Title	Study on monaural speech enhancement by restoring instantaneous amplitude and instantaneous phase
Author(s)	Vo, Duc Duy
Citation	
Issue Date	2022-12
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/18166
Rights	
Description	Supervisor:鵜木 祐史, 先端科学技術研究科, 修士(情報科学)

Master's Thesis

Study on monaural speech enhancement by restoring instantaneous amplitude and instantaneous phase

VO Duc Duy

Supervisor UNOKI Masashi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

December, 2022

Abstract

Speech is one of the most essential and important means for humans to communicate with each other. With the advancement of science and technology, speech now can be transmitted from a far distance to help connect people around the world through phones or web meetings,... Furthermore, it can also be used as an interface for humans to communicate with machines via automatic speech recognition (ASR) systems. However, in real environments, speech is contaminated by noise, reducing its quality and intelligibility such that it heavily affects the performance of these systems. In order to address this issue, efficient speech enhancement algorithms are needed for both human hearing and ASR systems.

Many techniques have been proposed to separate clean speech from the noisy mixture. Current state-of-the-art methods usually use short-time Fourier transform (STFT) as the means for feature extraction. The word recognition rate of ASR systems utilizing these techniques as the front-end still falls short of expectations, despite the fact that these methods can enhance the quality and intelligibility of noisy speech. Recent studies have showed that temporal envelope and temporal fine structure are crucial cues for speech perception and they also play a significant role in improving the speech intelligibility in noisy conditions. Therefore, speech enhancement by modifying instantaneous amplitude (IA) and instantaneous phase (IPh) extracted from an auditory filterbank is expected to have better improvement in quality, intelligibility as well as word recognition rate of ASR systems than STFT. On this basis, a speech enhancement method based on IA and IPh from the auditory filterbank was proposed. However, this method processed each channel independently, which could neglect important cross-channel information for ASR systems.

The purpose of this research is to investigate a model that can utilize cross-channel information of IA and IPh to explore the ability of this information in elevating the word recognition rate of ASR systems. This model revolves around vector-quantized variational autoencoder to estimate IA, and a complex convolution network to estimate IPh.

The efficacy of the proposed model will be estimated using three metrics: perceptual evaluation of speech quality, short-time objective intelligibility, and word error rate. The outcomes demonstrate that the proposed method can enhance quality, intelligibility of noisy speech and is competitive with some state-of-the-art methods. However, this method still cannot resolve the issue of high word error rate in ASR systems.

Contents

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and VNU - HCMC University of Sciences.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research background

One of the most fundamental and important ways that people interact with one another is through speech. In these modern days, instead of having to be physically close to each other to have a conversation, we can utilize the convenience of electronic devices such as phones, computers,... to be in touch with other people around the world.
With the breakthrough of sciences and technologies, speech is no longer a communication tool exclusively between humans anymore. The research and development activities have brought up new innovations that allow humans to communicate with machines to provide better life experiences. An important medium for human-machine interfaces is automatic speech recognition (ASR) systems.

However, there exists one problem that affects speech for human hearing and its related systems and still remains unsolved. In noisy conditions, speech signals are severely degraded in quality and intelligibility. The performance of ASR systems is also heavily reduced due to the damaged integrity of speech. To overcome this problem, researchers have been developing speech enhancement (SE) algorithms as the pre-processing system for noisy speech throughout the past decades. By processing noisy speech using these SE models, the quality and intelligibility of speech can be both improved. However, it is still not enough to elevate the word recognition rate for ASR systems since enhanced speech is usually over-suppressed or distorted [1]. Therefore, effective speech enhancement for both human hearing and ASR systems under noisy conditions is a challenging problem and worth paying attention to.

1.2 Problem statement

Numerous speech enhancement methods have been proposed for human hearing and robust ASR systems as a front-end. From classic unsupervised techniques such as spectral subtraction, minimum mean-squared error of short-time spectral amplitude, and statistical methods [2] to powerful supervised models such as deep learning networks [3]. Modern state-of-the-art methods, notably [4], usually process on short-time Fourier transform (STFT) domain. Yet the enhanced speech from these models still does not have a good performance on ASR systems since these models tend to overly suppress the noise and cause distortions in the output speech [1].

The temporal envelope and temporal fine structure are crucial cues for speech perception, according to several psycho-acoustical research [5, 6], and they also significantly contribute to speech intelligibility in noisy conditions [7]. Moreover, it has been found that the low-frequency range that is less than 2 kHz has more deciding factors to speech intelligibility [8]. These findings match with human auditory system since we process speech in the time domain and we are good at telling the differences between low-frequency sounds but barely able to distinguish high-frequency sounds. Therefore, speech enhancement by modifying instantaneous amplitude (IA) and instantaneous phase (IPh) extracted from an auditory filterbank is expected to have more improvement in quality as well as intelligibility than STFT [9]. On this basis, a speech enhancement method based on IA and IPh from the auditory filterbank was proposed in [9], which can improve sound quality as well as the intelligibility of speech under noisy conditions. However, this method processed each channel independently, which could neglect cross-channel information that contains spectral features. Since most ASR systems use spectral features extracted from the frequency domain [10], neglecting cross-channel information may reduce the word recognition rate of ASR systems.

1.3 Research purpose

This research aims to investigate a deep-learning-based speech enhancement framework that can utilize the cross-channel information of IA and IPh from the auditory filterbank to explore the ability of this information in elevating both the quality, intelligibility, and word recognition rate of ASR systems for the enhanced speech.

To achieve this research purpose, there are three sub-tasks in this study. The first sub-task is to investigate a pair of analysis-synthesis auditory filterbanks that can be integrated into a larger trainable network. Since the

analysis filterbank can be re-implemented from [9], the synthesis filterbank is expected to behave like an inverse of this analysis block. The second sub-task is to propose a deep learning model to estimate clean IA given noisy IA. This model needs to have abilities to utilize cross-channel information from the input. The final sub-task is to propose a module for phase correction with the same ability to utilize cross-channel information.

The novelty of this study is to utilize the spectral information across AF sub-bands of IA and IPh for speech enhancement tasks so that it could help improve the recognition rate of enhanced speech from ASR systems. In detail, this study considers the restoration of IA and IPh at a channel using the information of IA and IPh from all other channels. If the proposed method can improve both sound quality, speech intelligibility as well as word recognition rate of ASR systems, it can help elevate the quality of various applications such as hearing aids, ASR, mobile phones, and online web-meetings.

1.4 Structure of the thesis

The structure of this thesis is as follows:

- Chapter 1 presents the importance of speech enhancement in Section 1.1, explains the problems in Section 1.2, and describes the objectives, originality, and importance in Section 1.3.

- In chapter 2, a survey about typical methods for speech enhancement will be introduced in Section 2.1. An explanation for the target research is presented in Section 2.2, which includes a review of the analysis auditory filterbank in Section 2.2.1 and overview of this target method in Section 2.2.2.

- Chapter 3 will introduce the proposed method. First, the procedure to obtain the synthesis auditory model is described in Section 3.1. Then, Sections 3.2 and 3.3 will explain how to restore the instantaneous amplitude and instantaneous phase, respectively.

- Chapter 4 describes in detail how to implement the proposed method. Specifically, Sections 4.1 and 4.2 will explain two deep learning models used for instantaneous amplitude estimation and Section 4.3 will explain the model for instantaneous phase estimation.

- Chapter 5 explains the evaluations of the proposed method. Dataset information, objective metrics, and model configurations are described here. And the results and discussion are also presented.

- Chapter 6 summarizes this thesis and discusses the future works that still need to be done to make the proposed method better.

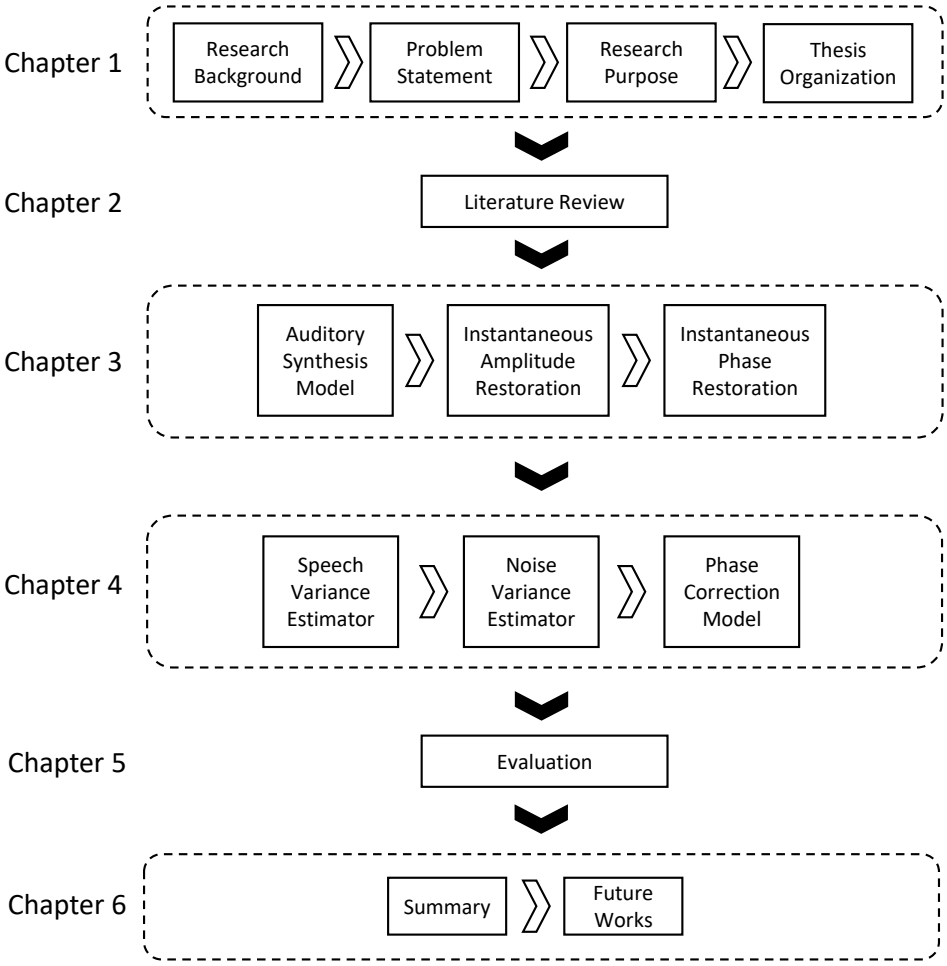


Figure 1.1: Structure of the thesis

Chapter 2

Literature review

2.1 Typical methods for speech enhancement

For single-channel speech enhancement tasks, several techniques have been presented. These techniques can be divided into two groups: unsupervised and supervised. For the unsupervised category, there is a wide range of notable algorithms. Although most of them are statistical models, there are still additional factors that need to be reviewed. The supervised algorithms, on the other hand, employ trainable models to suppress noise given a dataset of both clean and noisy speech in advance. The model will be given pairings of clean and noisy speech during the training phase, and an optimization algorithm (mainly gradient descent) is performed to adjust the model parameters in order to produce speech that is as similar to the clean speech as possible. In this section, both categories and their respective algorithms shall be reviewed.

For the unsupervised category, a well-known method is spectral subtraction initially proposed by Boll [11] for stationary noise suppression. Short-time Fourier transform (STFT) is used to convert the noisy speech to the frequency domain. The non-speech sections are used to estimate the noise magnitude spectrum, which is then subtracted from the noisy magnitude spectrum to get the estimated clean speech spectrum. However, this method produces musical noise artifacts in enhanced speech. Paliwal *et al.* [12], employed the modulation domain in an analysis-modification-synthesis (AMS) architecture to decrease the musical noise for enhanced speech in order to solve this issue.

For statistical-based speech enhancement algorithms, the Wiener filter works on an AMS scheme to minimize the mean squared error (MSE) between the predicted magnitude spectrum and clean speech spectrum [13]. Sev-

eral variations have been proposed over the years such as [14], which utilized relationships between time frames to reduce residual noise. And [15], which estimated a priori signal-to-noise ratio. These methods employed the STFT-based AMS models to obtain the estimated power spectrum from the noisy power spectrum. Other minimizing MSE-related methods focused on short-time spectral amplitude (MMSE-STSA). Some notable schemes for this method are [16], which can reduce musical noise in enhanced speech, and [17] - which estimated the underlying clean speech magnitude spectrum given the phase information. Additionally, techniques based on non-negative matrix factorization (NMF) have been introduced to separate components in the noisy spectrum. Li *et al.* [18] used the NMF to decompose the noisy spectrum into three parts: structured noise, raw speech estimate, and noise residual. The first two parts are then used in a non-negative sparse coding framework to predict a binary mask used to predict clean spectrum. Mohammed *et al.* [19] pre-trained a partial set of basis spectrum on clean speech in advance and adapt it to noisy spectrum in runtime.
Aside from the STFT domain as a means for feature extraction, many researchers have started incorporating human auditory system properties into their models. Such properties are captured and modeled as gammatone filterbank [20] and gammachirp filterbank [21]. Lin *et al.* [22] used a gammatone filterbank (GTFB) to extract features from the noisy speech and estimated the clean speech using the Wiener filter idea. On this basis, Kianyfar *et al.* [23] extended the idea and improved the performance of this model to adapt to non-stationary noise. Kortlang *et al.* [24] used a statistical model to predict the noise spectral density for the Wiener filter-based model and yielded comparable results in strong babble noise environment.

Over the past decade, deep learning-based models have been studied for speech enhancement and helped elevate the performance of this field of research since then. By utilizing a large scale of training data, these supervised algorithms can provide robust solutions for speech enhancement. Several methods processed the raw speech directly. Pandey *et al.* [25] used a convolutional neural network comprised of causal and dilated convolutional layers to map noisy waveform to enhanced waveform. A speech enhancement generative adversarial network (SEGAN) was introduced by Pascual *et al.* [26] to train two models: a generator G and a discriminator D. During the training phase, G will map noisy speech x to enhanced speech $\hat{x}$, while D will try to tell the differences between $\hat{x}$ and true clean speech y. Both G and D are simultaneously trained, G will try to perform better to fool D and D will try to perform better at distinguishing $\hat{x}$ and y until $\hat{x}$ is not distinguishable from true clean speech y. Both methods achieved high results in improving the quality and intelligibility of noisy speech.

Another approach is using STFT as a means for feature extraction and processing on the magnitude spectrum to learn the mask targets. These targets are ideal binary mask (IBM) [27,28], ideal ratio mask (IRM) [29,30] or spectral magnitude mask (SMM) [31]. By multiplying those masks with the noisy magnitude spectrum on an element-wise basis, the estimated clean magnitude spectrum can be obtained. This estimated magnitude spectrum is then incorporated with the noisy phase to form the enhanced speech using inverse short-time Fourier transform. In other words, these methods just leave the phase unprocessed. However, by experimenting on various cases of different combinations of noisy/clean amplitude/phase in the STFT domain, [32] and [33] have shown that the clean phase spectrum still yields significant information for speech enhancement. Motivated by this finding, researchers have started paying attention to the phase spectrum. Erdogan *et al.* incorporated phase information into the SMM and called it phase-sensitive mask [34]. The results of this method yielded a better estimation for clean speech than the original SMM.
Still, direct clean phase estimation remains a challenging task. In recently proposed deep-learning-based methods, researchers have started to focus on processing the complex STFT coefficients, which can deal with the phase in an indirect way. Lee *et al.* [35] proposed a model to predict a parametric complex-valued time-frequency mask (PCM), this mask is used to jointly estimate both STFT magnitude spectrum and phase spectrum. Li *et al.* [36] utilized long-short term memory (LSTM) layers with delayed output to estimate a complex ideal ratio mask, which is then applied to the complex STFT coefficient of noisy speech to obtain the clean one. Sun *et al.* [37] used a U-net model with the complex spectrum as input to predict both STFT magnitude spectrum and phase spectrum. These techniques have opened a new approach to deal with the phase information, which has been considered difficult since the wrapped phase does not have any clear patterns to be distinguished from the noisy one, and the unwrapped phase is unbounded.

2.2 Speech enhancement on instantaneous amplitude and instantaneous phase

2.2.1 Auditory filterbank

The gammatone filterbank is designed after the cochlear of humans auditory system [38]. Each sub-band of this filterbank corresponds to a position on

the basilar membrane. The impulse response of this filter is expressed as

$$g(t) = at^{n-1}e^{-2\pi b_f ERB(f_c)t}\cos 2\pi f_c t,\ t \geq 0 \tag{2.1}$$

where a, n and b_f are filter amplitude, filter order, and bandwidth coefficient, respectively. Center frequency f_c and equivalent rectangular bandwidth $ERB(f_c)$ for this frequency are defined as follows

$$ERB(f_c) = 24.7 + 0.108 f_c \tag{2.2}$$

In case $n = 4$, b_f=1.019, Eq. (2.1) represents human's auditory filter [39]. This filterbank characteristics are illustrated in Fig. 2.1.

To obtain the instantaneous amplitude (IA) and instantaneous phase (IPh) of a random signal *x(t)*, the Hilbert transform is applied on Eq. (2.1) to derive the analytic representation of the filter

$$\zeta(t) = at^{n-1}e^{j2\pi f_c t - 2\pi b_f ERB(f_c)t}, \tag{2.3}$$

By using $\zeta(t)$ to filter *x(t)*, we can obtain

$$X(k,t) = |X(k,t)|e^{(j\arg(X(k,t)))}, \tag{2.4}$$

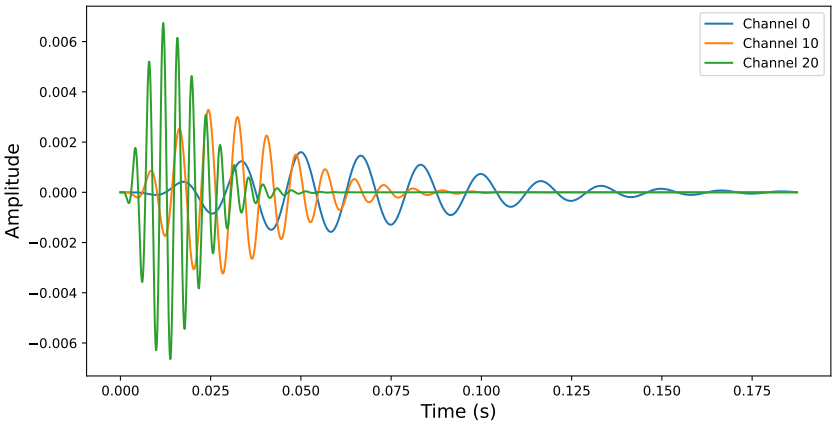
with $|X(k,t)|$ being IA and $\Phi(k,t) = \arg(X(k,t))$ being the phase spectrum. k, t is the channel and sample index, respectively. Then we have IPh defined as

$$\phi(k,t) = \Phi(k,t) - 2\pi f_k t, \tag{2.5}$$

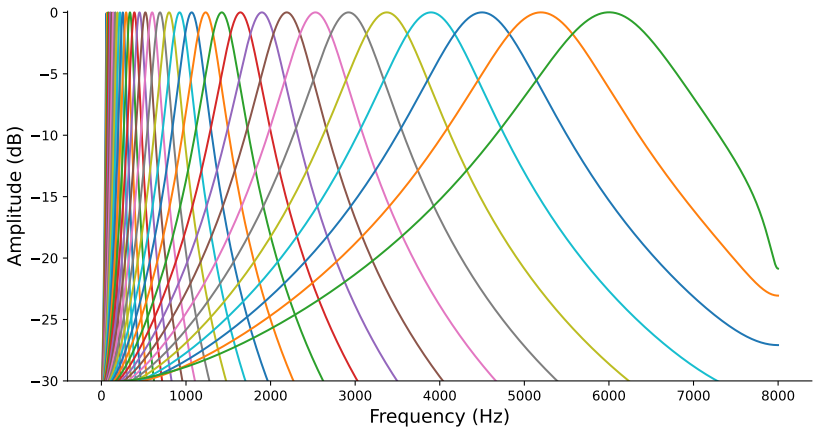
where f_k is the center frequency at channel k.
IA $|X(k,t)|$ and IPh $\phi(k,t)$ will be used as feature targets for our speech enhancement task. Also, from this point onward, if nothing else is mentioned, gammatone filterbank will be regarded as analysis model and inverse gammatone filterbank will be regarded as the synthesis model.

2.2.2 Kalman filter-based method for instantaneous amplitude and instantaneous phase estimation

Nower *et al.* [9] have proposed a model for speech enhancement focusing on instantaneous amplitude (IA) and instantaneous phase (IPh) restoration. In this analysis-modification-synthesis (AMS) scheme, IA and IPh are extracted from the noisy speech by an analysis gammatone filterbank into K channels. Then, on one channel at a time, the linear prediction (LP) coefficients of clean speech are calculated by two methods: non-blind and blind.



(a) Gammatone impulse responses



(b) Gammatone frequency responses

Figure 2.1: Gammatone filterbank characteristics (33 channels in total). Top: impulse responses in the 0th, 10th, and 20th channel (center frequencies are 60 Hz, 123 Hz, and 253 Hz, respectively); bottom: magnitude frequency responses in all 33 channels

In the non-blind method, clean speech is available for LP coefficients calculation. The purpose of this non-blind scenario is to identify the upper improvement limit for this scheme. For the blind method, clean speech is not available. An investigation that LP coefficients of IA and IPh in each channel have similarities in values and are not dependent on speaker identity or the content of the speech was conducted and verified. The LP coefficients are pre-trained on a closed dataset and then inferred during the enhancement

process.

These LP coefficients will be passed to a Kalman filter to estimate clean IA and IPh on a recursive basis. Based on the state of earlier time steps, the state of the present time step can be approximated. Finally, the estimated IA and IPh are combined together to form the real signal in the gammatone domain and this real signal is synthesized back to a waveform by using the inverse gammatone filterbank. This methods overall block diagram is shown in Fig. 2.2.

However, one thing to be noted is this method processed each channel inde-

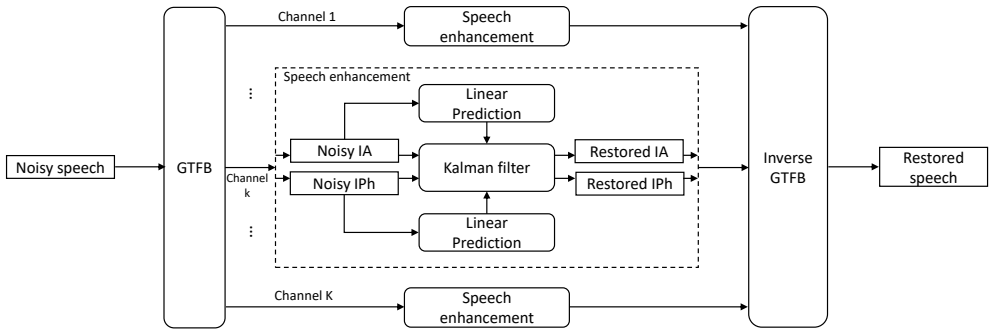


Figure 2.2: Block diagram of Kalman filter-based method for instantaneous amplitude and instantaneous phase estimation

pendently, which neglected cross-channel information that contains spectral features. Most ASR systems use spectral features extracted from the frequency domain [10], so neglecting cross-channel information may reduce the word recognition rate of ASR systems.

Chapter 3

Proposed Model

3.1 Auditory filterbank synthesis model

In order to convert the target features in this work (IA and IPh) back to the time domain, a proper synthesis model must be investigated. In the target method [9], time-reversal impulse responses of the analysis model was used as the synthesis model. However, since the method in this work is focusing on deep learning, both the analysis and synthesis models are required to be integrated as part of a larger trainable network for backpropagation algorithm [40].

Our analysis and synthesis models utilize the structure of convolutional neural network (CNN) [41]. The analysis model is implemented using the same idea as [38] while the synthesis model is initialized to be zeros in all channel. To train the synthesis network, a signal *x(t)* is passed through the analysis network, followed by Hilbert transform to obtain complex signal $X_k(t)$. This complex signal will be used to obtain IA and IPh for the speech enhancement task later. But for the purpose of obtaining the synthesis model in this section, $X_k(t)$ is transformed back to real values $\hat{X}_k(t)$ using

$$\hat{X}_k(t) = |X_k(t)| \cos\left(\arg\left(X_k(t)\right)\right) \tag{3.1}$$

The synthesis model is trained to map $\hat{X}_k(t)$ to the output $\hat{x}(t)$ so that $\hat{x}(t)$ is the same as input $x(t)$. Figure 3.1 illustrates this process.

The impulse responses of this synthesis model after training is showed in Fig. 3.2. This trained synthesis models will be used for speech enhancement task in the following sections.

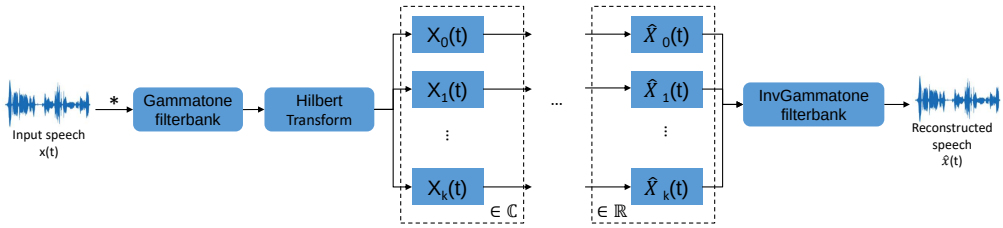


Figure 3.1: Analysis and synthesis models

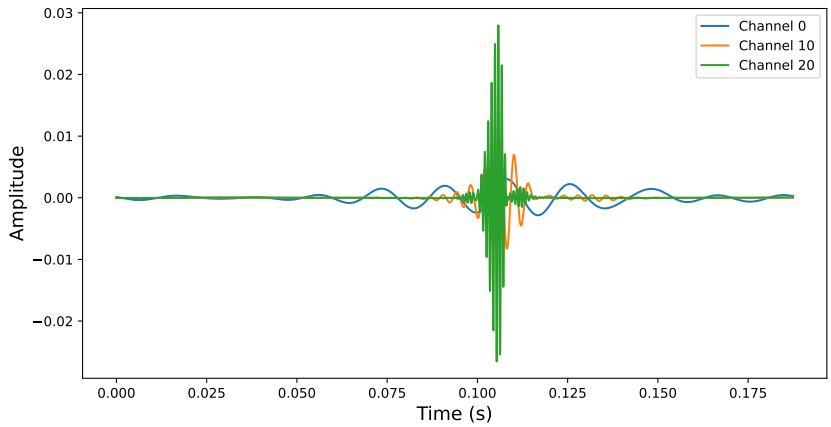


Figure 3.2: Impulse responses of the trained inverse gammatone filterbank at channel 0, 10 and 20 (center frequencies are 60Hz, 123Hz and 253Hz, respectively)

3.2 Instantaneous amplitude restoration

In the time domain, the noisy speech $x(n)$ can be represented as the summation of speech $s(n)$ and noise $e(n)$

$$x(n) = s(n) + e(n), \tag{3.2}$$

It is assumed that the signals $s(n)$ and $e(n)$ have a mean of zero, and this assumption can be recognized by subtracting each of them with their corresponding means at the beginning of the procedure. Therefore, we also have $x(n)$ to be zero mean.
The noisy speech $x(n)$ is filtered by a K-channel gammatone filterbank $g(\tau)$ described in section 2.2.1. The result of this process is given by

$$X(k,n) = x(n) * g(k,\tau) = S(k,n) + E(k,n), \tag{3.3}$$

where $*$ denotes the convolution operation. When clean speech is the input, the analysis filter produces $S(k,n)$; when pure noise is the input, it produces $E(k,n)$. Now if we assume that the signal at channel k - $X(k,n)$, can be denoised by a mask h

$$\hat{X}(k,n) = hX(k,n), \tag{3.4}$$

To calculate this mask h, we consider the mean squared error (MSE) between $\hat{X}(k,n)$ and $S(k,n)$

$$Err = E[(\hat{X}(k,n) - S(k,n))^2] = E[(hX(k,n) - S(k,n))^2], \tag{3.5}$$

In order to reduce this error Err, we set the derivative of 3.5 to zero with regard to h,

$$\begin{aligned} \frac{\partial Err}{\partial h} = E\left[\frac{\partial (hX(k,n) - S(k,n))^2}{\partial h}\right] = 0 \\ E\left[2h(S(k,n) + E(k,n))^2 - 2(S^2(k,n) + E(k,n)S(k,n))\right] = 0 \end{aligned} \tag{3.6}$$

If we assume that the speech component $S(k,n)$ and the noise component $E(k,n)$ are uncorrelated to each other, and they also have zero mean, then finally

$$h = E\left[\frac{S^2(k,n)}{S^2(k,n) + E^2(k,n)}\right] = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{3.7}$$

where σ_s^2 is the speech variance and σ_e^2 is the noise variance.

The estimated instantaneous amplitude of clean speech $\hat{A}_S(k,n)$ can be obtained using

$$\hat{A}_S(k,n) = A_X(k,n) \odot h^\gamma \tag{3.8}$$

with $\odot$ being the element-wise multiplication operator, $A_X(k,n)$ is the IA of noisy speech, and γ is an adjustable parameter used to scale the mask. During experimentation, $\gamma = 0.5$ is chosen since it gives the best results for speech perception. Then Eq. 3.8 is similar to an ideal ratio mask in [3].

Motivated by [42], a method that estimated a Wiener filter that has similar form to the mask h. In this study, we use vector-quantized variational autoencoder (VQ-VAE) model to estimate the speech variance σ_s^2. And a convolution neural network (CNN) to estimate the noise variance σ_e^2. Detailed for each model will be described in chapter 4.

3.3 Instantaneous phase restoration

Due to the warping property, direct mapping from noisy phase to estimated clean phase is difficult. To overcome this problem, several studies have proposed using a complex convolution network (CCN) to predict an additive

mask Φ_m for phase correction [4,42]. Using this mask, estimated clean phase $\hat{\Phi}_S$ can be realized by

$$\hat{\Phi}_S = \Phi_X + \Phi_m, \tag{3.9}$$

where $\hat{\Phi}_S, \Phi_X$ are estimated clean phase and noisy phase, respectively. This CCN is a sub model of the modification block in an analysis-modification-synthesis (AMS) scheme. To train this model, the scale-invariant signal-to-noise ratio (SI-SNR) is used as the optimization function for the AMS process. This function minimizes the angle between the clean speech and the enhanced speech, hence it can help estimate the amount of correction for noisy phase.

Motivated by this idea, we investigate two setups in this study to estimate the phase additive mask Φ_m:

- Noisy complex gammatone coefficients as input for the network to learn the additive mask, this setup has the same idea for phase correction as [4, 42].
- Instantaneous phase as input for the network model to learn the additive mask.

With these setups, stretched SI-SNR (SSISNR) is used as the optimization function since this function has fewer local maxima than SI-SNR [37].

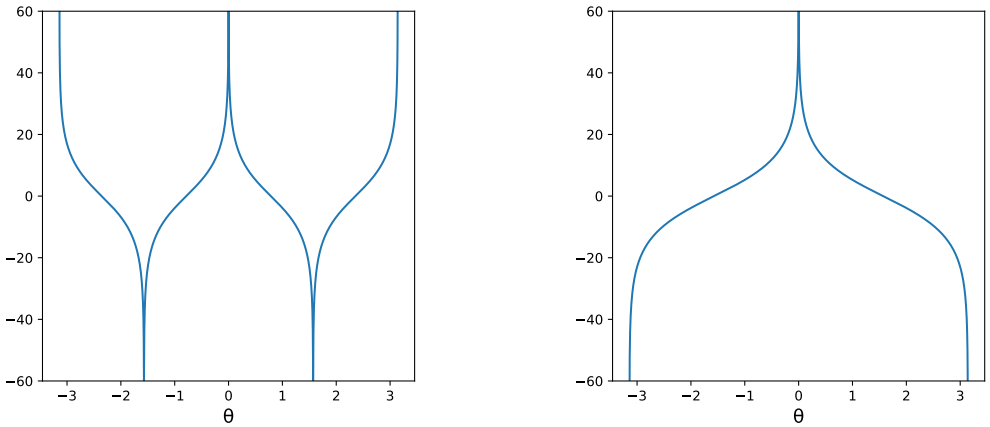
$$\text{SSISNR} = 10 \log_{10} \frac{1 + \cos\theta}{1 - \cos\theta} \tag{3.10}$$

$$\cos\theta = \frac{x \cdot \hat{x}}{|x||\hat{x}|} \tag{3.11}$$

where θ is the angle between the enhanced speech $\hat{x}$ and the clean speech x. As illustrated in Fig. 3.3, using SISNR could make the algorithm think $-\pi$ and π are the optimal points but what we actually want is 0. Using SSISNR can avoid this issue.

Finally, the estimated clean instantaneous phase $\hat{\phi_S}$ is obtained by adding the noisy instantaneous phase $\hat{\phi_S}$ with this mask

$$\hat{\phi}_S = \phi_X + \Phi_m, \tag{3.12}$$



(a) SISNR

(b) SSISNR

Figure 3.3: SISNR and SSISNR functions

Chapter 4

Implementation

4.1 Speech variance estimator model

4.1.1 Vector-quantized Variational Autoencoder

First introduced in 2017, vector-quantized variational autoencoder (VQ-VAE) [43] is a generative model that works on discrete latent distribution of data. Basically, it works based on the idea of learning from experience. This model's structure consists of:

- An encoder network used to extract useful information from input x to discrete latent space h_x.
- A vector-quantization (VQ) block that takes in the output of the encoder h_x. This block consists of a codebook and a quantizer. The codebook defines the latent embedding space $e \in \mathbb{R}^{K \times D}$ while the quantizer finds an embedding set v_x from the codebook such that the distance from set h_x to $\{e_k \in \mathbb{R}^D, k \in 1, 2, ..., K\}$ is smallest. Then this block will output this set v_x to the decoder.

$$v_x = \{e_k, \text{ where k=argmin}_j \|h_x - e_j\|_2^2\} \tag{4.1}$$

- A decoder that reconstructs $\hat{x}$ from the embedding set v_x.

Fig. 4.1 illustrates this process. During the training phase, the latent codebook will approach the true discrete latent distribution of the original dataset x and the encoder network will also try to make its output approach this distribution. After the training phase, we can consider the codebook in VQ block to be the knowledge learned from the dataset x and this knowledge will be used to generate future data.

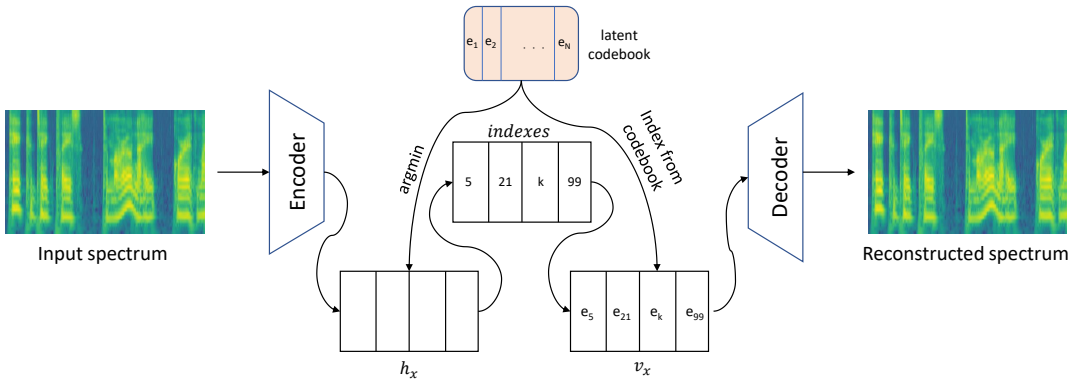


Figure 4.1: Overview of the VQ-VAE model

The optimization function for this model is described as

$$\mathcal{L} = d(x, \hat{x}) + \|\text{sg}(h_x) - v_x\|^2 + \delta\|h_x - \text{sg}(v_x)\|^2, \tag{4.2}$$

where $d(x, \hat{x})$ is the distance between x and $\hat{x}$, sg$(\cdot)$ is the stop-gradient operator, and δ is set to 0.25 to be the same as the original paper [43].
In the quantization process, backpropagation algorithm cannot run in Eq. 4.1 because there is no gradient here. Therefore, the straight-through reparameterization trick is used to directly copy the gradient from the decoder input v_x back to the encoder output h_x to solve this problem.

$$\begin{aligned} h_x &= \text{Enc}(x), \\ v_x &= \text{Quantizer}(h_x), \\ \hat{v}_x &= h_x + \text{sg}(v_x - h_x), \\ \hat{x} &= \text{Dec}(\hat{v}_x), \end{aligned}$$

where Enc$(\cdot)$, Quantizer$(\cdot)$, and Dec$(\cdot)$ are the encoder function, quantizer function, and decoder function, respectively.

4.1.2 Speech variance estimation using VQ-VAE-based model

As explained in Section 3.2, to restore IA, we need to estimate speech variance σ_s^2 and noise variance σ_e^2 for the mask h. To use the VQ-VAE for speech variance estimation, the latent codebook should be robust to noise. However, when we used direct noisy input to train VQ-VAE, we observed that the latent codebook has very low perplexity, which means that the patterns

of encoded clean speech features could not be learned in the codebook. Hence the decoder does not have enough information to reconstruct the speech spectrogram.
Since the main functionality of the codebook is to provide sufficient knowledge for the decoder to reconstruct the clean spectrogram. Direct training on noisy speech could make the codebook learn false information. Therefore, a pre-training process on clean speech was conducted. In other words, in the pre-training step, input to the VQ-VAE will be clean speech features S to train the codebook. Then, in the main training step, the codebook with pre-trained knowledge will be frozen, and only the encoder and decoder will be further trained on noisy features X.

The model used in this study is illustrated in Fig. 4.2. It has two VQ-VAE levels. Only level 1 takes in input X and output $\hat{X}$. Encoders and decoders are based on Wavenet model [44]. A block of Wavenet modules is put between two CNN layers to construct encoder and decoder with the activation functions being the differences between them. The detailed structures for encoder-decoder are illustrated in Fig. 4.3.

One thing to be noted that is the input to decoder 1 is concatenated from

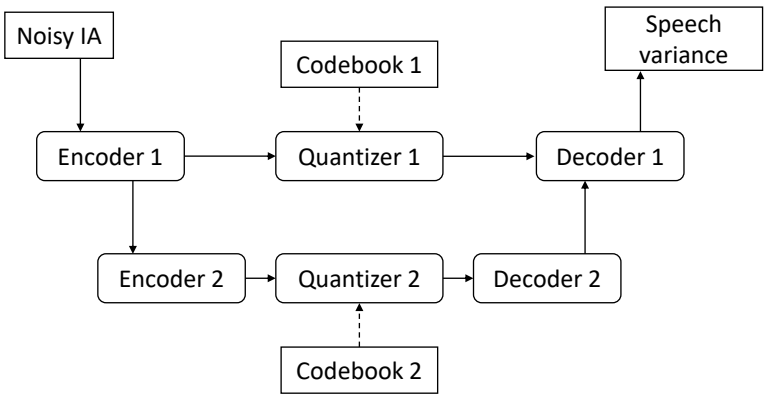


Figure 4.2: Structure of VQ-VAE model used to estimate speech variance in main training phase

the output of quantizer 1 and decoder 2 to reconstruct the speech variance. Since there are two steps in training this model: pre-training step and main-training step, we also have two optimization functions for each stage

$$\mathcal{L}_{pre}(\Theta) = \mathrm{LSD}(A_S^2, \hat{\sigma}_s^2) + \sum_{i \in \mathcal{U}} \|v_S^i - \mathrm{sg}(h_S^i)\|^2 + \delta \sum_{i \in \mathcal{U}} \|\mathrm{sg}(v_S^i) - h_S^i\|^2, \quad (4.3)$$

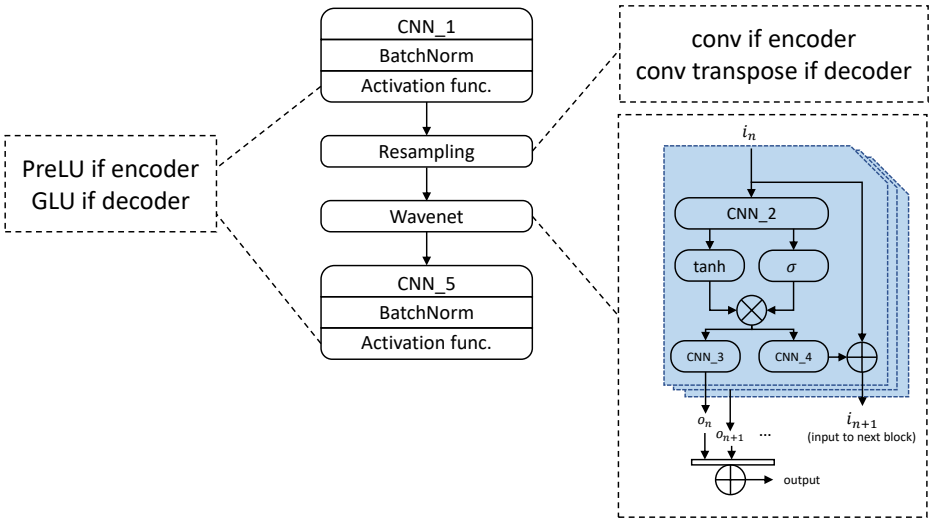


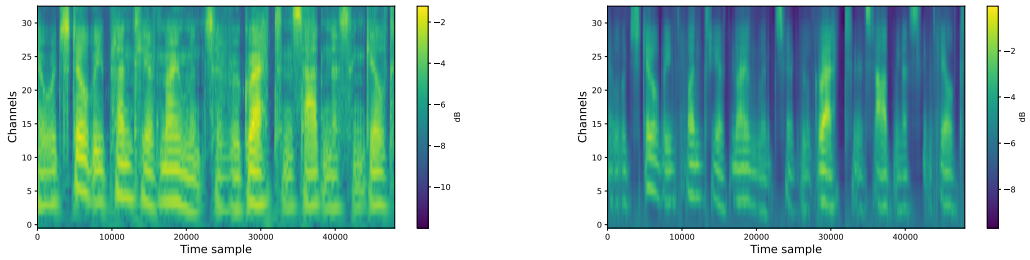
Figure 4.3: Wavenet-based structure encoder-decoder.

$$\mathcal{L}_{main}(\Theta) = \mathrm{LSD}(A_S^2, \hat{\sigma}_s^2) + \delta \sum_{i \in \mathcal{U}} \|\mathrm{sg}(v_S^i) - h_X^i\|^2, \tag{4.4}$$

where $\mathrm{LSD}(A_S^2, \hat{\sigma}_s^2)$ is log spectral distance between squared clean IA A_S^2 and estimated speech variance $\hat{\sigma}_s^2$, $\mathcal{U}$={level 1, level 2}, and Θ indicates all parameters of all sub-models. h^i is the output of the encoders, v^i is the quantized value as in Eq. 4.1 and $\mathrm{sg}(\cdot)$ is the stop gradient operator. δ is 0.25, the same value as in the original paper [43]. The difference between Eq. 4.3 and Eq. 4.4 because in main-training stage, only the encoders and decoders are fine-tuned while the codebooks are frozen. Therefore $\sum_{i \in \mathcal{U}} \|v^i - \mathrm{sg}(h^i)\|^2$ part, which is used to train the codebook in main-training phase is not needed.
Fig. 4.4 illustrates the results of this process. The speech variance $\hat{\sigma}_s^2$ is now can be obtained.

4.2 Noise variance estimator model

To estimate the noise variance σ_e^2, this study utilizes the noise-variance estimation model of [42]. This model includes one encoder and one decoder stacked on top of each other. Their structures are already described in Fig.



(a) Squared speech IA (b) Predicted speech variance

Figure 4.4: Squared-value of speech IA and predicted speech variance

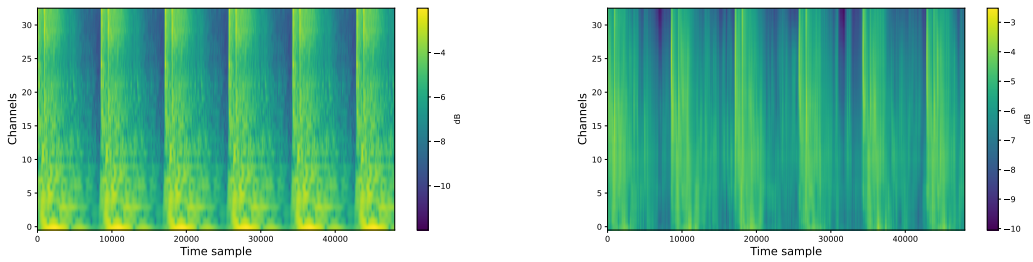
4.3. The input to this model is

$$I = \max(A_X^2(k,n) - \hat{\sigma}_s^2, 0) \tag{4.5}$$

Log spectral distance will also be chosen as the optimization function for this model. The predicted noise variance is illustrated in Fig. 4.5.

$$\mathcal{L}_{noise}(\Theta) = \mathrm{LSD}(A_E^2, \hat{\sigma}_e^2) \tag{4.6}$$

The noise variance $\hat{\sigma}_e^2$ is now also can be obtained. Using Eq. 3.7 and 3.8, we can obtain estimated clean instantaneous amplitude $\hat{A}_S(k,n)$. The overall process to restore IA can be summarized in Fig. 4.6



(a) Squared noise IA (b) Predicted noise variance

Figure 4.5: Squared-value of noise IA and predicted noise variance

4.3 Phase correction model

A deep complex network is used to estimate the additive phase mask Φ_m for phase correction. This network structure consists of a complex convolution

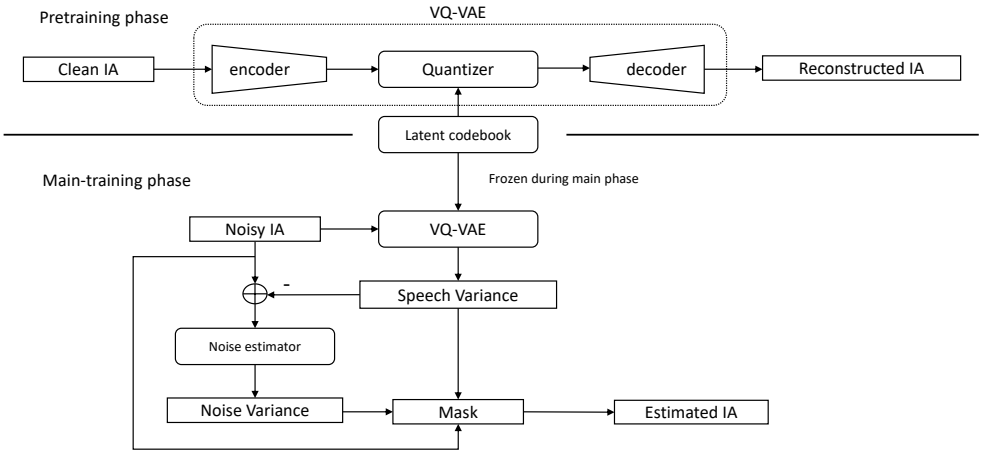


Figure 4.6: Block diagrams for instantaneous amplitude estimation.

network (CCN), followed by a block of Wavenet modules, followed by two LSTM layers. Finally, another CCN will calculate the complex output for this model. The structure of CCN is described in [4]. Fig. 4.7 illustrated this model.

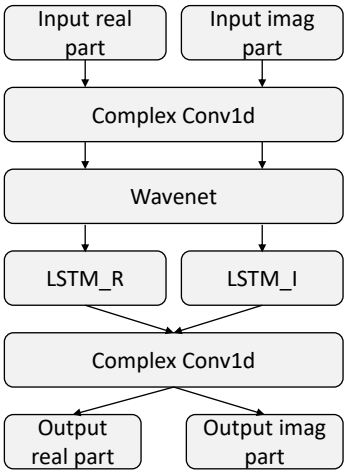


Figure 4.7: Deep complex network structure for estimating additive phase mask.

As explained in section 3.3, we use two types of input for this model in order to estimate Φ_m:

- Noisy complex coefficients as input:

By using the noisy complex gammatone coefficients $X(k,n)$ and the instantaneous amplitude mask h in Eq. 3.7. The input to the phase correction model is

$$\text{Input}_{\Re} = \Re(X(k,n)) \odot h$$
$$\text{Input}_{\Im} = \Im(X(k,n)) \odot h$$

This setup has the same idea as [42], which is: in non-speech sections, the amplitude of clean speech is very small or nearly zero. Then if the mask h is optimal then it will scale the values of complex coefficients to zero. This will limit the misbehavior of phase in those non-speech sections and will lead to a better phase estimation.

- Instantaneous phase as input for the network model to learn the additive mask.

$$\text{Input}_{\Re} = \cos\left[\phi(k,n)\right]$$
$$\text{Input}_{\Im} = \sin\left[\phi(k,n)\right]$$

From the complex output of this model, which consists of the real part o_r and imaginary part o_i, the additive phase mask is obtained by

$$\Phi_m = \arctan \frac{o_i}{o_r}, \tag{4.7}$$

This mask is then used to obtain the estimated instantaneous phase (IPh) as

$$\hat{\phi}_S = \phi_X + \Phi_m, \tag{4.8}$$

where $\hat{\phi}_S$ is estimated IPh, ϕ_X is noisy IPh and Φ_m is the additive mask, respectively.

Chapter 5

Evaluation

5.1 Dataset

The dataset used in all training and testing activities in this study is the Microsoft Scalable Noisy Speech Dataset (MS-SNSD), which was used in Interspeech 2019 Deep Noise Suppression Challenge [45]. It consists of two separate sub-sets for training and testing.
For the training set, clean speech includes over 23,000 clips of 56 speakers (28 male and 28 female speakers) reading short sentences, each clip is about 3 seconds long on average. And noise dataset includes 14 types of noise: air conditioner, announcements, washer/dryer, car noise, copy machine, door shutting, neighbor speaking, munching sound, babble, neighbor speaking, squeaky chair, traffic road, typing, vacuum cleaner.
For testing, 1000 clips from 20 speakers were recorded as clean speech dataset, each clip is about 10 seconds long. And the noise set still has the same noise types from the training set but from different recordings.
Data augmentation was used on the MS-SNSD dataset, clean speech utterances is scaled to -25 dB to improve robustness for training. To create noisy speech for training/testing purposes, a random clean speech utterance s and a random noise clip e were picked to form the noisy speech x utterance as

$$x = s + \lambda * e, \tag{5.1}$$

where λ denotes signal-to-noise ratio gain according to *snr* level, and is calculated as

$$\lambda = \sqrt{\frac{\|s\|^2}{\|e\|^2 10^{snr/10}}} \tag{5.2}$$

In the training phase, the *snr* value is randomly picked in the range [-5, 20] to create one noisy utterance. In the testing phase, we divided the testing

dataset mentioned above into five *snr* levels: -5, 0, 5, 10, and 15. So for each *snr* level, there are 200 utterances.

5.2 Evaluation metrics

Four objective metrics are used to assess the performance of the proposed method:

- Perceptual Evaluation of Speech Quality (PESQ) [46]
  A common metric used for evaluating speech quality of the target signal compared to a reference signal. The range of values is from -0.5 to 4.5, with greater values indicating better quality.

- Short-time Objective Intelligibility (STOI) [47]
  Metric used to assess speech intelligibility of the target signal compared to a reference signal. The range of values is from 0 to 1, with greater values indicating better intelligibility.

- Signal-to-Noise Ratio (SNR)
  SNR is a metric used to measure the level of speech s to the level of noise e in decibels. SNR = 0 means speech power is equal to noise power in the mixture. SNR > 0 means there is more speech than noise and SNR < 0 means there is more noise than speech.

  $$\mathrm{SNR} = 10 \log_{10} \frac{\|s\|^2}{\|s - \hat{s}\|^2} \tag{5.3}$$

  where s and $\hat{s}$ are the clean speech and the enhanced speech, respectively.

- Word Error Rate (WER)
  A metric used to evaluate how well the enhanced speech can be recognized by an ASR system compared to the clean speech. Consider two transcriptions A and B, which are output by the ASR system when the input is clean speech and reference speech (either noisy speech or enhanced speech), respectively. WER is calculated by

  $$\mathrm{WER} = \frac{S + D + I}{N} \tag{5.4}$$

  where S, D and I are the numbers of substitutions, deletions, and insertions in order to turn A to B. N is the number of words in A. The lower WER is, the better performance of reference speech in ASR

system. The ASR system used for evaluations in this study is Google Speech Recognition (GSR) API[1]. The GSR is widely used in many other practical speech-to-text-related systems, hence if the proposed method can work well on GSR, it can be directly applied on real-life applications.

All four metrics above will be used to evaluate the proposed synthesis filterbank while only PESQ, STOI, and WER will be used to evaluate the proposed speech enhancement model.

5.3 Experimental configurations

The analysis-synthesis gammatone filterbanks in this study use 33 channels. The analysis filterbank, which is implemented as in [38], has center frequency in channel 0 $f_0 = 60$ Hz and channel 33 $f_{33} = 6000$ Hz. Sampling frequency used is 16 kHz.

Both pre-train and main-train stages of VQ-VAE model use the same setups, the initial learning rate is set to $4e$-4 and decays by the factor of 0.992 to a minimum of $1e$-6. Batch size used is 4, inputs to the model are clipped to 3 seconds. Parameters for this model are described in Table 5.1 with regards to the overall structure in Fig. 4.2 and encoder/decoder structure in Fig. 4.3. The hyperparameters are in the format (*filterLength, stride*). Dilation rates for the Wavenet blocks in both encoders are (1, 2, 4, 1, 2, 4), and in decoders are (1, 2, 4). The number of embeddings in both codebooks is 1024.

For the noise estimator model, the number of filters in both encoder and decoder are 33, which means the number of channels is kept unchanged when the data goes through this model. Dilation rates for the Wavenet blocks in both encoder and decoder are (1, 2, 4).

For the phase mask estimator, input to this model is obtained by concatenating the real part and imaginary part as described in Section 4.3, so the input has the size (66 $\times$ T) and this size will keep unchanged throughout this model. The hidden size of both LSTM layers is 64. Dilation rates for the Wavenet blocks are (1, 2, 4, 1, 2, 4).

[1]https://cloud.google.com/speech-to-text

Table 5.1: Architecture of the VQ-VAE model. N denotes the total time samples of the input speech. And hyperparameters are represented in (*filterLength, stride*) format.

Block	Layer name	Input size	Hyperparameter	Output size
encoder 1	CNN_1	33 x N	5, 1	64 x N
	Resampling	64 x N	8, 2	64 x (N / 2)
	CNN_2, 3, 4	64 x (N / 2)	12, 1	64 x (N / 2)
	CNN_5	64 x (N / 2)	12, 1	64 x (N / 2)
encoder 2	CNN_1	64 x (N / 2)	5, 1	128 x (N / 2)
	Resampling	128 x (N / 2)	8, 2	128 x (N / 4)
	CNN_2, 3, 4	128 x (N / 4)	12, 1	128 x (N / 4)
	CNN_5	128 x (N / 4)	12, 1	128 x (N / 4)
decoder 2	CNN_1	128 x (N / 4)	5, 1	64 x (N / 4)
	Resampling	64 x (N / 4)	8, 2	64 x (N / 2)
	CNN_2, 3, 4	64 x (N / 2)	12, 1	64 x (N / 2)
	CNN_5	64 x (N / 2)	12, 1	64 x (N / 2)
decoder 1	CNN_1	128 x (N / 2)	5, 1	33 x (N / 2)
	Resampling	33 x (N / 2)	8, 2	33 x N
	CNN_2, 3, 4	33 x N	12, 1	33 x N
	CNN_5	33 x N	12, 1	33 x N

5.4 Results

5.4.1 Evaluation on the performance of the synthesis model

An assessment step is required to confirm that the synthesis model, which is described in section 3.1, can correctly transform speech signal from the gammatone domain back to the time domain. A set of 200 clean speech utterances is used as input to the analysis-synthesis models as described in Fig. 3.1. The reconstructed speech is then evaluated by four metrics mentioned in Section 5.2 to assess the abilities of the proposed synthesis model and is compared with the time-reversal version used in the target research [9]. The results are showed in Table 5.2.

Reconstructed speech using the proposed synthesis model has more quality

Table 5.2: Objective evaluation results for reconstructed speech using analysis-synthesis filterbanks compared with original speech in two cases: time-reversal and proposed synthesis model

Synthesis model	PESQ	STOI	SNR	WER
Time-reversal	3.99	1.0	8.7	**0.06**
Proposed	**4.37**	1.0	**20.68**	0.07

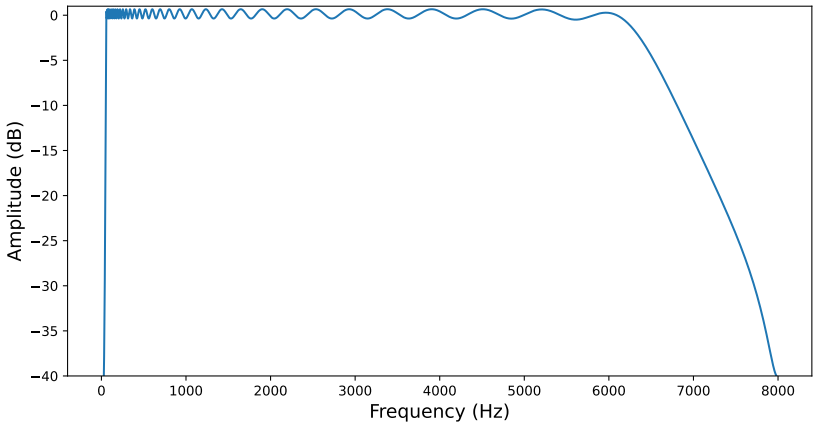
and SNR than the time-reversal synthesis model. However, the word error rate is slightly worse.

An unit impulse function is used as the input to this analysis-synthesis system to further investigate its characteristics. The output of this process is transformed to the frequency domain using fast Fourier transform. The result of this procedure is showed in Fig. 5.1.

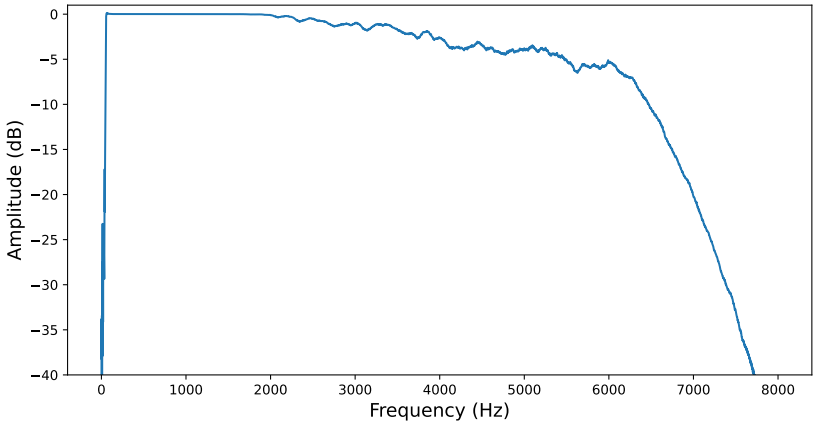
Overall, the trained synthesis filterbank still retains the initial properties for the analysis-synthesis process. Which are:

- Sampling frequency is 16 kHz, hence cutoff frequency is 8 kHz (Nyquist frequency).
- The center frequency ranges from 60 Hz at the lowest to 6 kHz at the highest.

Although it is not as perfect analytically as the time-reversal version, this proposed synthesis filterbank still performs quite well in retaining speech quality, and intelligibility. And the word error rate is just slightly worse than the time-reversal version. So we think it is an acceptable result.



(a) Time-reversal synthesis filterbank



(b) Trained synthesis filterbank

Figure 5.1: Frequency characteristics of the analysis-synthesis filterbanks. Top: when the synthesis filterbank is time-reversal of analysis filterbank. Bottom: when the synthesis filterbank is trained using proposed method.

5.4.2 Evaluation on the effectiveness of enhanced IA, IPh for speech quality, intelligibility and word error rate

The effectiveness of this proposed method is assessed in two steps in this section. In the first step, we will evaluate the effectiveness of the enhanced IA and corrected phase. Step two will involve a comparison of the proposed approach with other relevant methods. The metrics used in this section

are PESQ for speech quality, STOI for speech intelligibility, and WER for word recognition rate of ASR system. The purpose of the first stage is to evaluate the effectiveness of each sub-model (VQ-VAE, noise estimator, phase correction estimator) in this proposed method. The goal of the second stage is to compare the overall performance with other methods.

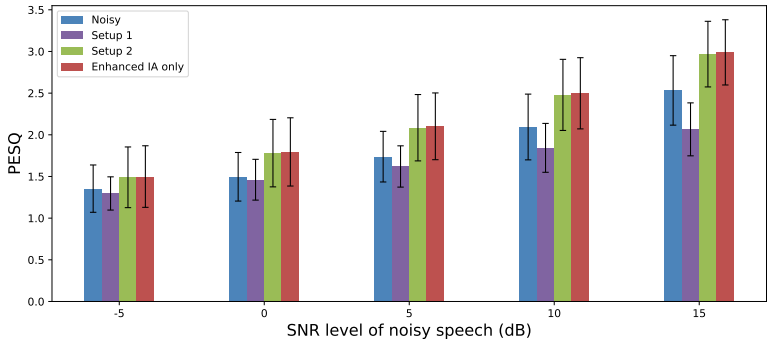
For the first stage of evaluation, we compare the noisy speech with the proposed method in 3 cases:

- Enhanced IA with complex gammatone coefficients as input for phase correction (setup 1).
- Enhanced IA with instantaneous phase as input for phase correction (setup 2).
- Enhanced IA only (noisy phase).

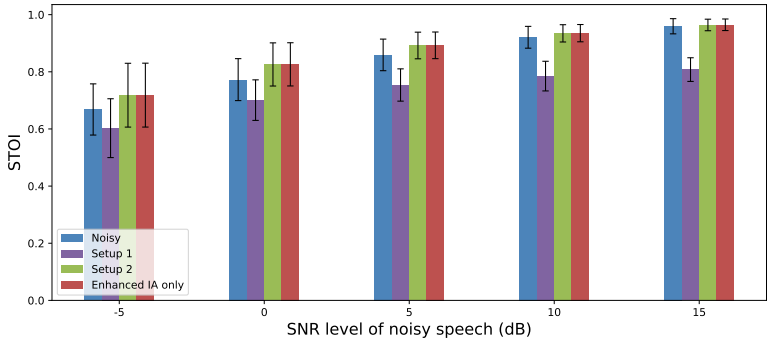
The results are shown in Fig. 5.2. As we can see, using the enhanced IA can increase speech quality and intelligibility compared with noisy speech. However, current schemes for phase correction do not have any improvements since setup 1 completely nullifies the effectiveness of enhanced IA and makes the enhanced speech perform worse than the noisy speech. Setup 2 performs better than setup 1 but it still does not effectively enhance the phase since the results are not better than noisy phase. For the performance of enhanced speech in ASR system, current method still cannot reduce the word error rate of noisy speech.

In the second stage of evaluation, we compare the proposed method with three other methods: DCCRN [4], VQ-VAE on STFT domain (STFT-VQVAE) [42] and the current baseline method of deep noise suppression challenge (DNS-baseline) [48]. Since the correction of phase is not good (as showed in the first stage), we will use enhanced IA with noisy phase to form the enhanced speech and compare with these three methods. The results are shown in Fig. 5.3. In terms of speech quality (PESQ) and intelligibility (STOI), the proposed method is better than DCCRN and STFT-VQVAE in most cases, from very-noisy speech (SNR = -5 dB) to less-noisy speech (SNR = 15 dB). However, the baseline method of DNS Challenge is still better.

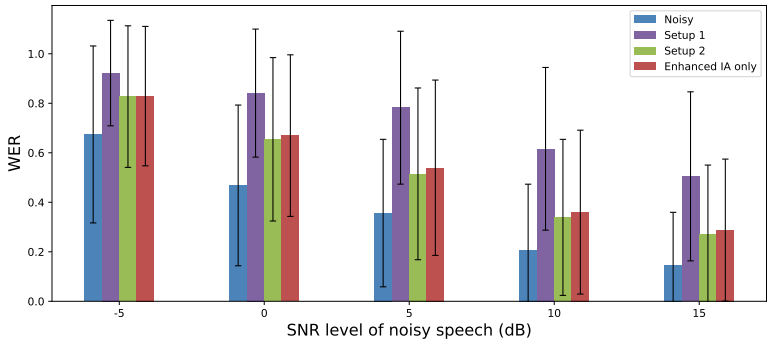
Regarding the performance of enhanced speech in ASR system, all four methods perform worse than noisy speech since word error rates (WERs) are all higher than WER of noisy speech. The proposed method performs the worst regardless of PESQ and STOI results are higher than DCCRN and STFT-VQVAE.



(a) PESQ

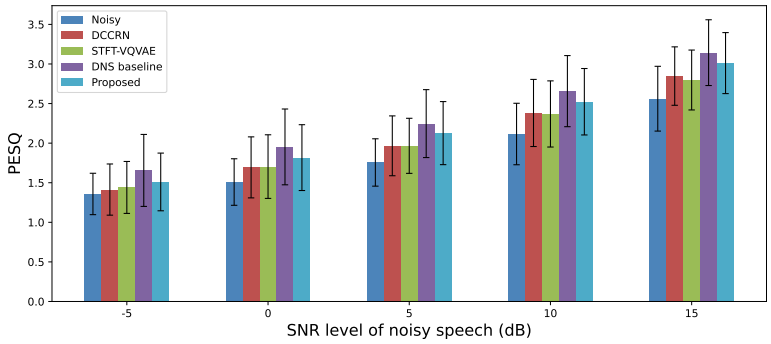


(b) STOI

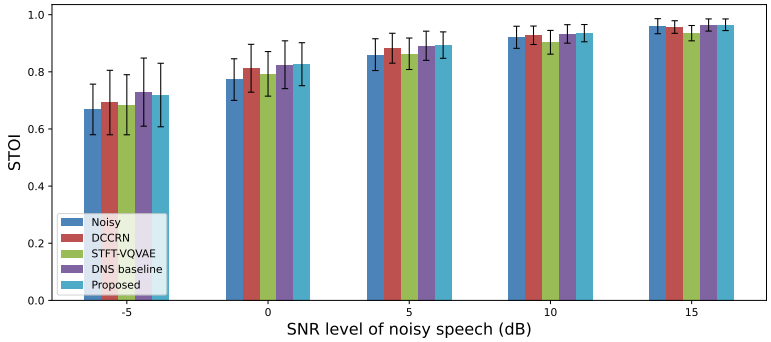


(c) WER

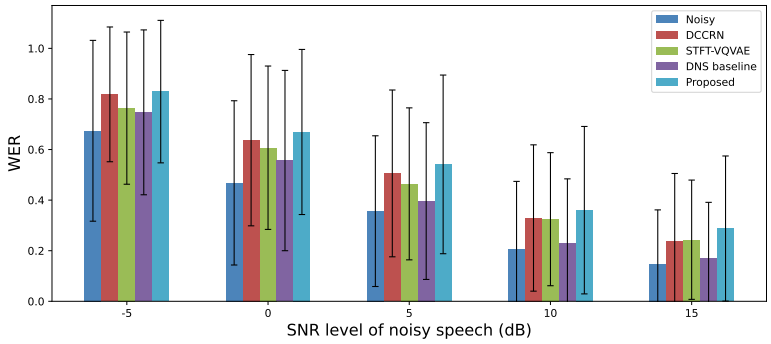
Figure 5.2: Evaluation results for setup 1 (phase correction using complex gammatone coefficients as input), setup 2 (phase correction using instantaneous phase as input) and enhanced instantaneous amplitude only



(a) PESQ



(b) STOI



(c) WER

Figure 5.3: PESQ, STOI and WER of the proposed method in comparison to DCCRN, STFT-VQVAE and current baseline method of DNS Challenge.

5.5 Discussion

From the evaluation results, the proposed method can improve quality as well as intelligibility of speech under noisy environments by enhancing instantaneous amplitude. Using current deep learning techniques for phase enhancement, however, does not perform very well. Regarding the gammatone domain with current techniques for phase correction, using instantaneous phase as input features seems to be better than complex coefficients since complex coefficients tend to make the network to over-predict the amount of correction and further damage the speech quality and intelligibility.

Although the currently proposed scheme for phase correction does not work well for objective evaluation metrics like PESQ and STOI, it is still promising for ASR system. As shown in Fig. 5.2c, using instantaneous phase as input for phase correction model can reduce the word error rate compared to just using noisy phase. With a more proper model, instantaneous phase could be a strong feature for improving the performance of enhanced speech in ASR systems.

Regarding the performance comparison with other methods, by using auditory filterbank features, the VQ-VAE model performs better than STFT features in terms of PESQ and STOI. However, the proposed method is still not better than the current baseline model of the DNS challenge. In the evaluation phase, only DCCRN and STFT-VQVAE were trained on the same dataset[2] as the proposed method, while the DNS baseline was trained on a much bigger dataset[3]. This could also be a factor that lead to the current results.

As for the performance of enhanced speech in ASR system, all four methods are worse than noisy speech. Again, the main reason for this could be the over-suppression of each method that causes the enhanced speech to have distortions and lose its naturalness.

[2]https://github.com/microsoft/MS-SNSD
[3]https://github.com/microsoft/DNS-Challenge

Chapter 6

Conclusion

6.1 Summary

In this study, a method for speech enhancement in noisy environments is proposed by restoring instantaneous amplitude (IA) and instantaneous phase (IPh), which are extracted from the gammatone filterbank. Due to the main method being a deep learning model and both input and output waveforms contributing to the optimization process, the backpropagation algorithm should be able to run on analysis and synthesis gammatone filterbanks. Therefore, we proposed a method to train the synthesis model while the parameters of the analysis model are initialized to be the same as in the target method [9]. The results of this newly implemented analysis-synthesis block showed that they are acceptable as a means for feature extraction.

As for the speech enhancement model. The cross-channel information in both IA and IPh is utilized by using convolutional layers to estimate clean IA and clean IPh with the expectation that this information could help elevate the performance of enhanced speech in ASR systems. To restore IA, we estimate a mask that is similar to the ideal ratio mask [3]. To obtain this mask, two parameters need to be estimated: speech variance and noise variance.

By pre-training VQ-VAE model on clean speech dataset, the codebook can capture the characteristics of speech in the latent space. These characteristics should remain the same in noisy speech. During main-training phase, the latent codebook is frozen to keep the pre-trained speech information unchanged. Speech variance is obtained from noisy speech using this VQ-VAE model. On the other hand, noise variance is obtained by a convolutional neural network. With these two parameters, estimated IA can be recognized.

For obtaining clean instantaneous phase, being motivated by [4, 37, 42], the

complex convolution network with the SSISNR optimization function is proposed to estimate the additive mask for phase correction. We used two kinds of input to this model: the complex gammatone coefficients and complex components of noisy instantaneous phase.

The proposed model is objectively evaluated by three metrics: PESQ (for speech quality), STOI (for speech intelligibility), and WER (to evaluate the word recognition rate of ASR systems). The results showed that by enhancing speech using auditory filterbank features, the proposed method performs better than some referenced methods for enhancing the quality and intelligibility of noisy speech. However, the performance of enhanced speech in ASR systems is still deteriorated. Furthermore, the results showed that the instantaneous phase is a better feature for phase estimation than the complex coefficients. With a more suitable model, enhanced instantaneous phase can help improve the word recognition rate for ASR systems.

6.2 Future works

Despite using auditory filterbank features, the proposed method still cannot increase the performance of enhanced speech in ASR systems although speech quality and intelligibility were improved. This issue could be caused by two reasons. First, the encoders in VQ-VAE models cannot capture the latent characteristics of speech in the mixture well enough. And second, speech quality and intelligibility do not reflect how well an utterance performs in ASR systems. Therefore, a more proper architecture for the encoder should also be investigated. In the evaluation step, only one ASR system was used, so the proposed method should also be evaluated on other ASR systems as well - such as Microsoft Azure Speech[1] or IBM Speech to text[2],... - to avoid bias in evaluation results. On the other hand, research on which factors of speech affect the word recognition rate should be carried out. In addition, the phase information was not enhanced much using the proposed model. So an effective method for clean phase estimation should also be brought into the discussion.

[1]https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/
[2]https://www.ibm.com/cloud/watson-speech-to-text

Bibliography

[1] Sawata R., Kashiwagi Y. and Takahashi S., "Improving Character Error Rate is Not Equal to Having Clean Speech: Speech Enhancement for ASR Systems with Black-Box Acoustic Models," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 991-995, 2022.

[2] Saleem, N., Khattak, M. I., Verdú, E., "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms," International Journal of Interactive Multimedia and Artificial Intelligence, 78-89, 2020.

[3] Wang D., Chen J., "Supervised Speech Separation Based on Deep Learning: An Overview," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702-1726, Oct. 2018.

[4] Yanxin H., Yun L., Shubo L., Mengtao X., Shimin Z., Yihui F., Jian W., Bihong Z., and Lei X., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in Proceedings of Interspeech, 2472– 2476, 2020.

[5] Drullman, R., "Temporal envelope and fine structure cues for speech intelligibility." J. Acoust. Soc. Am. 97 (1), 585–592, 1995.

[6] Moore, B. C. J., "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people." J. Assoc. Res. Otolaryngol. 9 (4), 399–406, 2008.

[7] Swaminathan, J., Heinz, M.G., "Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise". J. Neurosci. 32 (5), 1747–1756, 2012.

[8] Rory A. D., Claus P. J., and Tom F. (1996). "Frequency importance functions for words, sentences, and continuous discourse," Journal of Speech, Language, and Hearing Research, vol. 39, no. 4, pp. 714–723.

[9] Nower, N., Liu, Y., Unoki, M., "Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement". Speech Commun. 70, 13–27, 2015.

[10] Shrawankar U. and Thakare V., "Feature Extraction for a Speech Recognition System in Noisy Environment: A Study," Second International Conference on Computer Engineering and Applications, 358-361, 2010.

[11] Boll, S. F. , "Suppression of acoustic noise in speech using spectral subtraction." IEEE Trans. on Acoustic, Speech and Signal Processing ASSP-27, 1979.

[12] Paliwal K., Wójcicki K., and Schwerin B., "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Communication, vol. 52, no. 5, pp. 450-475, 2010.

[13] Lim J. S., Oppenheim A. V., "Enhancement and bandwidth compression of noisy speech," Proceedings of the IEEE, vol. 67, no. 12, pp. 1586-1604, 1979.

[14] Ding H., Soon Y., Koh S. N., and Yeo C. K., "A spectral filtering method based on hybrid wiener filters for speech enhancement," Speech Communication, vol. 51, no. 3, pp. 259-267, 2009.

[15] Scalart, P., Filho, J.V., "Speech enhancement based on a priori signal to noise estimation". In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processings (ICASSP), pp. 629–623, 1996.

[16] Ephraim, Y. , Malah, D. , "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". IEEE Trans. Acoust. Speech Signal Process, 32 (6), 1109–1211, 1984.

[17] Gerkmann T. and Krawczyk M., "MMSE-optimal spectral amplitude estimation given the STFT-phase," IEEE Signal Processing Letters, vol. 20, no. 2, pp. 129-132, 2012.

[18] Li Y., Zhang X., Sun M., Min G., "Unsupervised monaural speech enhancement using robust NMF with low-rank and sparse constraints," 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 1-4, 2015.

[19] Mohammed S., Tashev I., "A statistical approach to semisupervised speech enhancement with low-order non-negative matrix factorization," in IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 546–550, 2017.

[20] Patterson, R.D., Allerhand, M., Giguere, C., "Timedomain modelling of peripheral auditory processing: a modular architecture and a software platform." J. Acoust. Soc. Amer. 98, 1890-1894, 1995.

[21] Irino T., Patterson R. D., "A time-domain, level-dependent auditory filter: The gammachirp," The Journal of the Acoustical Society of America, vol. 101, pp. 412–419, 1997.

[22] Lin, L., Holmes, W.H., Ambikairajah, E. "Speech enhancement based on a perceptual modification of wiener filtering". Proc. 7th International Conference on Spoken Language Processing, 781-784, 2002.

[23] Kianyfar A., Abutalebi H. R., "Improved speech enhancement method based on auditory filterbank and fast noise estimation," 7'th International Symposium on Telecommunications, pp. 441-445, 2014.

[24] Kortlang S., Ewert S. D. and Gerkmann T., "Single channel noise reduction based on an auditory filterbank," 14th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 283-287, 2014.

[25] Pandey A., Wang D., "TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain," 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6875-6879, 2019,

[26] Pascual S., Bonafonte A., Serra J., "SEGAN: Speech enhancement generative adversarial network," in Proceedings of Interspeech. pp. 3642-3646, 2017.

[27] Hu G., Wang D. L., "Speech segregation based on pitch tracking and amplitude modulation," in Proceedings of IEEE WASPAA, pp. 79-82, 2001.

[28] Hu G., Wang D. L., "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Net., vol. 15, pp. 1135-1150, 2004.

[29] Narayanan A., Wang D. L., "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in Proceedings of ICASSP, pp. 7092-7096, 2013.

[30] Hummersone C., Stokes T., Brooks T., "On the ideal ratio mask as the goal of computational auditory scene analysis," in Blind Source Separation, G.R. Naik and W. Wang, Ed., Berlin: Springer, pp. 349-368, 2014.

[31] Wang Y., Narayanan A., Wang D. L., "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 22, pp. 1849-1858, 2014.

[32] Shannon, B.J., Paliwal, K.K., "Role of phase estimation in speech enhancement." In: Proceedings of IEEE SAPA@ INTERSPEECH, pp. 1427–1430, 2006.

[33] Paliwal, K.K., Wojcicki, K., Shannon, B., "The importance of phase in speech enhancement." Speech Commun. 53 (4), 465–494, 2011.

[34] Erdogan H., Hershey J., Watanabe S., Le Roux J., "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in Proceedings of ICASSP, pp. 708-712, 2015.

[35] Lee J., Kang H. G., "A Joint Learning Algorithm for Complex-Valued T-F Masks in Deep Learning-Based Single-Channel Speech Enhancement Systems," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 1098-1108, 2019.

[36] Li, X., Horaud, R. "Online Monaural Speech Enhancement Using Delayed Subband LSTM." Proc. Interspeech 2020, 2462-2466, 2020.

[37] Sun, Y., Yang, L., Zhu, H., Hao, J., "Funnel Deep Complex U-Net for Phase-Aware Speech Enhancement". Proc. Interspeech 2021, 161-165, 2021.

[38] Unoki, M., Akagi, M., "A method of signal extraction from noisy signal based on auditory scene analysis". Speech Commun. 27 (3), 261– 279, 1999.

[39] Patterson, R. D. and Holdsworth, J. L., "A functional model of neural activity patterns and auditory images," Advances in Speech, Hearing and Language Processing , (W. A. Ainsworth, ed.), Vol 3. JAI Press, London, 1991.

[40] Goodfellow I., Bengio Y., Courville A., "Deep Learning". MIT Press, 2016.

[41] Nemer E., "Audio Cochleogram with Analysis and Synthesis Banks Using 1D Convolutional Networks," 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), pp. 42-47, 2021.

[42] Ho T. V., Nguyen Q. H., Akagi M. and Unoki M., "Vector-quantized Variational Autoencoder for Phase-aware Speech Enhancement". Proc. Interspeech 2022, 176-180, 2022.

[43] Oord A. v. d., Vinyals O., Kavukcuoglu K., "Neural Discrete Representation Learning." in Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6309–6318, 2017.

[44] Oord A. v. d., Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., Kalchbrenner N., Senior A., Kavukcuoglu K., "WaveNet: A Generative Model for Raw Audio". Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), 125, 2016.

[45] Reddy C. K. A., Beyrami E., Pool J., Cutler R., Srinivasan S., Gehrke J., "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," in Interspeech 2019, pp. 1816–1820, 2019.

[46] Rix A. W., Beerends J. G., Hollier M. P., Hekstra A. P., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), pp. 749-752 vol.2, 2001.

[47] Taal C. H., Hendriks R. C., Heusdens R., Jensen J., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4214-4217, 2010.

[48] Braun, S., Tashev, I. "Data Augmentation and Loss Normalization for Deep Noise Suppression." In: Karpov, A., Potapova, R. (eds) Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science, vol 12335. Springer, Cham. 2020.