide higher ratings than NH peers for the same speech stimuli, corrupted by either background noises or reverberations. Similar patterns are found on the quality ratings for different degrees of phase distortion, which suggests that HI listeners have higher tolerance for phase distortion and noise/reverberation. After phase-insensitive speech enhancement (i.e., IRM), both groups of listeners can differentiate the degree of phase distortion that remained in the enhanced speech, indicating potential benefits from a phase-aware speech enhancement algorithm. We conjecture that these HI listeners are able to notice the phase distortions because (1) They have good temporal fine structure (TFS) sensitivity, or (2) TFS and phase cues are weighted higher for quality tasks than recognition tasks. We further obtain objective speech quality from two metrics (i.e., PESQ and HASQI) and investigate their correlations with human ratings. PESQ provides closer correlations to subjective ratings than HASQI, especially for enhanced speech. This is likely caused by the low-pass filtering introduced in the experiment design (i.e., low-

pass filtering is applied to all speech stimuli to ensure perceived quality is not dominated by high-frequency hearing loss), which makes PESQ more fitted in the narrow-band condition.

The influence of phase estimation across spectral bands is further investigated in Chapter 5. A phase-insensitive and another phase-aware speech enhancement systems are adopted to generate enhanced speech stimuli with different phase components, the enhanced speech stimuli are then filtered and merged together in a sense that low-frequency components in the filtered-merged stimuli are from phase-aware enhanced speech and high-frequency components are from phase-insensitive enhanced speech. Six octave cutoff frequencies are used to generate the filtered merged speech, including 0 Hz (i.e., full-band phase-insensitive enhanced speech), 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. These stimuli are paired with the full-band phase-aware enhanced speech to conduct a pairwise comparison study. Experimental results suggest that merged filtered speech stimuli with high cutoff frequencies (e.g., 2000 Hz and 4000 Hz) result in similar perceived quality. There is also a significant main effect of cutoff frequencies on the human preference, and that low-frequency phase estimation is most important to the perceived quality. HASQI is used to simulate the perceived speech quality for HI listeners and similar patterns are found, where most benefits are coming from phase estimation at lower-frequency regions. Based on these findings, we further propose a hybrid speech framework that adopts different phase estimation strategies across frequency bands. Phase-aware technique is applied for lower-frequency (i.e., 0 to 4000 Hz) processing whereas the high-frequency region (i.e., 4000 to 8000 Hz) adopts a phase-insensitive enhancement scheme. Experimental results suggest that the proposed hybrid-net can achieve comparable performance (i.e., speech quality) to a full-band phase-aware network but with much reduced model size and computations involved.

6.2 Future Work

In this dissertation, we have investigated several aspects of speech enhancement algorithms from a human perception point of view. Although significant progress has been made

on developing more powerful and efficient speech enhancement algorithms, there are still many promising future directions that worth exploring and we attempt to list a few of them below:

- *Evaluation on hearing-aids:* Although many speech enhancement systems with improved network complexity have been proposed, it is not clear how do these algorithms perform on actual hearing-aid devices. Some recent studies adopted mobile phone as an intermediate interface with digital hearing-aid devices [13, 210, 211]. In such way, more powerful speech enhancement algorithms could be implemented on low-resource platforms that enable better noise reduction performance than conventional methods. In Chapter 3, we investigated the objective performance of several speech enhancement systems for simulated HI listeners, however, it remains unclear on the actual performance of these algorithms for HI listeners when they are implemented in hearing-aids. Furthermore, can they meet runtime latency requirements for hearing-aids? These are the questions that need to be solved in the future study.

- *Training objective:* Existing speech enhancement algorithms are often optimized to fully recover the clean speech given the noisy input, where the objective function (e.g., MSE between estimated and oracle components) and performance metrics (e.g., PESQ, STOI, word error rate (WER)) are mismatched. In most cases, the speech enhancement algorithms can lead to improvement in speech quality and intelligibility. However, a trained system with the smallest loss value (e.g., MSE) may not necessarily yield the best performance in speech quality or ASR accuracy (i.e., WER). Prior studies have investigated using feedback from objective metrics as training criterion to update the speech enhancement system [212, 213, 214], while some other works [140, 138, 215] used perceptual loss terms and substantial improvements were observed. It is interesting to see more future works on this direction. Another idea is to directly incorporate responses from human listeners into training of speech enhancement systems, thus enabling the optimization of a speech enhancement system

towards a specific user (either NH or HI).

- *Model generalization:* Most of the current speech enhancement algorithms follow a supervised training scheme, where the systems are provided with a degraded and clean reference pair during training stage. However, it is often difficult or expensive to obtain clean reference speech in real-world environments. Therefore, most of the existing studies tend to generate a simulated dataset by mixing up clean speech with different noises at some specified SNR levels. However, the trained systems could potentially be deployed in an unseen environment that was not considered in the original design, which may degrade the systems' performance. Some self-supervised speech enhancement systems have been proposed recently [216, 217] that demonstrate comparable performance with supervised approaches while gaining superior generalization ability. It seems to be a promising future direction for improving the model generalization ability, especially when limited training materials are available.

- *Model compression:* A plethora of deep learning speech enhancement systems have been proposed over the past decade. Improved performance is often achieved with increased number of layers and more complicated network architectures. However, a light and efficient speech enhancement model is needed for applications on low-resource platforms, such as hearing-aid devices. Therefore, it is important to investigate model compression for speech enhancement. Recent studies have explored different ways to reduce the model complexity, including weight pruning, weight quantization [218, 219], and matrix product operators [220]. It is worth investigating these techniques for low-complexity speech enhancement. Another interesting direction could be injecting the human auditory design into speech enhancement systems for improved efficiency, where less computation power should be allocated to unimportant aspects of speech (e.g., phase component at high frequencies).

- *Incorporating visual cues:* Humans are able to exploit the audio-visual nature of the

speech to suppress noises and enhance the target speech. However, most of the existing speech enhancement algorithms only take in speech signals and do not consider the visual cues. Recent studies [221, 222, 54] suggested that incorporating additional visual cues could further enhance the performance of speech enhancement systems. It would be interesting to see if such audio-visual based speech enhancement algorithms can be successfully implemented for hearing-aids and other low-resource applications.

- *Preserving spatial cues:* Spatial cues (e.g., interaural time difference (ITD) and interaural level difference (ILD)) are important for human listeners to separate out different sound sources in a noisy environment [223, 224]. Any modifications to the signals presented to ears have the potential to distort such spatial information and it is crucial for speech enhancement algorithms to preserve this information to allow better human perception of the acoustic scene. Several studies [225, 226, 227] have demonstrated that better performance can be achieved by preserving the spatial cues on simulated binaural speech dataset, but it is unclear how do these approaches generalize to real-world conversations. Furthermore, it is important to properly define the evaluation criteria and training objectives for such binaural speech enhancement systems.

REFERENCES

[1] Joyce Vliegen and Andrew J Oxenham. “Sequential stream segregation in the absence of spectral cues”. In: *The Journal of the Acoustical Society of America* 105.1 (1999), pp. 339–346.

[2] Li Xu and Bryan E Pfingst. “Spectral and temporal cues for speech recognition: Implications for auditory prostheses”. In: *Hearing research* 242.1-2 (2008), pp. 132–140.

[3] Andreu Paredes-Gallardo et al. “The role of temporal cues in voluntary stream segregation for cochlear implant users”. In: *Trends in hearing* 22 (2018).

[4] Felix Weninger et al. “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR”. In: *International conference on latent variable analysis and signal separation*. Springer. 2015, pp. 91–99.

[5] Sunit Sivasankaran et al. “Robust ASR using neural network based speech enhancement and feature simulation”. In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE. 2015, pp. 482–489.

[6] Xueliang Zhang, Zhong-Qiu Wang, and DeLiang Wang. “A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 276–280.

[7] Zhuohuang Zhang et al. “ADL-MVDR: All deep learning MVDR beamformer for target speech separation”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 6089–6093.

[8] Zhuohuang Zhang et al. “Multi-channel multi-frame ADL-MVDR for target speech separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021).

[9] Yong Xu et al. “Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation”. In: *INTERSPEECH* (2021), pp. 3076–3080.

[10] Yixuan Zhang et al. “Continuous Speech Separation with Recurrent Selective Attention Network”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6017–6021.

[11] Hamid Sheikhzadeh, Robert L Brennan, and Hossein Sameti. “Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications”. In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 1995, pp. 808–811.

[12] Fatemeh Saki and Nasser Kehtarnavaz. “Automatic switching between noise classification and speech enhancement for hearing aid devices”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 736–739.

[13] Issa Panahi, Nasser Kehtarnavaz, and Linda Thibodeau. “Smartphone-based noise adaptive speech enhancement for hearing aid applications”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 85–88.

[14] Seon Man Kim. “Hearing aid speech enhancement using phase difference-controlled dual-microphone generalized sidelobe canceller”. In: *IEEE Access* 7 (2019), pp. 130 663–130671.

[15] Soha A Nossier et al. “Enhanced smart hearing aid using deep neural networks”. In: *Alexandria Engineering Journal* 58.2 (2019), pp. 539–550.

[16] Seung Ho Han et al. “Noise reduction for VoIP speech codecs using modified Wiener Filter”. In: *Advances and Innovations in Systems, Computing Sciences and Software Engineering*. Springer, 2007, pp. 393–397.

[17] Tokunbo Ogunfunmi, Roberto Togneri, and Madihally Narasimha. *Speech and audio processing for coding, enhancement and recognition*. Springer, 2015.

[18] Jiantao Liu et al. “Speech enhancement with stacked frames and deep neural network for VoIP applications”. In: *17th International Conference on Optical Communications and Networks (ICOCN2018)*. Vol. 11048. International Society for Optics and Photonics. 2019, p. 1104808.

[19] Yariv Ephraim. “Statistical-model-based speech enhancement systems”. In: *Proceedings of the IEEE* 80.10 (1992), pp. 1526–1555.

[20] Israel Cohen. “Relaxed statistical model for speech enhancement and a priori SNR estimation”. In: *IEEE Transactions on Speech and Audio Processing* 13.5 (2005), pp. 870–881.

[21] Jae-Hun Choi and Joon-Hyuk Chang. “On using acoustic environment classification for statistical model-based speech enhancement”. In: *Speech Communication* 54.3 (2012), pp. 477–490.

[22] Marwa A Abd El-Fattah et al. “Speech enhancement with an adaptive Wiener filter”. In: *International Journal of Speech Technology* 17.1 (2014), pp. 53–64.

[23] Tim Van den Bogaert et al. “Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids”. In: *The Journal of the Acoustical Society of America* 125.1 (2009), pp. 360–371.

[24] Navneet Upadhyay and Rahul Kumar Jaiswal. “Single channel speech enhancement: using Wiener filtering with recursive noise estimation”. In: *Procedia Computer Science* 84 (2016), pp. 22–30.

[25] Bhiksha Raj, Rita Singh, and Tuomas Virtanen. “Phoneme-dependent NMF for speech enhancement in monaural mixtures”. In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[26] Nasser Mohammadiha, Jalil Taghia, and Arne Leijon. “Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions”. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2012, pp. 4561–4564.

[27] Minje Kim and Paris Smaragdis. “Mixtures of local dictionaries for unsupervised speech enhancement”. In: *IEEE Signal processing letters* 22.3 (2014), pp. 293–297.

[28] Hao-Teng Fan et al. “Speech enhancement using segmental nonnegative matrix factorization”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 4483–4487.

[29] Kisoo Kwon, Jong Won Shin, and Nam Soo Kim. “NMF-based speech enhancement using bases update”. In: *IEEE Signal Processing Letters* 22.4 (2014), pp. 450–454.

[30] Tuomas Virtanen, Jort Florent Gemmeke, and Bhiksha Raj. “Active-set Newton algorithm for overcomplete non-negative representations of audio”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.11 (2013), pp. 2277–2289.

[31] Tuomas Virtanen, Bhiksha Raj, Jort F Gemmeke, et al. “Active-set newton algorithm for non-negative sparse coding of audio”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 3092–3096.

[32] Shinichi Tamura and Alex Waibel. “Noise reduction using connectionist models”. In: *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE. 1988, pp. 553–556.

[33] Shinichi Tamura. “An analysis of a noise reduction neural network”. In: *International Conference on Acoustics, Speech, and Signal Processing,* IEEE. 1989, pp. 2001–2004.

[34] Gibak Kim et al. “An algorithm that improves speech intelligibility in noise for normal-hearing listeners”. In: *The Journal of the Acoustical Society of America* 126.3 (2009), pp. 1486–1494.

[35] Kun Han and DeLiang Wang. “A classification based approach to speech segregation”. In: *The Journal of the Acoustical Society of America* 132.5 (2012), pp. 3475–3483.

[36] Yuxuan Wang and DeLiang Wang. “Towards scaling up classification-based speech separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.7 (2013), pp. 1381–1390.

[37] Johan AK Suykens and Joos Vandewalle. “Least squares support vector machine classifiers”. In: *Neural processing letters* 9.3 (1999), pp. 293–300.

[38] Xiaolin Huang, Lei Shi, and Johan AK Suykens. “Support vector machine classifier with pinball loss”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.5 (2013), pp. 984–997.

[39] Wenhua Shi et al. “Deep neural network and noise classification-based speech enhancement”. In: *Modern Physics Letters B* 31.19-21 (2017), p. 1740096.

[40] Yariv Ephraim and David Malah. “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator”. In: *IEEE transactions on acoustics, speech, and signal processing* 33.2 (1985), pp. 443–445.

[41] Yi Hu and Philipos C Loizou. “Subjective comparison and evaluation of speech enhancement algorithms”. In: *Speech communication* 49.7-8 (2007), pp. 588–601.

[42] DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[43] DeLiang Wang. “On ideal binary mask as the computational goal of auditory scene analysis”. In: *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[44] Eric W Healy et al. “An algorithm to improve speech recognition in noise for hearing-impaired listeners”. In: *The Journal of the Acoustical Society of America* 134.4 (2013), pp. 3029–3038.

[45] DeLiang Wang. “Time-frequency masking for speech separation and its potential for hearing aid design”. In: *Trends in amplification* 12.4 (2008), pp. 332–353.

[46] Valerie Hanson and Kofi Odame. “Real-time embedded implementation of the binary mask algorithm for hearing prosthetics”. In: *IEEE transactions on biomedical circuits and systems* 8.4 (2013), pp. 465–473.

[47] DeLiang Wang. “Deep learning reinvents the hearing aid”. In: *IEEE spectrum* 54.3 (2017), pp. 32–37.

[48] Xugang Lu et al. “Speech enhancement based on deep denoising autoencoder.” In: *Interspeech*. Vol. 2013. 2013, pp. 436–440.

[49] Yong Xu et al. “A regression approach to speech enhancement based on deep neural networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.1 (2014), pp. 7–19.

[50] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. “On training targets for supervised speech separation”. In: *IEEE/ACM transactions on audio, speech, and language processing* 22.12 (2014), pp. 1849–1858.

[51] Yi Luo and Nima Mesgarani. “Tasnet: time-domain audio separation network for real-time, single-channel speech separation”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 696–700.

[52] Yi Luo and Nima Mesgarani. “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation”. In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019), pp. 1256–1266.

[53] Ashutosh Pandey and DeLiang Wang. “On cross-corpus generalization of deep learning based speech enhancement”. In: *IEEE/ACM transactions on audio, speech, and language processing* 28 (2020), pp. 2489–2499.

[54] Daniel Michelsanti et al. “An overview of deep-learning-based audio-visual speech enhancement and separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021).

[55] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. “Complex ratio masking for monaural speech separation”. In: *IEEE/ACM transactions on audio, speech, and language processing* 24.3 (2016), pp. 483–492.

[56] Tian Gao et al. “SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement.” In: *INTERSPEECH*. 2016, pp. 3713–3717.

[57] Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. “Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (2016), pp. 153–167.

[58] Hakan Erdogan et al. “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 708–712.

[59] Zhuo Chen et al. “Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks”. In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[60] Lei Sun et al. “Multiple-target deep learning for LSTM-RNN based speech enhancement”. In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE. 2017, pp. 136–140.

[61] Shuai Nie et al. “Deep noise tracking network: a hybrid signal processing/deep learning approach to speech enhancement.” In: *Interspeech*. 2018, pp. 3219–3223.

[62] Yi Hu and Philipos C Loizou. “Techniques for estimating the ideal binary mask”. In: *Proc. 11th Int. Workshop Acoust. Echo Noise Control*. Citeseer. 2008, pp. 154–157.

[63] Yi Jiang, Hong Zhou, and Zhenming Feng. “Performance analysis of ideal binary masks in speech enhancement”. In: *2011 4th International Congress on Image and Signal Processing*. Vol. 5. IEEE. 2011, pp. 2422–2425.

[64] Arun Narayanan and DeLiang Wang. “Ideal ratio mask estimation using deep neural networks for robust speech recognition”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 7092–7096.

[65] Feng Bao and Waleed H Abdulla. “A new ratio mask representation for CASA-based speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2018), pp. 7–19.

[66] Christian Lorenzi and Brian CJ Moore. “Role of temporal envelope and fine structure cues in speech perception: A review”. In: *Proceedings of the International Symposium on Auditory and Audiological Research*. Vol. 1. 2007, pp. 263–272.

[67] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. “The importance of phase in speech enhancement”. In: *speech communication* 53.4 (2011), pp. 465–494.

[68] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. “Complex ratio masking for joint enhancement of magnitude and phase”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 5220–5224.

[69] Hyeong-Seok Choi et al. “Phase-aware speech enhancement with deep complex u-net”. In: *International Conference on Learning Representations*. 2018.

[70] Yanxin Hu et al. “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement”. In: *arXiv preprint arXiv:2008.00264* (2020).

[71] Ke Tan and DeLiang Wang. “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6865–6869.

[72] Ke Tan and DeLiang Wang. “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), pp. 380–390.

[73] Dacheng Yin et al. “PHASEN: A phase-and-harmonics-aware speech enhancement network”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 9458–9465.

[74] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. “Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR”. In: *IEEE/ACM transactions on audio, speech, and language processing* 28 (2020), pp. 1778–1787.

[75] Andong Li et al. “Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1829–1843.

[76] Guochen Yu et al. “Dual-branch Attention-In-Attention Transformer for single-channel speech enhancement”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7847–7851.

[77] Gyuseok Park et al. “Speech enhancement for hearing aids with deep learning on environmental noises”. In: *Applied Sciences* 10.17 (2020), p. 6077.

[78] Daniel Stoller, Sebastian Ewert, and Simon Dixon. “Wave-u-net: A multi-scale neural network for end-to-end audio source separation”. In: *arXiv preprint arXiv: 1806.03185* (2018).

[79] Chuanxin Tang et al. “Joint time-frequency and time domain learning for speech enhancement”. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 3816–3822.

[80] Soha A Nossier et al. “A comparative study of time and frequency domain approaches to deep learning based speech enhancement”. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.

[81] Jiaqi Su, Adam Finkelstein, and Zeyu Jin. “Perceptually-motivated environment-specific speech enhancement”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 7015–7019.

[82] Zhong-Qiu Wang and DeLiang Wang. “All-Neural Multi-Channel Speech Enhancement.” In: *Interspeech*. 2018, pp. 3234–3238.

[83] Rongzhi Gu et al. “Multi-modal multi-channel target speech separation”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.3 (2020), pp. 530–541.

[84] Hassan Taherian et al. “Robust speaker recognition based on single-channel and multi-channel speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1293–1302.

[85] Andong Li et al. “Embedding and Beamforming: All-neural Causal Beamformer for Multichannel Speech Enhancement”. In: *arXiv preprint arXiv:2109.00265* (2021).

[86] Zhuohuang Zhang et al. “All-neural beamformer for continuous speech separation”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6032–6036.

[87] Antony W Rix et al. “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”. In: *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE. 2001, pp. 749–752.

[88] Cees H Taal et al. “A short-time objective intelligibility measure for time-frequency weighted noisy speech”. In: *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2010, pp. 4214–4217.

[89] Emily Buss, Joseph W Hall III, and John H Grose. “Temporal fine-structure cues to speech and pure tone modulation in observers with sensorineural hearing loss”. In: *Ear and hearing* 25.3 (2004), pp. 242–250.

[90] Brian CJ Moore, Brian R Glasberg, and Kathryn Hopkins. “Frequency discrimination of complex tones by hearing-impaired subjects: Evidence for loss of ability to use temporal fine structure”. In: *Hearing research* 222.1-2 (2006), pp. 16–27.

[91] Christian Lorenzi et al. “Speech perception problems of the hearing impaired reflect inability to use temporal fine structure”. In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18866–18869.

[92] Kathryn Hopkins, Brian CJ Moore, and Michael A Stone. “Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech”. In: *The Journal of the Acoustical Society of America* 123.2 (2008), pp. 1140–1153.

[93] James M Kates and Kathryn H Arehart. “The hearing-aid speech quality index (HASQI) version 2”. In: *Journal of the Audio Engineering Society* 62.3 (2014), pp. 99–117.

[94] Felix Weninger et al. “Discriminatively trained recurrent neural networks for single-channel speech separation”. In: *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2014, pp. 577–581.

[95] Richard A Schmiedt. “The physiology of cochlear presbycusis”. In: *The aging auditory system*. Springer, 2010, pp. 9–38.

[96] Radiocommunication Sector ITU. *Recommendation bs. 1534-2: Method for the subjective assessment of intermediate quality level of audio systems*. 2014.

[97] Ke Tan and DeLiang Wang. “A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement.” In: *Interspeech*. 2018, pp. 3229–3233.

[98] Joerg Meyer and Klaus Uwe Simmer. “Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction”. In: *1997 IEEE international conference on acoustics, speech, and signal processing*. Vol. 2. IEEE. 1997, pp. 1167–1170.

[99] Md Kamrul Hasan, Sayeef Salahuddin, and M Rezwan Khan. “A modified a priori SNR for speech enhancement using spectral subtraction rules”. In: *IEEE signal processing letters* 11.4 (2004), pp. 450–453.

[100] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain”. In: *Speech communication* 52.5 (2010), pp. 450–475.

[101] Sunil Kamath, Philipos Loizou, et al. “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.” In: *ICASSP*. Vol. 4. Citeseer. 2002, pp. 44164–44164.

[102] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[103] Pascal Scalart et al. “Speech enhancement based on a priori signal to noise estimation”. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 2. IEEE. 1996, pp. 629–632.

[104] Israel Cohen. “Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models”. In: *Signal processing* 86.4 (2006), pp. 698–709.

[105] Richard C Hendriks, Richard Heusdens, and Jesper Jensen. “MMSE based noise PSD tracking with low complexity”. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, pp. 4266–4269.

[106] Timo Gerkmann and Richard C Hendriks. “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2011), pp. 1383–1393.

[107] Yariv Ephraim and David Malah. “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator”. In: *IEEE Transactions on acoustics, speech, and signal processing* 32.6 (1984), pp. 1109–1121.

[108] Jan S Erkelens et al. “Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.6 (2007), pp. 1741–1752.

[109] Jonathan Le Roux, John R Hershey, and Felix Weninger. “Deep NMF for speech separation”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, pp. 66–70.

[110] Shuai Nie et al. “Exploiting spectro-temporal structures using NMF for DNN-based supervised speech separation”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 469–473.

[111] Sean UN Wood et al. “Blind speech separation and enhancement with GCC-NMF”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4 (2017), pp. 745–755.

[112] Yuxuan Wang and DeLiang Wang. “A structure-preserving training target for supervised speech separation”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 6107–6111.

[113] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. “Reconstruction techniques for improving the perceptual quality of binary masked speech”. In: *The Journal of the Acoustical Society of America* 136.2 (2014), pp. 892–902.

[114] Yuxuan Wang and DeLiang Wang. “Cocktail party processing via structured prediction”. In: *Advances in Neural Information Processing Systems* 25 (2012), pp. 224–232.

[115] Yong Xu et al. “An experimental study on speech enhancement based on deep neural networks”. In: *IEEE Signal processing letters* 21.1 (2013), pp. 65–68.

[116] Martin Cooke. “A glimpsing model of speech perception in noise”. In: *The Journal of the Acoustical Society of America* 119.3 (2006), pp. 1562–1573.

[117] Ning Li and Philipos C Loizou. “Factors influencing glimpsing of speech in noise”. In: *The Journal of the Acoustical Society of America* 122.2 (2007), pp. 1165–1172.

[118] Ron J Weiss and Daniel PW Ellis. “Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking”. In: (2006).

[119] Gibak Kim and Philipos C Loizou. “Improving speech intelligibility in noise using environment-optimized algorithms”. In: *IEEE transactions on audio, speech, and language processing* 18.8 (2010), pp. 2080–2090.

[120] William Hartmann and Eric Fosler-Lussier. “Investigations into the incorporation of the ideal binary mask in ASR”. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2011, pp. 4804–4807.

[121] Douglas S Brungart et al. “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation”. In: *The Journal of the Acoustical Society of America* 120.6 (2006), pp. 4007–4018.

[122] Zhuohuang Zhang and Yi Shen. “Listener Preference on the Local Criterion for Ideal Binary-Masked Speech.” In: *INTERSPEECH*. 2019, pp. 1383–1387.

[123] Ulrik Kjems et al. “Role of mask pattern in intelligibility of ideal binary-masked noisy speech”. In: *The Journal of the Acoustical Society of America* 126.3 (2009), pp. 1415–1426.

[124] Jen-Cheng Hou et al. “Audio-visual speech enhancement using deep neural networks”. In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2016, pp. 1–6.

[125] Cassia Valentini Botinhao et al. “Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks”. In: *INTERSPEECH*. 2016, pp. 352–356.

[126] Morten Kolboek, Zheng-Hua Tan, and Jesper Jensen. “Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification”. In: *2016 IEEE spoken language technology workshop (SLT)*. IEEE. 2016, pp. 305–311.

[127] Khandokar Md Nayem and Donald S Williamson. “Incorporating intra-spectral dependencies with a recurrent output layer for improved speech enhancement”. In: *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2019, pp. 1–6.

[128] Szu-Wei Fu et al. “Raw waveform-based speech enhancement by fully convolutional networks”. In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2017, pp. 006–012.

[129] Se Rim Park and Jinwon Lee. “A Fully Convolutional Neural Network for Speech Enhancement”. In: *INTERSPEECH*. 2017.

[130] Soumitro Chakrabarty, DeLiang Wang, and Emanuël AP Habets. “Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks”. In: *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE. 2018, pp. 476–480.

[131] Ashutosh Pandey and DeLiang Wang. “A new framework for CNN-based speech enhancement in the time domain”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.7 (2019), pp. 1179–1188.

[132] Hynek Hermansky. “Perceptual linear predictive (PLP) analysis of speech”. In: *the Journal of the Acoustical Society of America* 87.4 (1990), pp. 1738–1752.

[133] Yang Shao and DeLiang Wang. “Robust speaker identification using auditory features and computational auditory scene analysis”. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2008, pp. 1589–1592.

[134] Hynek Hermansky and Nelson Morgan. “RASTA processing of speech”. In: *IEEE transactions on speech and audio processing* 2.4 (1994), pp. 578–589.

[135] Szu-Wei Fu et al. “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning”. In: *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*. IEEE. 2017, pp. 1–6.

[136] Yan Zhao et al. “Perceptually guided speech enhancement using deep neural networks”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5074–5078.

[137] Szu-Wei Fu, Chien-Feng Liao, and Yu Tsao. “Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality”. In: *IEEE Signal Processing Letters* 27 (2019), pp. 26–30.

[138] Ziyue Zhao, Samy Elshamy, and Tim Fingscheidt. “A perceptual weighting filter loss for DNN training in speech enhancement”. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2019, pp. 229–233.

[139] Masaki Kawanaka et al. “Stable training of DNN for speech enhancement based on perceptually-motivated black-box cost function”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7524–7528.

[140] Juan Manuel Martín-Doñas et al. “A deep learning loss function based on the perceptual evaluation of the speech quality”. In: *IEEE Signal processing letters* 25.11 (2018), pp. 1680–1684.

[141] Morten Kolbæk et al. “On loss functions for supervised monaural time-domain speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 825–838.

[142] Jean-Marc Valin et al. “A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech”. In: (2020), pp. 2482–2486.

[143] Zhuohuang Zhang, Yi Shen, and Donald S Williamson. “Objective comparison of speech enhancement algorithms with hearing loss simulation”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6845–6849.

[144] Zhuohuang Zhang, Donald S Williamson, and Yi Shen. “Impact of amplification on speech enhancement algorithms using an objective evaluation metric”. In: *International Congress on Acoustics (ICA)*. 2019.

[145] Helen Glyde et al. “Problems hearing in noise in older adults: a review of spatial processing disorder”. In: *Trends in amplification* 15.3 (2011), pp. 116–126.

[146] Cédric Févotte, Jonathan Le Roux, and John R Hershey. “Non-negative dynamical system with application to speech and audio”. In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2013, pp. 3158–3162.

[147] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. “Supervised and unsupervised speech enhancement using nonnegative matrix factorization”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (2013), pp. 2140–2151.

[148] Josef Shargorodsky et al. “Change in prevalence of hearing loss in US adolescents”. In: *Jama* 304.7 (2010), pp. 772–778.

[149] Ee-Munn Chia et al. “Hearing impairment and health-related quality of life: the Blue Mountains Hearing Study”. In: *Ear and hearing* 28.2 (2007), pp. 187–195.

[150] Abigail A Kressner, David V Anderson, and Christopher J Rozell. “Robustness of the hearing aid speech quality index (HASQI)”. In: *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2011, pp. 209–212.

[151] Abigail A Kressner, David V Anderson, and Christopher J Rozell. “Evaluating the generalization of the hearing aid speech quality index (HASQI)”. In: *IEEE transactions on audio, speech, and language processing* 21.2 (2012), pp. 407–415.

[152] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323.

[153] Chia-Ping Chen and Jeff A Bilmes. “MVA processing of speech features”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1 (2006), pp. 257–270.

[154] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[155] Tijmen Tieleman, Geoffrey Hinton, et al. “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude”. In: *COURSERA: Neural networks for machine learning* 4.2 (2012), pp. 26–31.

[156] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. “Hybrid speech recognition with deep bidirectional LSTM”. In: *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE. 2013, pp. 273–278.

[157] EH Rothauser. “IEEE recommended practice for speech quality measurements”. In: *IEEE Trans. on Audio and Electroacoustics* 17 (1969), pp. 225–246.

[158] Michael Nilsson, Sigfrid D Soli, and Jean A Sullivan. “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise”. In: *The Journal of the Acoustical Society of America* 95.2 (1994), pp. 1085–1099.

[159] John S Garofolo et al. “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1”. In: *NASA STI/Recon technical report n* 93 (1993), p. 27403.

[160] Anthony J Spahr et al. “Development and validation of the AzBio sentence lists”. In: *Ear and hearing* 33.1 (2012), p. 112.

[161] Karol J Piczak. “ESC: Dataset for environmental sound classification”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 1015–1018.

[162] Yi Hu and Philipos C Loizou. “Evaluation of objective quality measures for speech enhancement”. In: *IEEE Transactions on audio, speech, and language processing* 16.1 (2007), pp. 229–238.

[163] Denis Byrne and Harvey Dillon. “The National Acoustic Laboratories’(NAL) new procedure for selecting the gain and frequency response of a hearing aid”. In: *Ear and hearing* 7.4 (1986), pp. 257–265.

[164] Zhuohuang Zhang, Donald S Williamson, and Yi Shen. “Investigation of phase distortion on perceived speech quality for hearing-impaired listeners”. In: *INTERSPEECH*. 2020, pp. 2512–2516.

[165] Yingyue Xu et al. “Distorting temporal fine structure by phase shifting and its effects on speech intelligibility and neural phase locking”. In: *Scientific reports* 7.1 (2017), pp. 1–9.

[166] Meet H Soni, Neil Shah, and Hemant A Patil. “Time-frequency masking-based speech enhancement using generative adversarial network”. In: *ICASSP*. IEEE. 2018, pp. 5039–5043.

[167] L. P. Yang and Q. J. Fu. “Spectral subtraction-based speech enhancement for cochlear implant patients in background noise”. In: *JASA* 117.3 (2005), pp. 1001–1004.

[168] R. C. Mathes and R. L. Miller. “Phase Effects in Monaural Perception”. In: *The Journal of the Acoustical Society of America* 19.5 (1947), pp. 780–797. eprint: https://doi.org/10.1121/1.1916623.

[169] James H. Craig and Lloyd A. Jeffress. “Effect of Phase on the Quality of a Two-Component Tone”. In: *The Journal of the Acoustical Society of America* 34.11 (1962), pp. 1752–1760. eprint: https://doi.org/10.1121/1.1909118.

[170] R. Plomp and H. J. M. Steeneken. “Effect of Phase on the Timbre of Complex Tones”. In: *The Journal of the Acoustical Society of America* 46.2B (1969), pp. 409–421. eprint: https://doi.org/10.1121/1.1911705.

[171] Jont B. Allen and David A. Berkley. “Image method for efficiently simulating small-room acoustics”. In: *The Journal of the Acoustical Society of America* 65.4 (1979), pp. 943–950. eprint: https://doi.org/10.1121/1.382599.

[172] Michael R Wirtzfeld et al. “Predicting the quality of enhanced wideband speech with a cochlear model”. In: *The Journal of the Acoustical Society of America* 142.3 (2017), EL319–EL325.

[173] Pejman Mowlaee, R Saiedi, and Rainer Martin. “Phase estimation for signal reconstruction in single-channel speech separation”. In: *Proceedings of the International Conference on Spoken Language Processing*. 2012, pp. 1–4.

[174] Martin Krawczyk and Timo Gerkmann. “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12 (2014), pp. 1931–1940.

[175] Chaslav V Pavlovic. “Derivation of primary parameters and procedures for use in speech intelligibility predictions”. In: *The Journal of the Acoustical Society of America* 82.2 (1987), pp. 413–422.

[176] S3 22-1997 ANSI. “Methods for calculation of the speech intelligibility index”. In: *American National Standard Institute* (1997).

[177] Jianfen Ma, Yi Hu, and Philipos C Loizou. “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions”. In: *The Journal of the Acoustical Society of America* 125.5 (2009), pp. 3387–3405.

[178] Eric W Healy, Sarah E Yoho, and Frédéric Apoux. “Band importance for sentences and words reexamined”. In: *The Journal of the Acoustical Society of America* 133.1 (2013), pp. 463–473.

[179] Rory A DePaolis, Claus P Janota, and Tom Frank. “Frequency importance functions for words, sentences, and continuous discourse”. In: *Journal of Speech, Language, and Hearing Research* 39.4 (1996), pp. 714–723.

[180] Lena LN Wong et al. “Development of the Cantonese speech intelligibility index”. In: *The Journal of the acoustical society of America* 121.4 (2007), pp. 2350–2361.

[181] In-Ki Jin et al. “The band-importance function for the Korean standard sentence lists for adults”. In: *Journal of audiology & otology* 20.2 (2016), p. 80.

[182] Dequan Wang and Jae Lim. “The unimportance of phase in speech enhancement”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30.4 (1982), pp. 679–681.

[183] Pejman Mowlaee and Rahim Saeidi. “Iterative closed-loop phase-aware single-channel speech enhancement”. In: *IEEE Signal Processing Letters* 20.12 (2013), pp. 1235–1239.

[184] Chaslav V Pavlovic. “Band importance functions for audiological applications”. In: *Ear and Hearing* 15.1 (1994), pp. 100–104.

[185] Herman JM Steeneken and Tammo Houtgast. “Mutual dependence of the octave-band weights in predicting speech intelligibility”. In: *Speech communication* 28.2 (1999), pp. 109–123.

[186] Andrew Brughera, Larisa Dunai, and William M Hartmann. “Human interaural time difference thresholds for sine tones: The high-frequency limit”. In: *The Journal of the Acoustical Society of America* 133.5 (2013), pp. 2839–2855.

[187] Eric Verschooten, Christian Desloovere, and Philip X Joris. “High-resolution frequency tuning but not temporal coding in the human cochlea”. In: *PLoS biology* 16.10 (2018), e2005164.

[188] AR Palmer and IJ Russell. “Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells”. In: *Hearing research* 24.1 (1986), pp. 1–15.

[189] PHILIP X Joris et al. “Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency”. In: *Journal of neurophysiology* 71.3 (1994), pp. 1022–1036.

[190] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. “Fast and accurate deep network learning by exponential linear units (elus)”. In: *arXiv preprint arXiv:1511.07289* (2015).

[191] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[192] Andrew Varga and Herman JM 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”. In: *Speech communication* 12.3 (1993), pp. 247–251.

[193] Alice E Milne et al. “An online headphone screening test based on dichotic pitch”. In: *Behavior Research Methods* 53.4 (2021), pp. 1551–1562.

[194] Elliot M Cramer and WH Huggins. “Creation of pitch through binaural interaction”. In: *The Journal of the Acoustical Society of America* 30.5 (1958), pp. 413–417.

[195] Maria Chait, David Poeppel, and Jonathan Z Simon. “Neural response correlates of detection of monaurally and binaurally created pitches in humans”. In: *Cerebral cortex* 16.6 (2006), pp. 835–848.

[196] PB Elliot and JA Swets. *Signal detection and recognition by human observers*. 1964.

[197] Neil A Macmillan and C Douglas Creelman. *Detection theory: A user’s guide*. Psychology press, 2004.

[198] Fred Attneave and Richard K Olson. “Pitch as a medium: A new approach to psychophysical scaling”. In: *The American journal of psychology* (1971), pp. 147–166.

[199] Michael S Osmanski, Xindong Song, and Xiaoqin Wang. “The role of harmonic resolvability in pitch perception in a vocal nonhuman primate, the common marmoset (Callithrix jacchus)”. In: *Journal of Neuroscience* 33.21 (2013), pp. 9161–9168.

[200] Eric D Young and Murray B Sachs. “Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers”. In: *The Journal of the Acoustical Society of America* 66.5 (1979), pp. 1381–1403.

[201] Andrew J Oxenham. “Pitch perception”. In: *Journal of Neuroscience* 32.39 (2012), pp. 13335–13338.

[202] Colette M McKay, Hugh J McDermott, and Robert P Carlyon. “Place and temporal cues in pitch perception: Are they truly independent?” In: *Acoustics Research Letters Online* 1.1 (2000), pp. 25–30.

[203] Tim Green, Andrew Faulkner, and Stuart Rosen. “Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants”. In: *The Journal of the Acoustical Society of America* 112.5 (2002), pp. 2155–2164.

[204] Trevor M Shackleton and Robert P Carlyon. “The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination”. In: *The Journal of the Acoustical Society of America* 95.6 (1994), pp. 3529–3540.

[205] Ray Meddis and Michael J Hewitt. “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification”. In: *The Journal of the Acoustical Society of America* 89.6 (1991), pp. 2866–2882.

[206] Xindong Song et al. “Complex pitch perception mechanisms are shared by humans and a New World monkey”. In: *Proceedings of the National Academy of Sciences* 113.3 (2016), pp. 781–786.

[207] Peter A Cariani and Bertrand Delgutte. “Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch”. In: *Journal of neurophysiology* 76.3 (1996), pp. 1717–1734.

[208] William A Yost. “The dominance region and ripple noise pitch: a test of the peripheral weighting model”. In: *The Journal of the Acoustical Society of America* 72.2 (1982), pp. 416–425.

[209] Geoffrey A Moore and Brian CJ Moore. “Perception of the low pitch of frequency-shifted complexes”. In: *The Journal of the Acoustical Society of America* 113.2 (2003), pp. 977–985.

[210] Yu Rao et al. “Smartphone-based real-time speech enhancement for improving hearing aids speech perception”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 5885–5888.

[211] Chandan Karadagur Ananda Reddy et al. “An individualized super-Gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device”. In: *IEEE signal processing letters* 24.11 (2017), pp. 1601–1605.

[212] Szu-Wei Fu et al. “Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2031–2041.

[213] Haoyu Li et al. “iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning”. In: *arXiv preprint arXiv: 2004.00932* (2020).

[214] Szu-Wei Fu et al. “MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement”. In: *arXiv preprint arXiv:2104.03538* (2021).

[215] Saurabh Kataria, Jesús Villalba, and Najim Dehak. “Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 7118–7122.

[216] Ryandhimas E Zezario et al. “Self-supervised denoising autoencoder with linear regression decoder for speech enhancement”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6669–6673.

[217] Ying Cheng et al. “Improving Multimodal Speech Enhancement by Incorporating Self-Supervised and Curriculum Learning”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 4285–4289.

[218] Jyun-Yi Wu et al. “Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques”. In: *IEEE Signal Processing Letters* 26.12 (2019), pp. 1887–1891.

[219] Ke Tan and DeLiang Wang. “Towards Model Compression for Deep Learning Based Speech Enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1785–1794.

[220] Xingwei Sun et al. “A model compression method with matrix product operators for speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2837–2847.

[221] Mandar Gogate et al. “CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement”. In: *Information Fusion* 63 (2020), pp. 273–285.

[222] Mostafa Sadeghi et al. “Audio-visual speech enhancement using conditional variational auto-encoders”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1788–1800.

[223] Jürgen Peissig and Birger Kollmeier. “Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners”. In: *The Journal of the Acoustical Society of America* 101.3 (1997), pp. 1660–1670.

[224] Monica L Hawley, Ruth Y Litovsky, and John F Culling. “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer”. In: *The Journal of the Acoustical Society of America* 115.2 (2004), pp. 833–843.

[225] Junfeng Li et al. “Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication”. In: *Speech Communication* 53.5 (2011), pp. 677–689.

[226] Joachim Thiemann et al. “Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene”. In: *EURASIP Journal on Advances in Signal Processing* 2016.1 (2016), pp. 1–11.

[227] Andreas I Koutrouvelis et al. “Binaural speech enhancement with spatial cue preservation utilising simultaneous masking”. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 598–602.

CURRICULUM VITAE

Education

Indiana University, Bloomington, IN, USA

- Ph.D., Double Major in Speech & Hearing Sciences and Computer Science 2022
- M.S., Computer Science 2021

University of Rochester, Rochester, NY, USA

- M.S., Electrical and Computer Engineering 2017

Beijing Institute of Technology, Beijing, China

- B.Eng., Opto-electrical Information Engineering 2015

Publications

- G. Yi, W. Xiao, Y. Xiao, B. Naderi, S. Möller, W. Wardah, G. Mittag, R. Cutler, **Z. Zhang**, D. S. Williamson, F. Chen, F. Yang, and S. Shang, "ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications." In arXiv preprint arXiv: 2203.16032 2022.
- **Z. Zhang**, T. Yoshioka, N. Kanda, Z. Chen, X. Wang, D. Wang, and S. E. Eskimez. "All-neural Beamformer for Continuous Speech Separation." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- D. Yun, Y. Shen, and **Z. Zhang**. "Feasibility of hearing aid gain self-adjustment using speech recognition." In *Journal of the Acoustical Society of Korea (JASK)*, 2022.

- **Z. Zhang**, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. S. Williamson, and D. Yu. "Multi-channel Multi-frame ADL-MVDR for Target Speech Separation." In *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2021.

- Y. Xu, **Z. Zhang**, M. Yu, S.-X. Zhang, and D. Yu. "Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation." In *INTERSPEECH*, 2021.

- **Z. Zhang**, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu. "ADL-MVDR: All deep learning MVDR beamformer for target speech separation." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

- **Z. Zhang**, P. Vyas, X. Dong, and D. S. Williamson. "An End-To-End Non-Intrusive Model for Subjective and Objective Real-World Speech Assessment Using a Multi-Task Framework." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

- **Z. Zhang**, D. S. Williamson, and Y. Shen. "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners." In *INTERSPEECH*, 2020.

- **Z. Zhang**, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, H. Song, and X. Li. "On loss functions and recurrency training for GAN-based speech enhancement systems." In *INTERSPEECH*, 2020.

- **Z. Zhang**, and Y. Shen. "Listener Preference on the Local Criterion for Ideal Binary-Masked Speech." In *INTERSPEECH*, 2019.

- **Z. Zhang**, D. S. Williamson, and Y. Shen. "Impact of amplification on speech enhancement algorithms using an objective evaluation metric." In *International Congress on Acoustics (ICA)*, 2019.

- **Z. Zhang**, Y. Shen, and D. S. Williamson. "Objective comparison of speech enhancement algorithms with hearing loss simulation." In *International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 2019.

- Y. Shen, C. Zhang, and **Z. Zhang**. "Feasibility of interleaved Bayesian adaptive procedures in estimating the equal-loudness contour." In *Journal of the Acoustical Society of America (JASA)*, 2018.

- Z. Tan, L. Zhao, **Z. Zhang**, B. Shan, J. Wang, and J. Xu. "Sulfur Passivation Enhancement for GaSb MOS Devices by Adding H_2O_2 to $(NH_4)_2S$ Solution." In *IEEE Semiconductor Interface Specialists Conference*, 2014.

- Q. Luo, Z. Xiao, P. Lu, **Z. Zhang**, and L. Zhao. "Mechanical design and kinematic analysis of wearable lumbodorsal therapeutic instrument." In *Journal of Mechanical & Electrical Engineering*, 2014.

- Y. Liu, Y. Song, Q. Hao, T. Tan, C. Liu, and **Z. Zhang**. "Design and implementation of a retina-like imaging system based on non-uniform lens array." In *International Symposium on Optoelectronic Technology and Application: Infrared Technology and Applications*, 2014.