Doctoral dissertation

Speech Enhancement based on Amplitude and Frequency Modulation Characteristics

NGUYEN, Huy Quoc

Supervisor UNOKI Masashi

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

September, 2024

Abstract

Clear and intelligible speech is vital for effective human communication, particularly in critical systems like airport communication, where low speech intelligibility can lead to severe consequences. Speech enhancement techniques are crucial for improving speech quality and intelligibility in various real-world applications such as telecommunication, hearing aids, and voice recognition systems, as well as in military and aviation communications where clarity is essential. Despite having a long history of studies, recent speech enhancement techniques still suffer speech over suppression and noise under suppression, distorting the enhanced speech signals, which sometimes have lower quality and intelligibility than the noisy speech itself.

Believing that the gap between mathematical/computational techniques and the nature of speech is the cause of this distortion, this study utilizes the concept of modulation for speech enhancement to build a bridge to connect this gap. The main objective of this research is to investigate the effectiveness of utilizing speech modulation characteristics for enhancement. This main objective contains three sub-objectives: to model the amplitude modulation characteristics for speech enhancement, derive the relationship between amplitude and instantaneous frequency modulation, and enhance speech using the derived relationship.

To achieve the first objective, a method to model the spectral modulation characteristics of speech in amplitude is proposed and applied for speech enhancement. In voiced speech, the speech power spectrum is amplitude-modulated, where the spectral fine structure is periodic with a period equal to the fundamental frequency. Thus, the proposed method constructs the categorical distribution of fundamental frequency to characterize spectral fine-structure characteristics of speech. Evaluating the Valentini et al. dataset, the results show that improving amplitude modulation characteristics im-

proves speech enhancement performance.

The analytical derivative method is proposed to extract instantaneous frequency deviation (IFD) to achieve the second objective. By deriving the principal value of the logarithm of the complex time-frequency representation, an equation connecting the amplitude to the IFD is established. Via single-tone frequency-modulated signals, the proposed method is verified to work correctly, which confirms the proposed equation's validity. As the established equation indicates, this result confirms a connection between amplitude and IFD.

The findings in the second objective provide two critical perspectives on IFD. First, although defined from the phase, IFD has a multiplicative connection with the amplitude, which allows real-valued processing. Second, computationally, the IFD can be derived instantaneously without a time difference. From these findings, a method to enhance speech via IFD is proposed to modify IFD by a learnable affine transform at the frame-wise level. Evaluating the Valentini et al. dataset, the results show that the proposed method improves speech enhancement performance, especially quality. Specifically, the proposed method achieves the Perceptual Evaluation of Speech Quality of about 2.87 and Short-Time Objective Intelligibility of 0.94, outperforming many state-of-the-art techniques in speech enhancement. Significantly, the proposed method improves up to 15% in a 2.5 dB signal-to-noise ratio. These results confirm the effectiveness of using IFD in speech enhancement based on its relationship with amplitude.

All the results confirm that utilizing speech's modulation characteristics can improve speech enhancement performance, satisfying the research objective. This research has established a solid base by showing how effective modulation characteristics can improve speech quality in noisy conditions. This research has practical applications in improving user experience in mobile calls, VoIP services, and video conferencing, as well as benefiting assistive technologies such as hearing aids and cochlear implants. Additionally, it contributes to advancements in audio signal processing, machine learning, artificial intelligence, and neuroscience.

Keywords: speech enhancement, amplitude modulation, spectral modulation, frequency modulation, instantaneous frequency deviation

Acknowledgment

I extend my heartfelt gratitude to Prof. Masashi Unoki for his invaluable guidance and unwavering support throughout my research journey. His expertise has been instrumental in enabling me to explore the fascinating field of science and complete this dissertation.

I am also profoundly grateful to my minor research supervisor, Prof. Minh-Le Nguyen, for his insightful feedback and contributions to my research. His perspectives and expertise have greatly enriched my work.

Furthermore, I would like to express my gratitude to Prof. Masato Akagi for his valuable feedback and comments during his tenure at JAIST. His insightful perspectives have made a significant contribution to my research.

I am genuinely thankful to my colleagues and friends who have accompanied me on this academic journey. Together, we engage in academic discussions, envision the future, and support one another. Your companionship has enriched this experience with knowledge and made it enjoyable.

Lastly, my sincere thanks go to my family, my mother and father, for their unwavering support and encouragement throughout my doctoral studies. Their belief in me has been a constant source of strength.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Speech enhancement

Speech is essential for conveying thoughts, emotions, and information, serving as a fundamental means of human communication. It allows people to express themselves, share ideas, and connect with others personally and professionally. In an era where digital communication is paramount, ensuring clear and intelligible speech transmission has become increasingly crucial. In critical systems, digital speech communication is vital in ensuring a precise and reliable transmission of crucial information. For example, airport communication requires a high-intelligibility communication system, including air traffic control instructions, pilot communications, and emergency announcements, where low intelligibility of speech causes serious consequences.

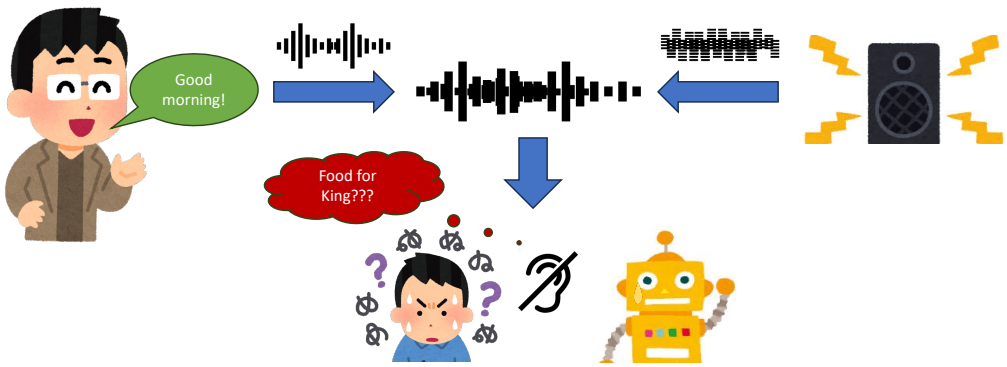

Figure 1.1: Illustration of speech enhancement problem.

Speech enhancement techniques, which aim to improve the quality and intelligibility of speech signals, are crucial to achieving such a high-intelligibility communication system. It involves using various algorithms and methods to reduce background noise, minimize distortion, and enhance the overall clarity of speech. Speech enhancement is vital in numerous real-world applications, such as telecommunications, hearing aids, voice recognition systems, and audio forensics. By effectively removing unwanted noise and improving essential speech components, speech enhancement techniques improve communication and user experience in various audio-related technologies. The applications of speech enhancement are extensive and varied, reflecting the ubiquity of speech communication in modern life. In telecommunications, enhanced speech quality can significantly improve user experience, particularly in noisy environments like public transportation or crowded areas. Voice-activated systems, such as virtual assistants (e.g., Siri, Alexa), rely on clear speech input to function accurately, making noise suppression essential for their reliability. In healthcare, hearing aids benefit from advanced speech enhancement techniques to provide users with higher-quality auditory signals, improving their quality of life. In addition, in military and aviation communications, where clarity is critical, speech enhancement can ensure effective and accurate information exchange.

1.2 Challenges

Although speech enhancement techniques have a long history of studies, recent speech enhancement techniques still suffer speech over suppression and noise under suppression, distorting the enhanced speech signals, which sometimes have lower quality and intelligibility than the noisy speech itself [1]. After reviewing several speech enhancement techniques to analyze the cause of why this happens, speech enhancement techniques can be divided into three approaches:

- Mathematical approach: The techniques in this category introduce the mathematical grounds on how to separate a signal from another signal, such as spectral subtraction techniques which try to subtract noise from the mixture [2–7], statistical techniques which try to formulate a conditional statistical estimation problem to estimate the speech given

the mixture [8–14], or subspace techniques which try to decompose the mixture into speech and noise components using different bases [15–19]. The foundation of these techniques comes from two mathematical assumptions: additivity, where the noise is additive, and speech-noise independence, where the noise and speech have zero correlation. However, these techniques are usually limited by some assumptions, such as stationary noise, and can hardly handle arbitrary or unknown noise.

- Computational approach: The techniques in this category are only developed recently and are based on two main factors: powerful numerical estimator (deep neural network) and large-scale dataset. These techniques extend previous techniques in the mathematical approach to improve their generalization [20–28].

- Nature-of-speech approach: These studies investigate the different characteristics of speech along time and frequency based on speech production and perception mechanism [29–36]. While these features are essential for speech, enhancing them from noisy speech is challenging due to the lack of mathematical framework.

In summary, the categorization above shows a gap between mathematical/computational techniques and the nature of speech. In other words, there needs to be a unified framework for enhancing essential features of speech using powerful computational and mathematical tools.

1.3 Motivation and research goals

Speech contains structural information such as linguistic information, emotion, and speaker identity. The variance of such information is much less than acoustic variability in speech signals. In other words, some *simple features* inside the enormous variance of the speech signal affect the quality and intelligibility of speech, and enhancing such features is enough to enhance speech quality.

As the signal in the waveform does not offer a clear picture of the variances of the signal components, this study focuses on the speech signal in a complex time-frequency representation, which provides a detailed view of

how the signal's spectral content, including amplitude and phase, evolves. Although the amplitude exhibits clear patterns that reflect the energy distribution of different frequency components at each moment, the phase appears more chaotic and less visually interpretable. However, both amplitude and phase are required to construct a high-quality speech waveform. Therefore, a research question is raised: 'Does enhancing the speech features in the time-frequency representation lead to improving speech enhancement performance?' This question can be decomposed into three sub-questions:

1. Can enhance the speech characteristics in amplitude improve speech enhancement performance?

2. Is there a relationship between the amplitude and phase?

3. Can the relationship between amplitude and phase help improve speech enhancement performance?

This study employs the concept of modulation to seek the answer to the research questions mentioned above. Modulation, including amplitude modulation and frequency modulation, refers to a technique in communication theory that allows the transmission of a low-frequency message signal over a long distance using a high-frequency carrier signal. Thus, the transmitted signal is high-frequency and complex for transmission, while the message is simple. From this analogy, speech can be considered as a sound wave modulated by the speech features, where the sound wave is the carrier with a large variability. Not only in analogy, speech signals are mathematically amplitude- and frequency-modulated in the time-frequency representation. The modulation characteristics of the amplitude appear in both the time and frequency axes, while the modulation characteristics of the frequency (or *instantaneous frequency deviation*) allow us to analyze the phase more effectively (see Chapter 3). These properties show the potential of the modulation concept to connect the gap between mathematical/computational speech enhancement techniques and the nature of speech.

Therefore, the main objective of this research is to investigate the effectiveness of utilizing modulation characteristics of speech for enhancement. This main objective contains three sub-objectives, each of which seeks the answer to the research questions above:

1. Model the amplitude modulation characteristics for speech enhancement
2. Derive the relationship between amplitude and the instantaneous frequency modulation
3. Enhance speech using the derived relationship

1.4 Novelty and significance

This study utilizes the concept of modulation for speech enhancement, building a bridge to connect the gap between mathematical/computational speech enhancement techniques and the nature of speech. In technical detail, this study proposes a method to quantify the spectral modulation characteristics of speech amplitude, serving as a loss function for deep learning-based speech enhancement. Also, this study establishes an equation connecting the amplitude and instantaneous frequency deviation, introducing a new viewpoint of instantaneous frequency deviation beyond the phase. Based on these findings, this study proposes a speech enhancement method to enhance the instantaneous frequency deviation of speech without circular data processing on the phase.

The findings of this study and the proposed techniques offer valuable insights and methodological advancements for signal processing research. They allow for modifying and analyzing the phase information of signals in complex-valued domains without wrapping issues. These findings can be applied to several types of signals beyond speech.

Furthermore, this study proposes a speech enhancement method that effectively removes noise from speech. The proposed method could be used to build a practical speech enhancement application, impacting various aspects of daily life, technology, and communication.

1.5 Organization of dissertation

This dissertation contains seven chapters, the rest of which are organized as follows:

- Chapter 2 introduces related work in speech enhancement, including speech enhancement techniques and the methods to evaluate the performance of these techniques.

- Chapter 3 explains the modulation theory, including the concept of modulation, and evidence of modulation in speech, including amplitude and frequency modulation.

- Chapter 4 proposes a method to enhance the spectral modulation characteristics in speech amplitude and applies the proposed method in speech enhancement using Valentini *et al.* dataset [37]. The chapter seeks to answer the first research question: 'Can enhancing the speech characteristics in amplitude improve speech enhancement performance?'

- Chapter 5 investigates the relationship between the amplitude and the instantaneous frequency deviation to reveal the relationship between the amplitude and the phase in the complex time-frequency representation and evaluate the relationship using single-tone frequency-modulated signals. The chapter seeks to answer the second research question, 'Is there a relationship between amplitude and phase?'

- Chapter 6 proposes a speech enhancement method by enhancing the instantaneous frequency deviation of speech. The method is evaluated using Valentini *et al.* dataset [37]. The chapter seeks to answer the third research question: 'Can the relationship between amplitude and phase help improve speech enhancement performance?'

- Chapter 7 summarizes the dissertation, including the insights and contributions revealed in the study.

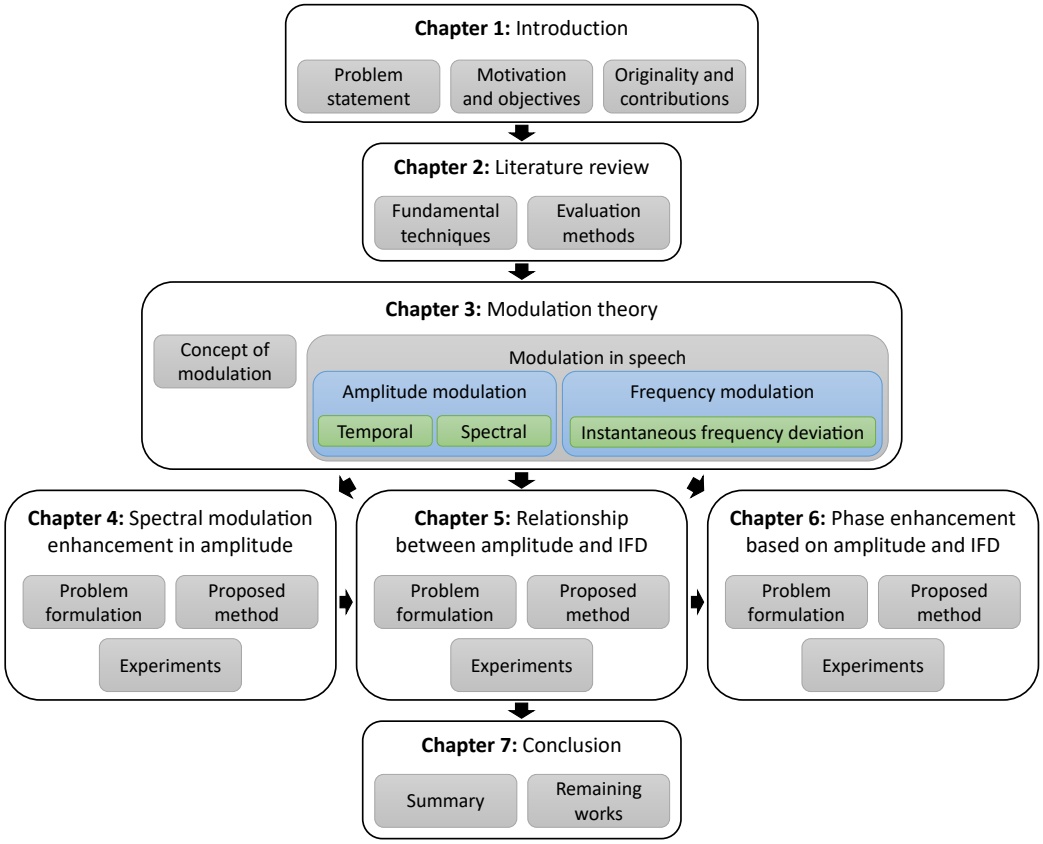

Figure 1.2: Structure of dissertation.

Chapter 2

Literature review

This chapter provides the fundamental principles for speech enhancement. The first section reviews the development of speech enhancement techniques in different aspects. In the second section, the chapter briefly introduces how to evaluate the performance of a speech enhancement method.

2.1 Techniques in speech enhancements

Speech quality and intelligibility deteriorate due to additive background noise. Researchers have proposed various methods for improving speech quality and intelligibility under such conditions. From traditional techniques, speech enhancement methods have evolved to handle multiple types of noise by incorporating deep neural networks. In addition, several studies investigate the nature of speech and its features. This section first introduces the traditional techniques in speech enhancement, then describes some feature enhancement techniques, and finally briefly reviews state-of-the-art speech enhancement methods incorporating deep learning techniques.

2.1.1 Traditional techniques: mathematical grounds

The additive assumption of the background noise gives the following stochastic process

$$y(t) = s(t) + n(t)\,, \tag{2.1}$$

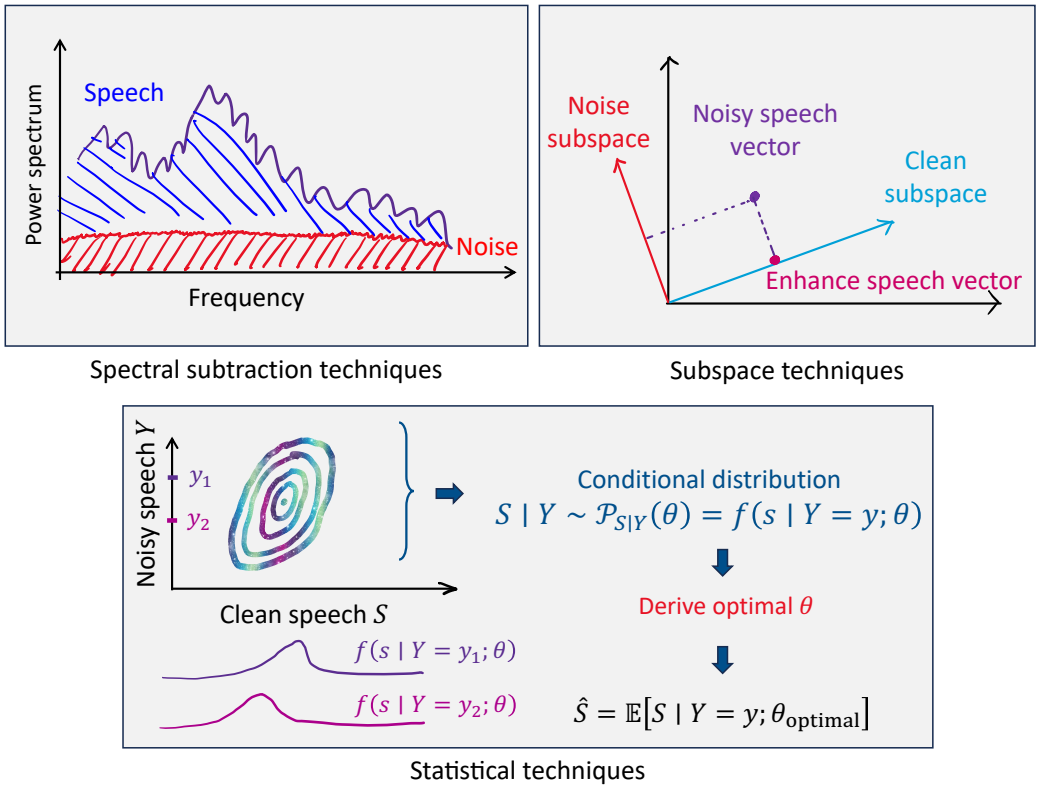

Figure 2.1: Tradition techniques in speech enhancement.

where $s(t)$, $n(t)$, and $y(t)$ are the clean, additive noise, and noisy observed signals in the time domain, respectively. Based on this assumption, several researchers have proposed several speech enhancement techniques to improve speech quality and intelligibility from the noisy observed signal. These techniques provide the mathematical ground and principles to construct speech enhancement methods, hence 'traditional.' According to Loizou [38], these techniques can be classified into three main classes (see Figure 2.1) as follows:

- Spectral-subtraction techniques are simple techniques that directly subtract the noise spectra from the noisy spectra as the noise is additive, such as Weiss *et al.*'s [2] or Boll's [3] spectral subtraction.

- Statistical techniques consider speech enhancement a statistical estimation problem to estimate the speech parameters given the noisy signal. Two representative techniques are minimum-mean-squared-error short-time spectral amplitude (MMSE-STSA) by Ephraim and Malah [10]

and Wiener filtering developed by Lim and Oppenheim [8, 9].

- Subspace techniques consider speech enhancement a linear decomposition problem. Specifically, the assumption is that the speech signals come from a subspace of an Euclidean space, which is orthogonal to the subspace forming the noise signals in the same Euclidean space. Some example techniques are the singular-value decomposition [15] or eigenvalue decomposition [18] in the time domain, or non-negative matrix factorization [19] in the time-frequency domain of power spectrogram.

Spectral-subtraction techniques

The spectral subtraction technique is one of the first and simplest noise reduction techniques. This technique comes from the intuition that the noise is additive, so the denoising process should be subtractive. This technique utilizes two important assumptions, which are

- The noise is assumed to be stationary; in other words, the noise spectrum does not significantly change as time evolves.
- The speech and noise signals are zero mean and stochastically independent, i.e., $\mathbb{E}\left[s(t)n(t+\tau)\right] = 0$ for all τ.

From the stationary-noise assumption, the spectral subtraction techniques involve two stages:

1. Noise estimation: estimating the noise power spectrum during the periods where the speech signal is absent,
2. Spectral subtraction: subtracting the noise power spectrum from the noisy power spectrum.

The foundation of spectral subtraction is as follows. Let $S(\omega)$, $N(\omega)$, and $Y(\omega)$ be the complex spectra of speech signal $s(t)$, noise signal $n(t)$, and noisy signal $y(t)$, respectively, obtained by the Fourier transform, i.e.,

$$S(\omega) = \mathcal{F}\left\{s(t)\right\}, \tag{2.2}$$

$$N(\omega) = \mathcal{F}\left\{n(t)\right\}, \tag{2.3}$$

$$Y(\omega) = \mathcal{F}\left\{y(t)\right\}. \tag{2.4}$$

where $\mathcal{F}\{f(t)\}$ denotes the Fourier transform of a continuous-time signal $f(t)$ as

$$\mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt\,. \tag{2.5}$$

As Fourier transform conserver additivity, the Eq. (2.1) becomes

$$Y(\omega) = S(\omega) + N(\omega)\,, \tag{2.6}$$

which allows to express the noisy power spectrum as follows

$$\begin{aligned} |Y(\omega)|^2 &= |S(\omega) + N(\omega)|^2 \\ &= |S(\omega)|^2 + |N(\omega)|^2 + S(\omega)N^*(\omega) + S^*(\omega)N(\omega)\,, \end{aligned} \tag{2.7}$$

where the notation u^* denotes the complex conjugate of $u \in \mathbb{C}$. Then, the clean speech power spectra can be estimated as follows

$$\left|\hat{S}(\omega)\right|^2 = |Y(\omega)|^2 - \mathbb{E}\left[|N(\omega)|^2 + S(\omega)N^*(\omega) + S^*(\omega)N(\omega)\right]\,. \tag{2.8}$$

Under the stochastical independence assumption, $\mathbb{E}[S(\omega)N^*(\omega)] = 0$ and $\mathbb{E}[S^*(\omega)N(\omega)] = 0$, which reduces Eq. (2.8) to become

$$\left|\hat{S}(\omega)\right|^2 = |Y(\omega)|^2 - \left|\hat{N}(\omega)\right|^2\,, \tag{2.9}$$

where $\left|\hat{N}(\omega)\right|^2$ is the estimated noise during the noise estimation stage. The time-domain signal can be reconstructed using inverse Fourier transform.

It is possible to generalize the spectral subtraction to a p-power as

$$\left|\hat{S}(\omega)\right|^p = |Y(\omega)|^p - \left|\hat{N}(\omega)\right|^p\,, \tag{2.10}$$

where $p = 1$ returns the Boll's spectral subtraction [3] in which the magnitude spectrum is subtracted. Besides, there are several extended versions of spectral subtraction to improve the enhancement performance, such as selective spectral subtraction [4, 5] or multi-band spectral subtraction [6, 7].

Although spectral subtraction techniques are simple and computationally efficient, the enhancement results may yield distortion, namely musical noise and reduced speech quality. In addition, spectral subtraction techniques have

two restrictions: the 'subtraction' operator in the 'spectral' domain, neither of which is necessary to be optimal for enhancement. Hence, later traditional techniques aim to provide better performance by changing the view on the enhancement problem.

Statistical techniques

Statistical techniques refer to the speech enhancement techniques that estimate the speech signal or its representation, such as the time-domain signal in the Wiener filtering [8] or magnitude spectrum in the MMSE-STSA [10]. In principle, if the estimation target is a random variable S given the noisy observation $Y = y$, then a point estimation of S can be obtained using the conditional expectation as follows

$$\begin{aligned} \hat{s}_{\text{optimal}} &= \mathbb{E}\left[S \mid Y = y\right] \\ &= \int s\, f_{S|Y}\left(s \mid y\,;\,\theta\right) ds\,, \end{aligned} \tag{2.11}$$

where $f_{S|Y}\left(s \mid y\,;\,\theta\right)$ is the conditional probability density function for the conditional distribution with some optional parameter θ as follows

$$S \mid Y \sim \mathcal{P}_{S|Y}\left(\theta\right). \tag{2.12}$$

Thus, the estimation problem becomes determining the optimal parameter θ. One of the most commonly used solutions for this problem is the Maximum Likelihood Estimation (MLE), which yields

$$\theta_{\text{optimal}} = \arg\max_{\theta} f_{S|Y}\left(s \mid y\,;\,\theta\right). \tag{2.13}$$

In speech enhancement, the result of deriving the optimal parameter θ_{optimal} determines the essential characteristics to enhance the speech signals. Different assumptions of $\mathcal{P}_{S|Y}\left(\theta\right)$ lead to different optimization criteria.

In Wiener filtering [8], the estimation target is the time-domain speech signal $s(t)$, the observed noisy signal for conditioning is $y(t)$, and the parameter is a time-domain linear filter $h(t)$; in other words, the assumption is that

filtering noisy signal $y(t)$ by $h(t)$ can obtain the clean speech signal, i.e.,

$$\begin{aligned}\hat{s}(t) &= y(t) * h(t) \\ &= \int_{-\infty}^{\infty} h(\eta) y(t-\eta) d\eta \,. \end{aligned} \tag{2.14}$$

Finally, in Wiener filtering, the conditional probability distribution is Gaussian distribution with the mean of $\hat{s}(t)$ and unit variance, i.e.,

$$s(t) \mid y(t) \sim \mathcal{N}\left(\hat{s}(t), 1\right) \,. \tag{2.15}$$

Using MLE, the optimal $h_{\text{optimal}}(t)$ follows the Wiener-Hopf equation, of which spectrum (also called spectral gain function) has the following formula

$$H_{\text{optimal}}(\omega) = \frac{\mathbb{E}\left[|S(\omega)|^2\right]}{\mathbb{E}\left[|S(\omega)|^2\right] + \mathbb{E}\left[|N(\omega)|^2\right]} \,, \tag{2.16}$$

when speech and noise signals are stochastically independent. This result introduces a different operation on the noisy spectrum, which is

$$\hat{S}(\omega) = Y(\omega) H_{\text{optimal}}(\omega) \,, \tag{2.17}$$

instead of literal subtraction using the spectral subtraction method. The Wiener filter has a vast range of applications in speech enhancement and source separation, which is one of the most common computational targets even in the modern deep learning techniques [26].

In MMSE-STSA [10], Ephraim and Malah believe that the magnitude spectrum contains more significant perception information than the time-domain signal. Therefore, MMSE-STSA keeps the noisy phase, uses the amplitude spectrum $|S(\omega)|$ as the estimation target, and utilizes the Rayleigh distribution for the conditional distribution. Similar to Wiener filtering, the result of this derivation is also a spectral gain function to apply on $|Y(\omega)|$.

Under the same principle, there have been several statistical techniques to produce the spectral gain function to modify the magnitude of the speech signal and improve the performance [11–14], which substantially reduce the residual noise and distortion. In summary, statistical techniques investigate the spectral gain function under different optimization conditions; however,

implementing such a function from the noisy observation is still ad-hoc.

Subspace techniques

Subspace techniques change the viewpoint of speech enhancement. Instead of trying to design some optimal enhancement characteristics, i.e., spectral gain function, like in statistical techniques, the subspace techniques provide the solution for decomposing the noisy signal into speech and noise signals. The noise is additive, hence linear. Therefore, the speech-noise independent assumption changes into a linear algebra assumption: an Euclidean space exists such that speech and noise stay separately in two mutually orthogonal subsets. Then, the speech can be obtained by simply zeroing out all components in the noise subspace. Consequently, with this approach, the dimension of the speech subspace should be smaller than the original dimension of the input, and therefore, dimensional reduction techniques are essential for subspace techniques.

Subspace techniques require representing the signal in some vector form and performing a dimensional reduction to obtain the approximated speech vector/matrix. For instance, Dendrinos *et al.* [15] and Jensen *et al.* [16] utilizes the windowed representation of the speech signal $s(t)$ as

$$s_{\text{window}}(\tau, t) = s(t + \tau)w(t)\,, \tag{2.18}$$

where $w(t)$ is some window function, such as a rectangular window. Then, the low-rank approximation of $s_{\text{window}}(\tau, t)$ is used for dimensional reduction and can be defined as

$$\hat{s}_{\text{window}}(\tau, t) = \sum_{k=1}^{p} u_{\text{S},\,k}(\tau)\sigma_{\text{S},\,k}v_{\text{S},\,k}(t)\,, \tag{2.19}$$

where p is the number of singular values, $\sigma_{\text{S},\,k}$ are the singular values, and $u_{\text{S},\,k}(\tau)$ and $v_{\text{S},\,k}(t)$ are the left and right singular vectors, respectively. This low-rank approximation projects the speech matrix into a subspace with a lower rank p. Then, the speech enhancement process is merely projecting the noisy matrix $y_{\text{window}}(\tau, t)$ into this speech subspace. When the noise is white noise, singular-value decomposition (SVD) can straightforwardly compute $u_{\text{S},\,k}(\tau)$, $\sigma_{\text{S},\,k}$, and $v_{\text{S},\,k}(t)$ from $y_{\text{window}}(\tau, t)$. For colored noise, some modifi-

cations are necessary to adjust the method [17, 39–41]. Besides SVD, eigenvalue decomposition (EVD) is also a candidate dimensional reduction, which utilizes the additivity in covariance matrix under the assumption of speech-noise independent, or more correctly, uncorrelation [18, 42, 43]. Non-negative matrix decomposition is another technique of which the speech vector is power spectrum [19] thanks to the additivity shown in Eq. (2.9) restricted by non-negative constraint.

Subspace techniques provide a new perspective on the speech enhancement problem, providing techniques for constructing the subspace of speech that preserves its most essential characteristics.

Summary

In summary, three classes of traditional speech enhancement techniques are spectral subtraction, statistical, and subspace techniques, which provide different points of view to the speech enhancement problem. According to Loizou [38], among the techniques, statistical techniques, especially Wiener filtering, consistently perform well, including speech quality and intelligibility, in many different noise conditions. Subspace techniques may give a low overall quality, yet they outperform in terms of intelligibility, revealing the physical meaning behind the obtained subspace. However, traditional techniques backlog some issues, one of which is dealing with various types of noise and various speaker characteristics in a single model. The deep learning techniques aim to resolve this problem in the next section.

2.1.2 Feature enhancement techniques: speech knowledge

Traditional techniques focus on the mathematical framework for speech enhancement or, more accurately, extracting the speech signal from the noisy signal. Their principles are the same for all kinds of signal enhancement, not only for speech. In contrast to traditional techniques, feature enhancement techniques focus on the target of enhancement: the speech signal. Specifically, these techniques try to investigate the meaningful features of speech strongly related to speech quality and intelligibility and how to modify them for enhancement.

Along spectral axis

The most common feature domain for speech enhancement is the complex time-frequency domain obtained from the Short-Time Fourier Transform (STFT) of a signal, of which the equation is as follows

$$\begin{aligned}\tilde{x}\left(\omega,\tau\right) &= \mathcal{F}\left\{x_{\text{window}}(t,\tau)\right\} \\ &= \int_{-\infty}^{\infty} x(t+\tau)w(t)e^{-i\omega t}dt\,, \end{aligned} \tag{2.20}$$

where $x(t)$ is the signal in the time domain and $w(t)$ is some window function [44]. A window function of length T_{w} is a symmetric, non-negative function that is zero-valued outside the interval $\left[-\frac{T_{\text{w}}}{2},\frac{T_{\text{w}}}{2}\right]$. STFT allows capturing the spectrum of several equal-length segments of the time-domain signal $x(t)$ at different positions τ. Most traditional techniques processing [3–14, 19] in the spectral domain use this technique in practical implementation.

In narrow-band STFT where the window length T_{w} is relatively long, around 20 ms to 40 ms. With this setting, when an audio segment contains a voiced sound, i.e., the sound caused by the vibration of the vocal cords, its amplitude spectral information $|\tilde{x}(\omega,\tau)|$ contains a spectral envelope - the broad shape of the spectrum, and spectral fine structure - the detailed, rapid variations in the spectrum [29–33]. The spectral envelope is closely related to the perceived characteristics of vowels and consonants. In contrast, the spectral fine structure contains information about subtle nuances such as pitch variations, timbral characteristics, and transient events like consonant sounds or percussive elements. Hu and Loizou [45] apply this knowledge to advance the Wiener filtering to enhance the spectral envelope and perceptually improve speech intelligibility. On the other hand, Malah and Cox [46] introduce comb filtering to enhance the spectral fine structure. Not only stop at the amplitude, Wakabayashi *et al.* [47, 48] considers the enhancement of phase spectrum to improve the harmonic characteristics in spectral fine structure.

Along temporal axis

The use of STFT described above is 'spectral analysis of a segment', which focuses on the features along the spectral domain. However, several previ-

ous studies [34–36] show that the temporal domain also contains important perceptual features for speech recognition. The mechanism of the human auditory system forms the basis for these characteristics. The cochlea performs frequency decomposition in the inner ear by converting sound vibrations into neural signals.The cochlea achieves this through a process known as tonotopic organization, where different frequencies of sound stimulate different regions along the length of the cochlea [49–51]. A model to approximate the frequency decomposition in the cochlea is the auditory filterbank, which decomposes the audio stimuli into several sub-band signals. Together with the Hilbert transform, a filterbank can return a complex time-frequency representation as follows:

$$\begin{aligned}\tilde{x}(\omega,\tau) &= (x * \psi_\omega)(\tau) + i\,\mathcal{H}\{(x * \psi_\omega)(\tau)\} \\ &= (x * (\psi_\omega + i\,\mathcal{H}\{\psi_\omega\}))(\tau) \\ &= \left(x * \tilde{\psi}_\omega\right)(\tau), \end{aligned} \tag{2.21}$$

Several auditory filter models, such as Gammatone [52] and Gammachirp [53] filter, have been proposed to mimic the shape and bandwidth of the cochlea filter. In addition, STFT is also a filterbank that maintains a constant bandwidth (and thus, not an auditory filterbank) where the filter $\tilde{\psi}_\omega(t) = w(t)e^{i\omega t}$ is a band-pass filter of which center frequency is ω (rad/s).

Along the temporal axis, the amplitude envelopes of the sub-band signals [34–36, 54, 55] play an essential role in speech recognition. Many researchers analyzed the frequency components of the temporal amplitude envelopes, also known as the modulation spectrum, and found that different modulation frequency bands corresponded to different information of speech, for example, 0.2 Hz to 0.5 Hz for sentence units, 1 Hz to 2 Hz for stressed syllables, 2 Hz to 3 Hz for words, 3 Hz to 6 Hz for syllables, and 10 Hz to 20 Hz for phonemes [35]. Moreover, the modulation spectrum is mainly independent of the center frequency of the sub-band. From these findings, several studies attempt to enhance the modulation spectrum characteristics of speech using techniques such as temporal modulation transfer function [34, 56–59]. For additive noise, some studies develop speech enhancement methods for filtering on the amplitude envelopes [60–62].

Summary

In summary, several studies reveal the nature of speech to clarify the enhancement target. These studies develop beyond speech enhancement in additive noise, as the knowledge they bring can help enhance speech signals in several different scenarios, for example, enhancing the speech signal so that it still has high quality and intelligibility when emitted to an adverse environment such as the train station.

2.1.3 Deep learning techniques: computational actualization

Deep learning techniques model a data distribution from several data samples using artificial neural networks, which are non-linear functions with millions of learnable parameters to estimate some arbitrary function. Some well-known architectures of artificial neural networks are convolution neural networks, recurrent neural networks, and transformers. The parameters of these artificial neural networks are estimated by minimizing some loss function between the ground-truth data and the networks' output via several iterations of adjustment based on the gradient of the loss function, which is called model training. Thus, deep learning techniques require training the model on some known data before applying it in the estimation tasks.

In speech enhancement, deep learning techniques play a robust role as estimators. They can be applied to estimate the speech signal, some enhancement characteristics such as the spectral gain function from statistical techniques, or the lower-dimension subspace using the principles from subspace techniques, given the clean-noisy pair stimuli data samples. Therefore, deep learning techniques result from actualizing the traditional techniques by non-linear estimation from data to handle the complex variation in speaker, pronunciation, and noise. With the current development of deep learning techniques, applying deep learning techniques to traditional techniques greatly improves enhancement performance.

There are two main processing domains with speech enhancement methods: time and time-frequency. Most speech enhancement methods in the time domain are based on powerful generative models, such as generative adversarial networks or diffusion probabilistic models, which attempt to generate a

clean waveform directly, given the noisy speech waveform or features [20–23]. Methods in the time-frequency domain operate on the time-frequency representation of the signal obtained by short-time Fourier transform (STFT) or wavelet transform. A time-frequency representation is complex-valued, which contains amplitude and phase features. Most initial methods can only enhance the amplitude features while leaving the phase unprocessed [24] because the phase features have a complicated pattern and are challenging to model. The estimation targets are either the amplitude features or masks to modify the spectra, such as the ideal binary mask [25] or ideal ratio mask [26]. By incorporating deep neural networks, several speech enhancement methods with this approach perform excellently [27, 28, 63].

Speech signals follow a speech-production process and include linguistic, paralinguistic, and nonlinguistic information [64], while noise signals can be arbitrary. Thus, several studies have utilized the distribution of speech [65–67]. These studies modeled the speech and noise variances separately and then constructed the ideal ratio mask based on Wiener filtering to enhance speech, in which a variational autoencoder (VAE) [68] modeled the distribution of speech and the non-negative matrix factorization [19] modeled the distribution of noise. With this approach, the generative speech model must satisfy noise-robustness and high-fidelity requirements to obtain high-quality enhanced speech.

In the time-frequency processing domain, besides the amplitude features, the importance of the phase features, with which speech quality improves significantly when the phase features are accurately estimated, has been clarified [69]. Therefore, several studies developed phase-aware enhancement methods [70–76], of which the most successful utilizes the concept of the complex ideal ratio mask (cIRM) [77]. With the development of complex-valued deep neural networks [78], estimating the cIRM from the complex-valued noisy time-frequency representation [73–76] becomes possible. However, the unbounded property of the cIRM makes it difficult for optimization due to the infinite search space [79]. The imaginary part of the cIRM also lacks learnable patterns for the neural network to explore [80].

2.2 Speech enhancement evaluations

Section 2.1 introduces several speech enhancement techniques, from traditional to modern ones. Before applying these techniques in practice, it is necessary to evaluate their performance. The target of speech enhancement is to improve speech quality and intelligibility, which are qualitative measures such as 'high quality' or 'low intelligibility.' Therefore, the performance evaluation methods of speech enhancement involve quantifying these qualitative measures.

Speech enhancement has two primary evaluation targets: quality and intelligibility. Both are distinct attributes of speech signals and should not be confused. Quality measures evaluate a speaker's utterance production, including factors like naturalness or hoarseness. In contrast, intelligibility measures assess how understandable the speech content is. For instance, speech signals synthesized from a few modulated noise bands using noise vocoding can have high intelligibility but low quality, making them sound mechanical [36].

Regarding evaluation conduction methods, there are two types: subjective and objective. Subjective evaluation methods refer to evaluations based on human perception and judgment, such as rating overall quality, naturalness, or background noise reduction by a group of listeners. Objective evaluation assesses the quality of speech enhancement systems based on measurable and quantifiable metrics designed to correlate well with subjective listening tests.

2.2.1 Subjective evaluations

Subjective evaluation methods provide valuable insights into how effective a speech enhancement system is from the perspective of the end user, which can sometimes differ from objective metrics. They are crucial parts of evaluating the performance of speech enhancement systems.

- ITU-T P.800 [81]: This standard describes methods and procedures for subjective evaluations of telephone connections' perceived transmission quality. It is typically applied in the telecommunications industry to evaluate analog and digital systems. ITU-T P.800 follows an Absolute Category Rating (ACR) methodology. ACR provides a simple

test paradigm for assessing stimulus on a single quality scale - or Mean Opinion Score (MOS), where each value from 1 to 5 is associated with a specific categorical description. However, typically, the scale is interpreted as in Table 2.1

Table 2.1: Score description in ITU-T P.800.

Score	Description
5	excellent (perfect quality, effortless conversation)
4	good (high quality, comfortable conversation)
3	fair (acceptable quality, possible conversation without much effort)
2	poor (low quality, possible conversation with effort)
1	bad (unacceptable quality, impossible conversation)

- ITU-T P. 807 [82]: This standard describes a subjective testing methodology for assessing speech intelligibility in communications settings, systems, and devices. The method provides a percent correct intelligibility score based on a two-alternative, forced-choice task where the stimulus is one of the two words from a pair of words, i.e., a test item. Half of the test items are rhyming word pairs (i.e., they differ only in the initial consonant), and half are alliterative word pairs (i.e., they differ only in the final consonant). The two critical consonants in each test item vary only in a distinctive feature. In addition to a score for overall intelligibility, the method provides scores for each of six characteristic features: voicing, nasality, sustention, sibilation, graveness, and compactness.

- ITU-T P. 835 [83]: This standard describes a subjective evaluation framework for evaluating speech communication systems incorporating speech enhancement algorithms. It is particularly appropriate for evaluating speech enhancement methods. The methodology uses separate MOS scales to estimate three dimensions of speech quality separately: signal distortion (SIG), noise distortion (BAK), and overall quality (OVRL).

  In ITU-T P. 835, each trial fed to the listeners contains three subsamples. Each subsample has the following sequential structure: a period of background noise alone, a period of speech with background noise,

Table 2.2: Score description (SIG and BAK) in ITU-T P.835.

Score	SIG description	BAK description
5	not distorted	not noticable
4	slightly distorted	slightly noticable
3	somewhat distorted	noticable but not intrusive
2	fairly distorted	somewhat intrusive
1	very distorted	very intrusive

a period of noise only, and an appropriate silent voting interval. For the first two subsamples, listeners rate either the SIG or the BAK, depending on the rating scale order specified for that trial. For the SIG, listeners are instructed to attend only to the speech signal and rate the speech on the MOS distortion scale, while for the BAK, subjects are instructed to attend only to the background and rate the background on the MOS intrusiveness scale. The score description of both scales are in Table 2.2. For the third subsample, subjects are instructed to listen to the speech signal, including background, and rate it the same way as ITU-T P.800 [81].

2.2.2 Objective evaluations

Objective evaluation methods for speech enhancement primarily involve quantifying the improvement in speech quality and intelligibility using measurable and quantifiable metrics. Most objective evaluation methods are intrusive, requiring measuring the distance or similarity from the target stimuli to a reference stimuli of perfect quality. Some common perceptual evaluations of speech quality and intelligibility are:

- Signal-to-noise ratio (SNR): SNR measures the ratio of the speech signal power to the noise power. Higher values of SNR indicate less noise distortion. In practical use, the SNR is in 10based logarithm scale using the following equation

$$\mathrm{SNR}\,(\mathrm{dB}) = 10\log_{10}\frac{\mathbb{E}\left[s^2(t)\right]}{\mathbb{E}\left[n^2(t)\right]}\,. \tag{2.22}$$

- Mel cepstral distortion (MCD) [84]: MCD quantifies the difference between two speech signals based on the Mel scale, a perceptual scale of pitches approximating the human ear's response more closely than the linearly-spaced frequency bands. In speech synthesis or recognition, MCD provides a way to compare a synthesized speech signal with a reference signal and determine how closely they match. Lower MCD values indicate a closer match and, thus, better performance of the speech synthesis or recognition system.

- Perceptual Evaluation of Speech Quality (PESQ) [85]: PESQ is a widely used objective metric that measures the similarity between the enhanced speech signal and the reference signal based on the perceptual quality of the speech signal. The scores given by PESQ are on the MOS scale, i.e., from 1 (bad) to 5 (excellent). PESQ is one of the standards recognized by ITU [86].

- Perceptual Objective Listening Quality Analysis (POLQA) [87]: POLQA is another speech perceptual evaluation varying from 1 to 5, similar to PESQ. POLQA advances PESQ and provides more accurate and reliable results, especially for modern wide-band speech codecs. POLQA is also standardized by ITU [88].

- Short-Time Objective Intelligibility (STOI) [89]: STOI computes the correlation between the clean and degraded speech signals based on short-time spectral features. The scores given by STOI are between 0 - low intelligibility and 1 - high intelligibility.

- Blind Source Separation Evaluation (BSS Eval) [90]: BSS Eval is a set of metrics used to evaluate the performance of blind source separation algorithms. With additive noise, speech enhancement can be considered blind source separation. In blind source separation, the estimated target (speech) signal is decomposed as a sum of four signals

$$\hat{s}(t) = s_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t)\,, \tag{2.23}$$

  where $s_{\text{target}}(t)$, $e_{\text{interf}}(t)$, and $e_{\text{noise}}(t)$ are allowed deformations of the clean speech signal $s(t)$, point-source noise signal, and isotropic noise

signal, respectively; while $e_{\text{artif}}(t)$ is the artifact yielded by the separation algorithm [91]. The output values of the BSS Eval include Signal-to-Distortion Ratio (SDR), Source-to-Inferences Ratio (SIR), Source-to-Noise Ratio (SNR), and Source-to-Artifacts Ratio (SAR).

- Perceptual Evaluation methods for Audio Source Separation (PEASS) [92]: PEASS is another set of metrics designed to predict the perceived quality of estimated source signals in the context of audio source separation1. PEASS aims to provide a more perceptually relevant evaluation of the performance of audio source separation algorithms with four metrics: Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), Artifact-related Perceptual Score (APS), and Overall Perceptual Score (OPS).

- Virtual Speech Quality Objective Listener (ViSQOL) [93]: ViSQOL is an objective, full-reference metric for perceived audio quality using the spectro-temporal measure of similarity between a reference and a test speech signal to produce a Mean Opinion Score - Listening Quality Objective (MOS-LQO) score from 1 (the worst quality) to 5 (the best quality). ViSQOL has been designed to be robust for quality issues associated with Voice over IP (VoIP) transmission. ViSQOL competes well with other standard metrics like PESQ and POLQA, offering an alternative to POLQA in predicting speech quality in VoIP scenarios.

Overall, objective evaluation metrics provide a reliable and quantitative means of evaluating the performance of speech enhancement algorithms. However, it is essential to note that these metrics may only sometimes correlate well with subjective listening tests, which are the ultimate measure of speech enhancement performance.

Chapter 3

Modulation theory of speech communication

The waveform of speech signals has a complex structure that fluctuates over time, which seems completely random. However, the language spoken only contains a finite set of phonemes, which restrains the linguistic information to be finite. Also, listeners can understand the content even when two different people pronounce the same sentence. Therefore, the speech signals must contain some structure essential for perceiving and recognizing speech, while the other information is just variation. The organs for speech production construct sound waves from those essential features, and the auditory system receives and decodes those features for speech perception. Then, how should the speech signals be dissected?

The structure described above is similar to a communication system that uses the *modulation* technique, such as a radio communication system. In communication theory, modulation is a technique for transmitting a low-frequency message signal over a longer distance by manipulating a high-frequency carrier signal so the carrier signal can 'carry the message' over a longer distance. The construction process of speech signals is similar to the modulation process, where the sound wave is the carrier that carries essential speech information to the listeners.

This chapter provides the background to support the idea that speech communication processes are modulation-based. The first section briefly introduces the modulation concept. The second section introduces the mod-

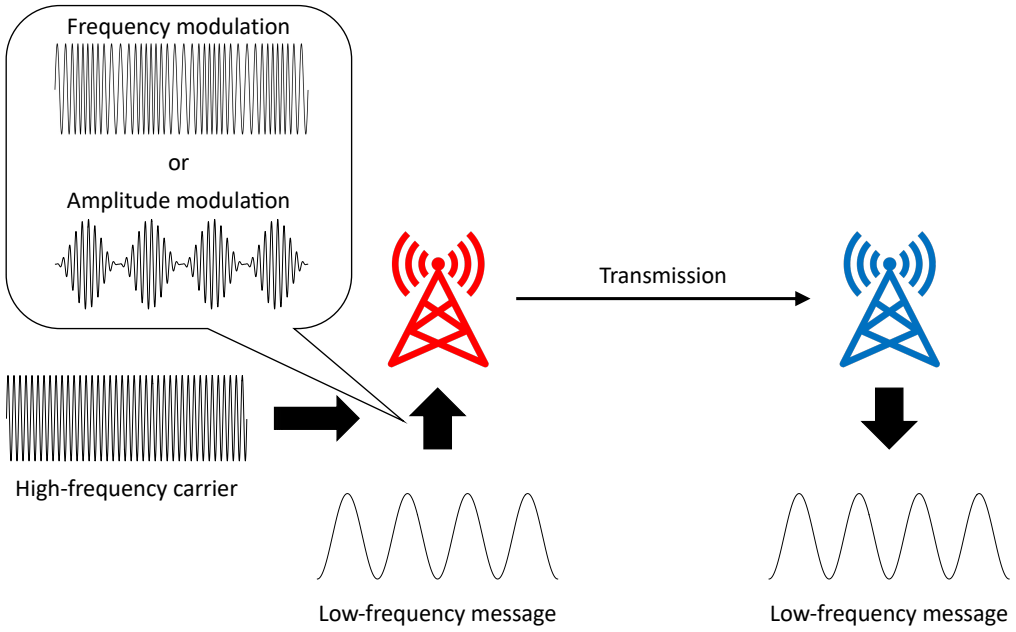

Figure 3.1: Diagram of modulation process in radio transmission.

ulation assumption of speech, which provides the view of speech signals as modulated signals along different axes and in different attributes.

3.1 Concept of modulation

Modulation is an essential process employed in communication theory. It involves manipulating a carrier signal's characteristics, such as amplitude, frequency, and phase, with a modulating wave. This technique converts data as a message signal into electrical or digital signals that transmit over a communication medium. The primary function of modulation is to enhance the strength of signals for maximum reach, which is critical in modern communication systems. Radio transmission to transmit sound signals is a typical example of modulation, of which the process diagram is in Figure 3.1. The sound wave can only travel in the air for a short distance, so the radio system applies modulation to transmit it at a longer distance, enabling communication across a village, a city, or even a country.

Modulation is a complex process involving mathematical operations to modulate the carrier signal. The modulating signal is superimposed onto

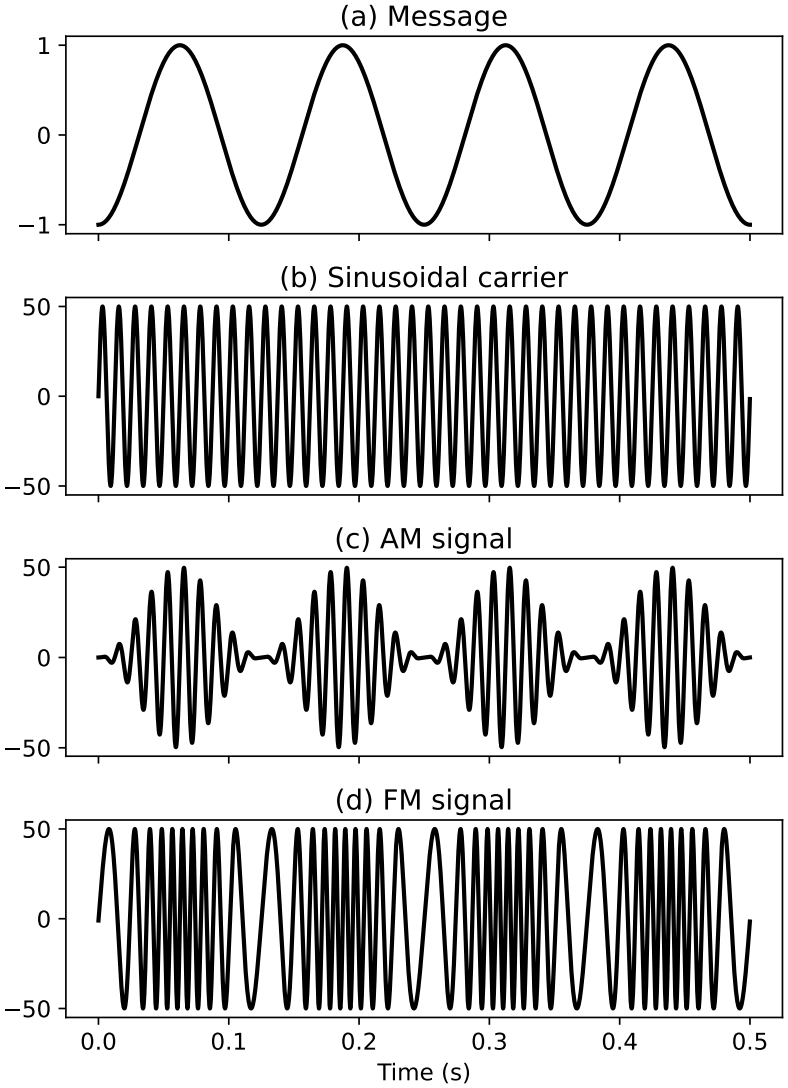

Figure 3.2: Example of AM and FM signals with a carrier frequency of $f_c = 80\,\text{Hz}$ carrying an $8\,\text{Hz}$ message $m(t)$: (a) message signal, (b) carrier signal, (c) AM signal, and (d) FM signal.

the carrier signal, resulting in the production of a modulated signal. The modulated signal transmits to the intended receiver over a communication channel, such as a cable, fiber-optic line, or wireless network.

Two main types of modulation techniques used in practice communication systems are amplitude modulation and frequency modulation:

- Amplitude Modulation (AM): Only the amplitude of the carrier signal is varied to represent the message to the signals. In contrast, the signal's phase and frequency are kept unchanged. The mathematical description of an AM signal is as follows

$$x_{\mathrm{AM}}(t) = A\left(\frac{m(t)+1}{2}\right)\cos\left(2\pi f_c t\right), \tag{3.1}$$

  where

  - A is the peak amplitude of the carrier signal,
  - $m(t)$ is the message signal of which range is between -1 and 1,
  - f_{c} is the frequency (in Hertz) of the carrier signal.

  The instantaneous amplitude of the sinusoidal carrier is modified by a multiplicative term $\frac{m(t)+1}{2}$, which is non-negative, to construct the AM signal $x_{\mathrm{AM}}(t)$. Figure 3.2 (c) illustrates an AM signal with a carrier frequency $f_{\mathrm{c}} = 80\,\mathrm{Hz}$ carrying a $8\,\mathrm{hz}$ sinusoidal message.

- Frequency Modulation (FM): Only the carrier signal frequency is varied to represent the message, while the signal's phase and amplitude are kept unchanged. The formula of an FM signal is as follows

$$x_{\mathrm{FM}}(t) = A\cos\left(2\pi f_{\mathrm{c}} t + 2\pi f_{\Delta}\int_0^t m(\tau)d\tau\right), \tag{3.2}$$

  where

  - A is the amplitude of the carrier signal,
  - $m(t)$ is the message signal of which range is between -1 and 1,
  - f_{c} is the frequency (in Hertz) of the carrier signal,

– f_Δ is the peak frequency deviation (in Hertz), controlling the maximum modification to the instantaneous frequency of the carrier signal away from its frequency f_c in the modulation process.

This definition modifies the instantaneous of the sinusoidal carrier by an additive term $f_\Delta m(t)$ away from the carrier frequency f_c. Figure 3.2 (d) illustrates an FM signal with a carrier frequency $f_c = 80\,\text{Hz}$ carrying a 8 Hz sinusoidal message.

After the transmission, the receiver must extract the message signal from the transmitted signal, called demodulation. In practice, the communication system is a broadcasting system; in other words, several senders are sending messages simultaneously. Each sender uses a unique carrier frequency, which helps distinguish the senders from one another. The receiver first 'listens' to a target sender by filtering at the specific carrier frequency, keeping the signals of which frequency stays within a band around the carrier frequency. Then, depending on the modulation type, i.e., AM or FM, the receiver extracts the message signal using AM or FM demodulation techniques.

Modulation is a crucial technique in communication theory that enables data transmission over a communication medium. However, within the context of this dissertation, modulation is considered a general principle and a mathematical framework to describe the human speech communication process. The following section will discuss this viewpoint.

3.2 Modulation assumption of speech

Modulation is the technique that helps transmit a low-frequency message signal using a high-frequency carrier signal. The analogy of the modulation-demodulation process can be applied to human speech communication to discuss the perceptual invariance against a large amount of acoustic variance [94]. As a consequence, the mathematical framework in modulation can help extract the characteristics of speech. The earlier section describes how human speech communication is similar to a modulation process and the related works that share a similar point of view. Then, the latter of this section provides evidence of modulation in amplitude and frequency by the extracted features using the modulation framework.

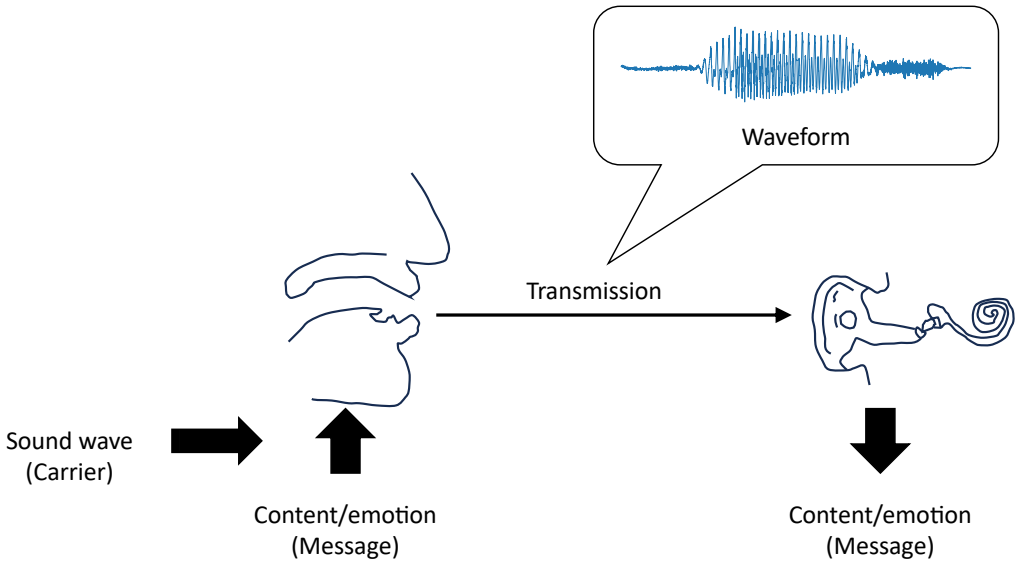

Figure 3.3: Speech communication process as a modulation-demodulation process.

3.2.1 Speech communication as a modulation process

Speech is a natural yet powerful way to convey human thoughts, ideas, and emotions to one another. It is the bridge that connects the realm of our minds to the external world. Through speech, we can share our unique perspectives, inspire others with our ideas, and express our emotions in a way that words on a page cannot capture. It is a fundamental aspect of human interaction, fostering understanding and empathy among individuals. Indeed, speech is not just a means of communication but a testament to the complexity and depth of human thought and emotion.

There are two aspects of speech communication: speech production and speech perception. The physiological mechanism of speech production is a complex process that involves several stages and parts of the human body [95], including:

1. Thought formation and word choice: The process of forming thoughts and choosing words begins in the mind.

2. Respiration: The air from the lungs is directed to the larynx.

3. Phonation: The larynx has vocal folds that remain open or vibrate

to produce sound. When the vocal folds vibrate, they convert the air breathed out into intermittent airflow pulses. This process results in a buzzing sound.

4. Articulation: The air from our lungs is shaped by different parts of our mouth and nose, called articulators. These include the tongue, lower jaw, lips, and velum, and they work together to create specific movements that change the resonance properties of the airway above our vocal cords.

5. Resonance: Resonators in the upper respiratory tract include pharyngeal, oral, and nasal cavities. These chambers transform laryngeal sounds into sounds with special linguistic functions.

This process is similar to the modulation process, where the first stage acts like a message synthesizer, while other stages construct a sound wave carrying the synthesized messages. On the other hand, in speech perception, the auditory system first transfers the sound wave as air vibration into the cochlea and vibrates the basilar membrane, causing the organ of Corti to move against the tectorial membrane, stimulating the generation of nerve impulses to the brain [96]. Within the cochlea, each membrane region is most affected by a specific frequency of vibrations, which helps the basilar membrane analyze the different frequencies of complex sounds [96]. This auditory system function is similar to demodulation, where there are several receivers, each of which keeps listening to a specific frequency.

The physiological process of speech production and speech perception behaves similarly to a modulation-demodulation system, which makes speech signals the result of modulating sound waves with speech information to travel in the air. Figure 3.3 visualizes such an analogy. Under this analogy, the later parts of this section review the messages hidden inside speech signals.

3.2.2 Amplitude modulation: temporal and spectral modulation

The speech signal is amplitude-modulated in both temporal and spectral axes (see Figure 3.4).

Temporal modulation

As aforementioned in Chapter 2.1.2, when using a filterbank to decompose a speech signal, each sub-band signal is amplitude-modulated [34–36]. In the temporal domain, AM manifests as fluctuations in the amplitude of the speech waveform over short time intervals, capturing dynamic changes in loudness and intensity. These temporal variations provide essential cues for segmenting speech into phonetic units, delineating syllables, and conveying prosodic features such as stress, emphasis, and intonation. Comprehending spoken language seems to hinge on a diverse range of syllable durations, spanning from 50 to 400 ms for American English, as evidenced by the modulation spectrum of the acoustic signal. The upper segment of the modulation spectrum, falling within the range of 6–20 Hz, primarily corresponds to unstressed syllables. In contrast, below 5 Hz, the lower segment predominantly denotes heavily stressed syllables [97].

Spectral modulation

AM also appears in the spectral domain. Specifically, the power spectrum of a voiced segment is amplitude-modulated. The source-filter model of speech production can explain this phenomenon. In this model, the vocal tract acts as a filter, shaping the excitation source into specific speech sounds by modifying spectral characteristics. The vocal tract's transfer function represents how the vocal tract shapes the excitation source to produce speech sounds. The excitation source, often called the glottal flow, is generated by the vibration of the vocal folds in the larynx. This vibration creates a periodic signal with a fundamental frequency corresponding to the pitch of the voice. The glottal flow serves as the input to the vocal tract filter, and its spectral characteristics determine the fundamental pitch and timbre of the speech sounds. The spectral envelope reflects the frequency response of the vocal tract filter, while the spectral fine structure captures the harmonic structure in the periodic glottal flow.

3.2.3 Frequency modulation: instantaneous frequency deviation (IFD)

While several studies investigate the amplitude modulation characteristics of speech, frequency modulation receives less attention. Maragos *et al.* introduces the concept of FM to model the non-linearities of speech resonances [98]. Several studies also show that FM appears in speech signal [99, 100], characterizing formant transitions [101] and strongly contributing to speech perception [102].

In speech processing, the message demodulated from the FM signal is called instantaneous frequency deviation (IFD) [103], which refers to the amount of deviation of the instantaneous frequency around the center frequency of each sub-band signal. Given a complex time-frequency representation $\tilde{x}(\omega, \tau)$ of a signal $x(t)$, from the construction of the FM signal in Eq. (3.2), the following equation can conventionally compute the IFD:

$$Q(\omega, \tau) = \frac{\partial \angle \tilde{x}(\omega, \tau)}{\partial \tau} - \omega \tag{3.3}$$

As shown in Figure 3.5, when analyzing a speech signal, while the phase of the complex time-frequency representation has a complex structure without specific patterns, the instantaneous frequency of each sub-band signal deviates around the center frequency with a similar pattern to the amplitude. In detail, when the instantaneous amplitude is locally high, the IFD tends to be zero. Although one can intuitively explain this relation between the temporal amplitude and frequency modulation, the detailed quantitative relation is still unclarified [103].

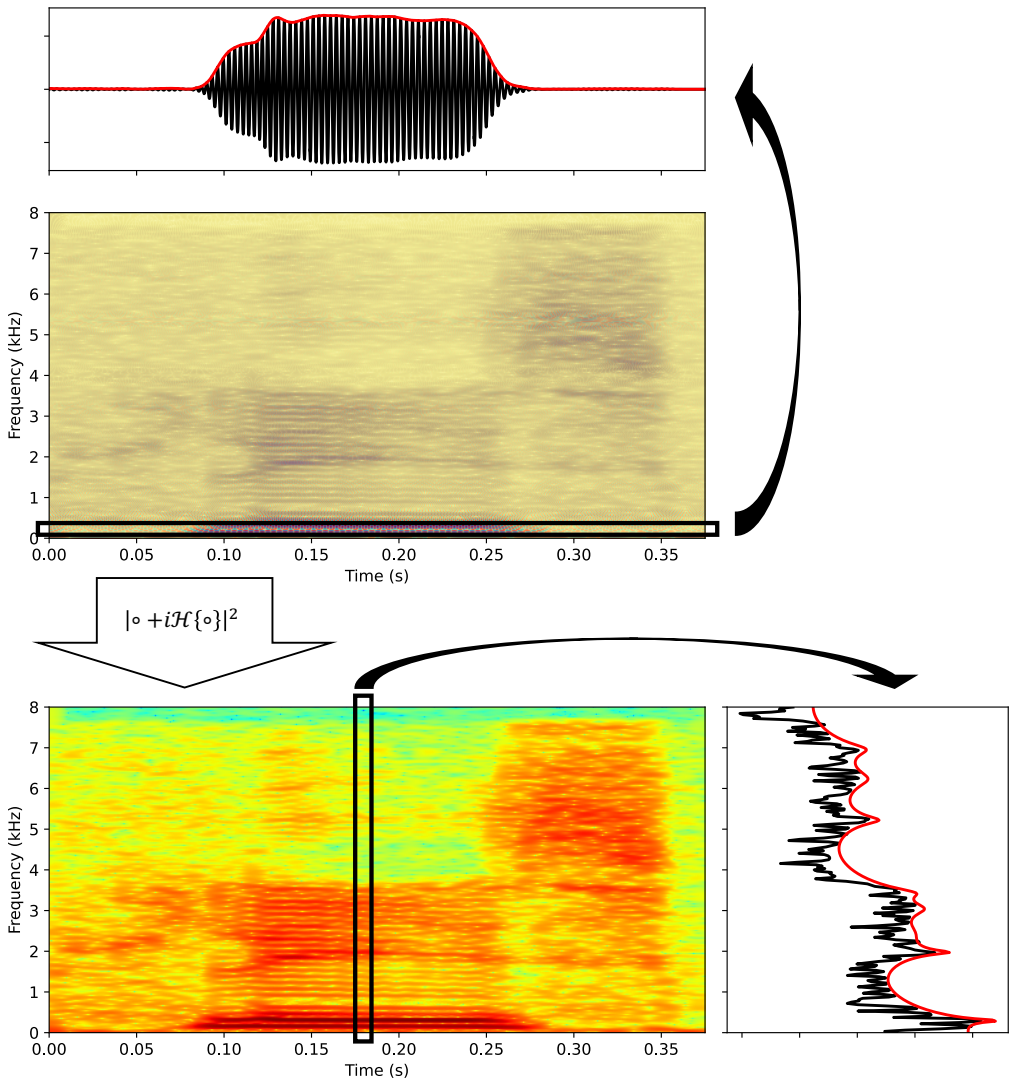

Figure 3.4: Amplitude modulation of speech in time and frequency axes.

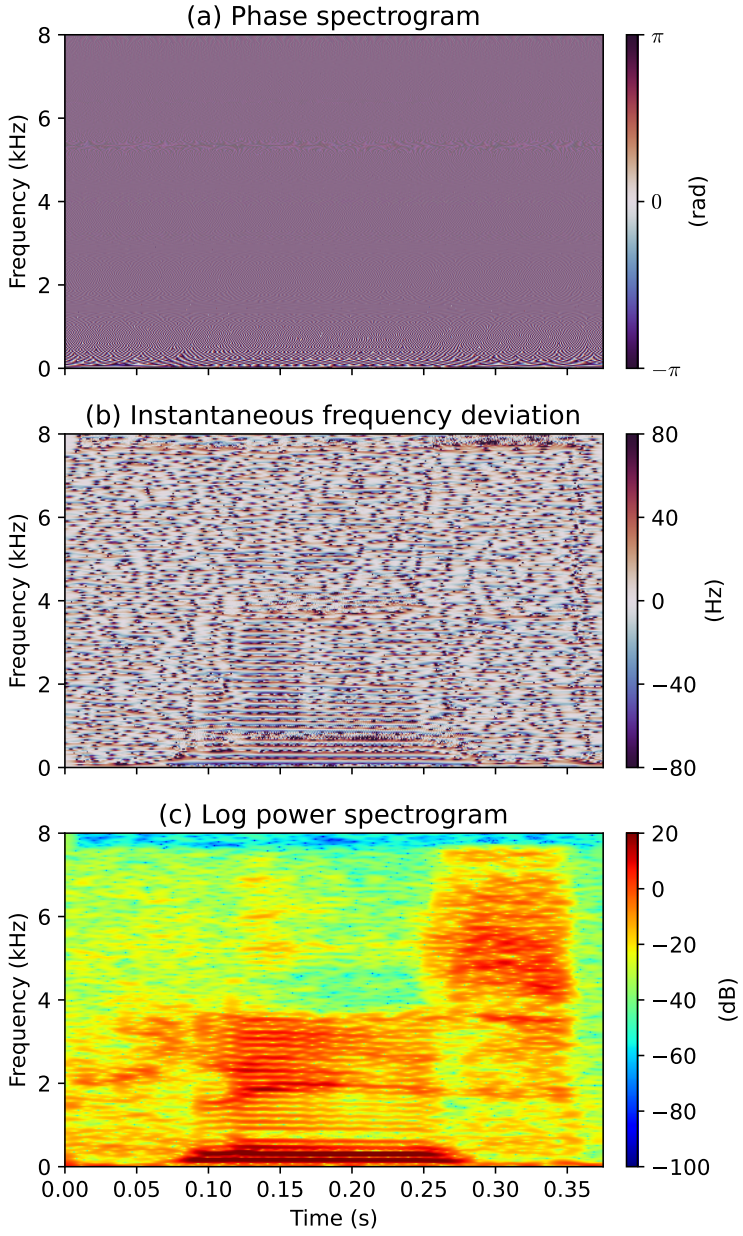

Figure 3.5: Time-frequency representation of speech: (a) phase spectrogram, (b) IFD spectrogram, and (c) log power spectrogram (amplitude). While phase spectrogram has a complex structure, IFD spectrogram has similar patterns to log power spectrogram.

Chapter 4

Spectral modulation characteristics enhancement in amplitude

In an attempt to answer the first research question, 'Does enhancing the speech characteristics in amplitude improve speech enhancement performance,' based on modulation theory, this chapter proposes a method to model the spectral modulation characteristics of speech and applies it to speech enhancement. The results show that (1) the proposed method effectively models the spectral modulation characteristics of speech and (2) applying the proposed method to speech enhancement improves the quality and intelligibility of enhanced speech, answering the first research question.

4.1 Problem formulation

As Chapter 3 mentions, the power spectra of voiced segments are amplitude-modulated along the spectral axis under the source-filter assumption. In the source-filter assumption, a voiced segment is the result of filtering the excitation source from the glottal by a linear acoustic filter - the vocal tract (see Figure 3.4). The glottal source, resulting from the vibration of the vocal cords, is not a pure tone but contains a fundamental tone with a fundamental frequency (F0) of f_0 Hz and a series of higher frequencies called upper harmonics, usually corresponding to a simple mathematical ratio of harmonics,

i.e., $2f_0, 3f_0, 4f_0, 5f_0, \ldots$. The vocal tract amplifies or attenuates this glottal source to determine the spoken phoneme and produces the sound wave of a voiced segment carrying that phoneme as a message. As convolution in time is multiplication in frequency, the spectrum of a voiced segment contains two multiplicative components: the spectral envelope, a smooth curve representing the vocal tract filter, and the spectral fine structure, a fluctuated structure with peaks appearing at a period of f_0 (in Hz).

Although it is easy to estimate the spectral envelope using subspace modeling such as linear prediction or even high-complexity methods such as deep learning, modeling spectral fine structure is a challenging problem for parametric estimation of speech spectra. The frequency bins are large in the narrow-band configurations where the harmonic appears. The spectrum estimation model estimates the spectral envelope and treats the spectral fine structure as unwanted noise when f_0 is low [104]. Unlike modulation in the time domain or defined in Eq. 3.1, the periodic peaks in the spectral fine structure are not sinusoidal; therefore, it is difficult to completely separate the spectral envelope and fine structure using Fourier analysis.

This chapter develops a method to solve the problem of spectral-fine-structure modeling. Section 4.2 introduces a novel loss function to encourage the model to learn the spectral fine structure using the discrete F0 distribution of F0 candidates. Applying this loss function, Section 4.3 proposes a spectral-fine-structure-aware speech enhancement method. Section 4.4 provides the evaluation results to verify whether modeling spectral fine structure is helpful for speech enhancement. Finally, section 4.5 concludes this chapter.

4.2 Modeling spectral fine-structure via discrete F0 distribution

4.2.1 Discrete F0 distribution

Given a log power spectrum $\boldsymbol{\rho} \triangleq \{\rho(f)\}_f$, the significance [104] of an F0 candidate ξ can be defined as

$$q_\xi = \sum_{k \geq 1} \frac{\rho\left(k\xi\right) - \rho\left(k\xi - \xi/2\right)}{\sqrt{k}}\,. \tag{4.1}$$

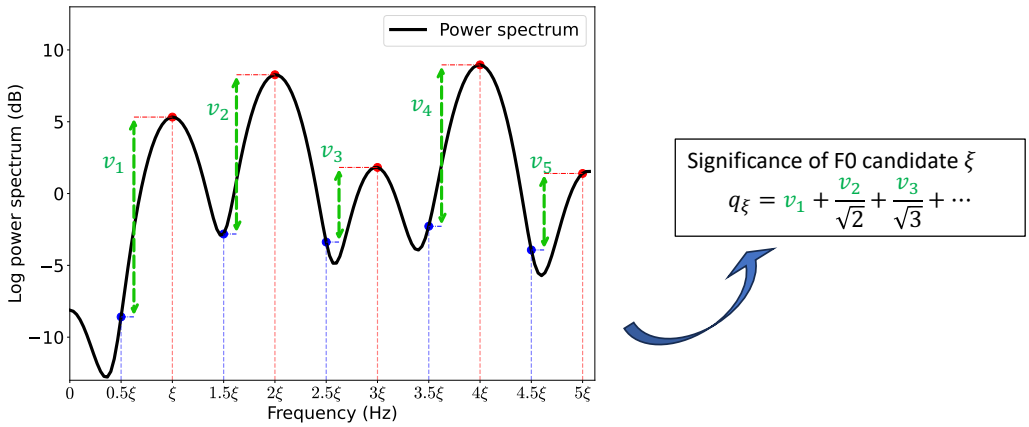

Figure 4.1: Subharmonic-to-harmonic algorithm to compute significance of an F0 candidate.

which refers to the distance between the peaks and the valleys right before it. Figure 4.1 visualizes the idea of the significance. The higher the significance q_ξ, the higher the chance that ξ is the F0 of $\rho(f)$. Therefore, given a set of C candidates $\{\xi_1, \ldots, \xi_C\}$ for F0, the F0 can be estimated as the candidate with the highest significance. In other words, the F0 distribution is the Dirac delta distribution at the candidate ξ_{k^*} where $k^* = \arg\max_k q_{\xi_k}$. As the $\arg\max$ is not differentiable, this section proposes an F0 distribution approximation using the Softmax function with temperature as follows

$$p_{\xi_k} = \frac{\exp\left(q_{\xi_k}/\iota\right)}{\sum_{k'=1}^{C} \exp\left(q_{\xi_{k'}}/\iota\right)}, \tag{4.2}$$

where $\iota > 0$ is the temperature parameter. The lower the temperature, the closer the approximated distribution to the Dirac delta distribution, while the higher the temperature, the more uniform the approximated distribution [105].

Using the approximated categorical distribution of F0 as above, the difference in F0 between two spectra can be quantified via the Kullback-Leibler (KL) divergence between their F0 distributions.

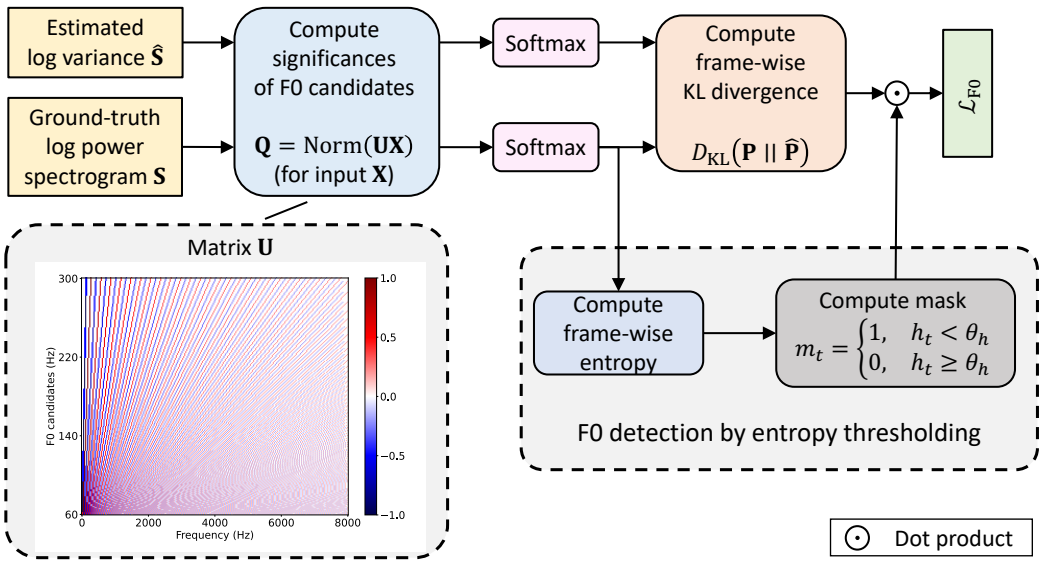

Figure 4.2: Block diagram of $\mathcal{L}_{\mathrm{F0}}$'s computation process.

4.2.2 Voiced and unvoiced cases via entropy thresholding

However, unvoiced or silent segments do not contain F0. Since there is no best F0 candidate, the approximated F0 distribution of an unvoiced segment is likely uniform, which means its entropy is high. Thus, this section proposes a loss function that quantifies the difference between a ground-truth spectrum $\boldsymbol{\rho}$ and an estimated spectrum $\hat{\boldsymbol{\rho}}$ based on their F0 distributions $\mathbf{p} = \begin{bmatrix} p_{\xi_1} & \cdots & p_{\xi_C} \end{bmatrix}^\top$ and $\hat{\mathbf{p}} = \begin{bmatrix} \hat{p}_{\xi_1} & \cdots & \hat{p}_{\xi_C} \end{bmatrix}^\top$ as follows

$$\mathcal{L}_{\mathrm{F0}}(\boldsymbol{\rho} \,\|\, \hat{\boldsymbol{\rho}}) = \begin{cases} \mathrm{D}_{\mathrm{KL}}(\mathbf{p} \,\|\, \hat{\mathbf{p}})\,, & \text{if } \mathrm{H}(\mathbf{p}) \leq \theta_{\mathrm{h}}\,, \\ 0\,, & \text{otherwise}\ , \end{cases} \tag{4.3}$$

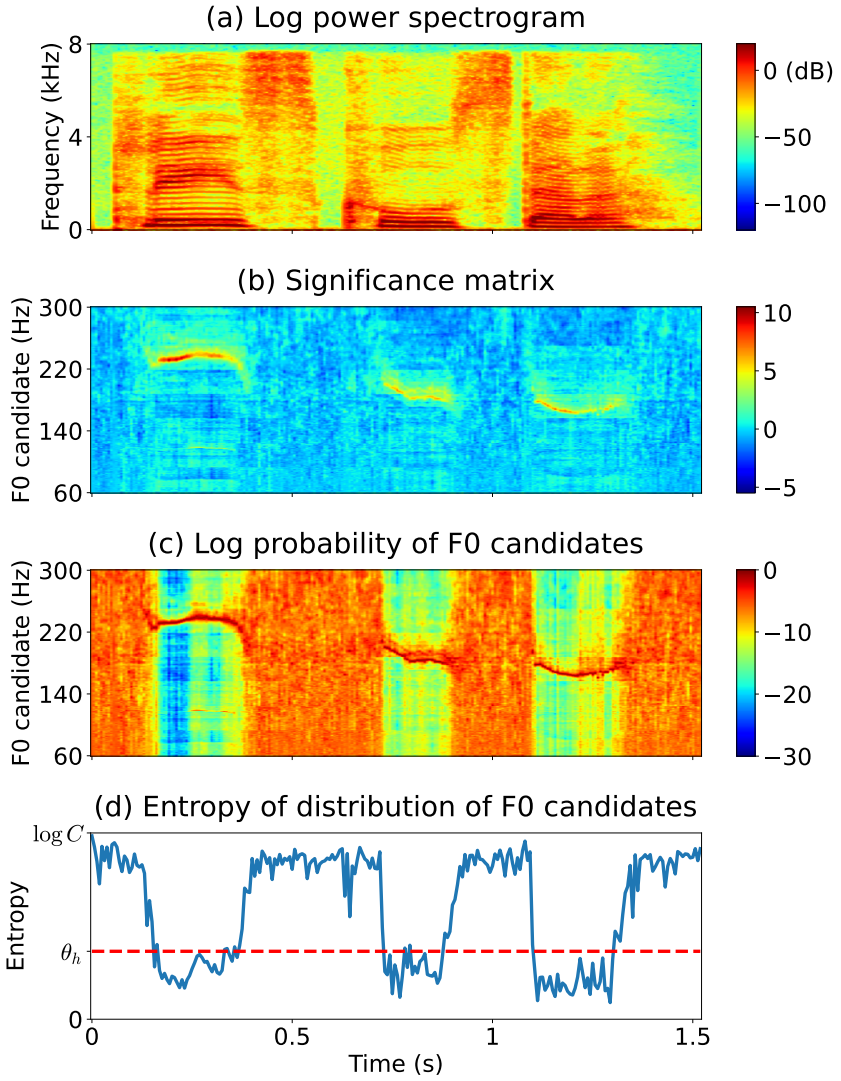

Figure 4.3: Visualization of outputs of each step in computing the discrete distribution of F0 candidates from the log power spectrogram of a speech signal: (a) the input log power spectrogram, (b) the significance matrix of which columns are significance vectors corresponding to the input, (c) the corresponding approximated log probability of F0 candidates, and (d) the corresponding entropy of these F0 distributions. The number of F0 candidates is $C = 241$, where the candidates are linearly located from 60 Hz to 300 Hz (resolution of 1 Hz).

where $\theta_{\mathrm{h}} \in [0, \log C)$ is a fixed threshold. The KL divergence $\mathrm{D}_{\mathrm{KL}}(\mathbf{p} \parallel \hat{\mathbf{p}})$ from $\hat{\mathbf{p}}$ to $\mathbf{p}$ and the entropy $\mathrm{H}(\mathbf{p})$ of $\mathbf{p}$ are respectively defined as

$$\mathrm{D}_{\mathrm{KL}}(\mathbf{p} \parallel \hat{\mathbf{p}}) = \sum_{k=1}^{C} p_{\xi_k} \log \frac{p_{\xi_k}}{\hat{p}_{\xi_k}} , \tag{4.4}$$

$$\mathrm{H}(\mathbf{p}) = -\sum_{k=1}^{C} p_{\xi_k} \log p_{\xi_k} . \tag{4.5}$$

Figure 4.2 illustrates the computation process in Eq. (4.3). Figure 4.3 illustrates the significance matrix and the log probabilities and entropy of the approximated distributions of a speech spectrogram to visualize the quantities in the equations (4.1), (4.2), and (4.5) for each frame.

4.3 Spectral-fine-structure-aware Wiener filter for speech enhancement

4.3.1 Mathematical assumptions

For additive noise, in the STFT domain, the noisy complex spectrogram $\mathbf{Y} \in \mathbb{C}^{K \times M}$ is the sum of the clean-speech complex spectrogram $\mathbf{S} \in \mathbb{C}^{K \times M}$ and the noise complex spectrogram $\mathbf{N} \in \mathbb{C}^{K \times M}$, where K is the number of frequency bins and M is the number of time frames. Let $\tilde{y}_{km}$, $\tilde{s}_{km}$, and $\tilde{n}_{km}$ represent the coefficients of $\mathbf{Y}$, $\mathbf{S}$, and $\mathbf{N}$, respectively, at a frequency bin index $0 \leq k < K$ and a frame index $0 \leq m < M$. The relation between $\tilde{y}_{km}$, $\tilde{s}_{km}$, and $\tilde{n}_{km}$ are described by the following equation

$$\tilde{y}_{km} = \tilde{s}_{km} + \tilde{n}_{km} . \tag{4.6}$$

Also, the complex coefficients of $\mathbf{S}$ and $\mathbf{N}$ are assumed to follow the circularly symmetric complex normal distribution; in other words,

$$\tilde{s}_{km} \sim \mathcal{N}_{\mathbb{C}} \left(0, \sigma^2_{\mathrm{S},\, km}\right) , \tag{4.7}$$

$$\tilde{n}_{km} \sim \mathcal{N}_{\mathbb{C}} \left(0, \sigma^2_{\mathrm{N},\, km}\right) , \tag{4.8}$$

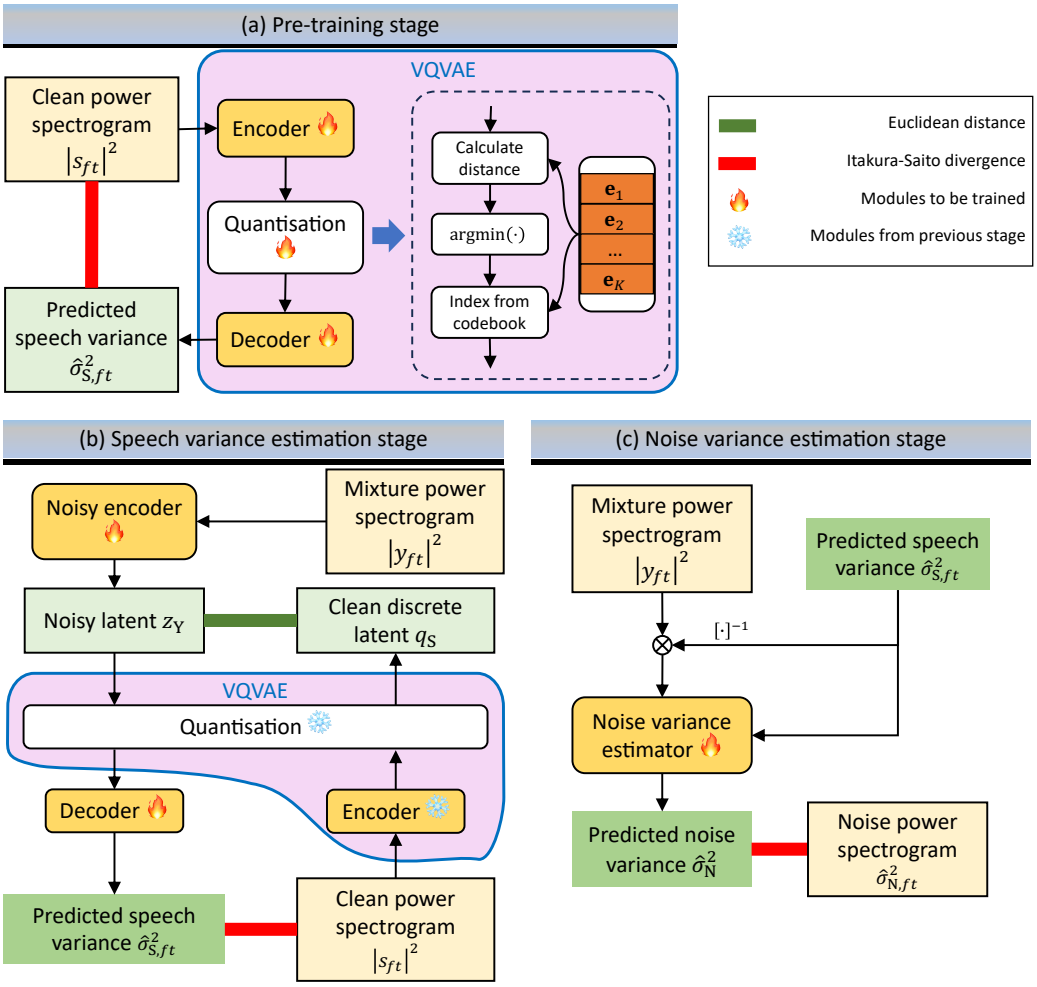

Figure 4.4: Training diagram.

where $\sigma^2_{\mathrm{S},\,km}$ and $\sigma^2_{\mathrm{N},\,km}$ represent the variances of speech and noise, respectively. Assuming that speech and noise are uncorrelated, the spectrogram coefficients of the noisy signal then also follow the complex normal distribution of which variance is the sum of the speech and noise variances:

$$\tilde{y}_{km} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma^2_{\mathrm{S},\,km} + \sigma^2_{\mathrm{N},\,km}\right) . \tag{4.9}$$

Under such constraints, the power spectrograms of speech, noise, and mixture signals follow the exponential distribution as follows

$$|\tilde{s}_{km}|^2 \sim \mathcal{E}xp\left(\sigma^2_{\mathrm{S},\,km}\right)\,, \tag{4.10}$$

$$|\tilde{n}_{km}|^2 \sim \mathcal{E}xp\left(\sigma^2_{\mathrm{N},\,km}\right)\,, \tag{4.11}$$

$$|\tilde{y}_{km}|^2 \sim \mathcal{E}xp\left(\sigma^2_{\mathrm{S},\,km} + \sigma^2_{\mathrm{N},\,km}\right)\,, \tag{4.12}$$

where $\mathcal{E}xp\,(\lambda)$ indicates the exponential distribution with the mean parameter $\lambda > 0$. The log-likelihood function for a sample v to belong to an exponential distribution $\mathcal{E}xp(\lambda)$ is

$$L\,(\lambda|v) = -\frac{v}{\lambda} - \log\lambda\,, \tag{4.13}$$

$$= -\mathrm{d}_{\mathrm{IS}}(v \parallel \lambda) + \mathrm{const}\,, \tag{4.14}$$

where d_{IS} is the Itakura-Saito (IS) divergence defined as

$$\mathrm{d}_{\mathrm{IS}}(v \parallel \lambda) = \frac{v}{\lambda} - \log\frac{v}{\lambda} - 1\,. \tag{4.15}$$

If the speech and noise variances are known, the spectrogram of clean speech can be estimated by applying the ideal ratio mask, of which equation is defined via the Wiener filter as

$$\hat{\tilde{s}}_{km} = \left(\frac{\sigma^2_{\mathrm{S},\,km}}{\sigma^2_{\mathrm{S},\,km} + \sigma^2_{\mathrm{N},\,km}}\right)\tilde{y}_{km}\,. \tag{4.16}$$

For the rest of the chapter, since v and λ are used as the power spectrogram coefficient and its variance parameter for a time-frequency bin, the following notation is conveniently defined

$$\mathrm{D}_{\mathrm{IS}}(\mathbf{X} \parallel \boldsymbol{\lambda}) = \sum_{k,\,m} \mathrm{d}_{\mathrm{IS}}(|\tilde{x}_{km}|^2 \parallel \lambda_{km})\,, \tag{4.17}$$

where $\mathbf{X} \in \mathbb{C}^{K\times M}$ is a complex spectrogram of which element-wise variances are defined in $\boldsymbol{\lambda} \in \mathbb{R}_+^{K\times M}$.

4.3.2 Spectral-fine-structure-aware speech variance estimation using vector-quantized variational autoencoder

Vector-quantized variational autoencoder

Vector-quantized variational autoencoder (VQVAE) is a framework for learning the probability distribution of a dataset. The VQVAE assumes that each observation $\mathbf{x}$ in a dataset $\mathcal{X}$ is stochastically generated from a latent variable $\mathbf{z} \in \mathbb{R}^D$ following the discrete uniform distribution on a set of pseudovectors in a codebook $\mathcal{C} = \left\{\mathbf{e}_1, \cdots, \mathbf{e}_{|\mathcal{C}|}\right\}$ via a conditional distribution (decoder) $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. The VQVAE approximates the intractable posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ as follows

$$\mathbf{z}|\mathbf{x} \sim q_{\boldsymbol{\phi},\mathcal{C}}(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{e}^*) \,, \tag{4.18}$$

where $\delta(\cdot)$ denotes the Dirac delta distribution and $\mathbf{e}^*$ is determined by an encoding-quantization process as follows

$$\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) \qquad \text{(encoding)}\,, \tag{4.19}$$

$$\mathbf{e}^* = \underset{\mathbf{e}\in\mathcal{C}}{\arg\min} \|\boldsymbol{\mu} - \mathbf{e}\|_2^2 \qquad \text{(quantization)}\,. \tag{4.20}$$

The settings of VQVAE resemble a communication system in which the encoder compacts the input vector $\mathbf{x}$ into a code $\boldsymbol{\mu}$, the quantization process maps $\boldsymbol{\mu}$ to the nearest code $\mathbf{e}^*$ in the codebook, and the decoder reconstructs the input. The parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\mathcal{C}$ can be obtained by minimizing

$$\begin{aligned}\mathcal{L}_{\text{VQVAE}}(\boldsymbol{\theta},\boldsymbol{\phi},\mathcal{C}) = -\mathbb{E}_{\hat{\mathbf{z}}\sim q_{\boldsymbol{\phi},\mathcal{C}}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\hat{\mathbf{z}})\right] \\ + \|\text{sg}(\boldsymbol{\mu}) - \mathbf{e}^*\|_2^2 + \beta\|\boldsymbol{\mu} - \text{sg}(\mathbf{e}^*)\|_2^2 \,,\end{aligned} \tag{4.21}$$

where $\text{sg}(\cdot)$ is the stop-gradient operator. As the sampling process of $\hat{\mathbf{z}}$ is not differentiable, the VQVAE uses the reparameterization trick as follows

$$\hat{\mathbf{z}} = \boldsymbol{\mu} + \text{sg}(\mathbf{e}^* - \boldsymbol{\mu}) \,. \tag{4.22}$$

Method for achieving noise-robustness speech variance estimation

The VQVAE utilizes a codebook to capture the data distribution, reducing the latent variable domain to a codebook $\mathcal{C}$ instead of $\mathbb{R}^D$. As a result, the VQVAE, when trained on clean speech, will only produce clean speech samples regardless of input. Applying this property, the VQVAE is initially pre-trained on clean speech. Then, a noisy encoder is trained to denoise speech in the latent space and refine the decoder for noise-robustness using the pre-trained codebook. Figure 4.4 illustrates this training process.

For pre-training, VQVAE is applied to model the distribution of the speech complex spectrogram $\mathbf{S}$ as

$$\tilde{s}_{km}|\hat{\mathbf{z}} \sim \mathcal{N}_{\mathbb{C}}\left(0, \hat{\sigma}^2_{\boldsymbol{\theta},\mathrm{S},km}\left(\hat{\mathbf{z}}\right)\right) , \tag{4.23}$$

which leads to the following loss function

$$\mathcal{L}_{\mathrm{S}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathrm{S}}, \mathcal{C}\right) \;=\; \mathrm{D}_{\mathrm{IS}}(\mathbf{S}\| \;\hat{\boldsymbol{\sigma}}^2_{\mathrm{S}}) + \left\|\mathrm{sg}\left(\boldsymbol{\mu}\right) - \mathbf{e}^*\right\|_2^2 + \beta\left\|\boldsymbol{\mu} - \mathrm{sg}\left(\mathbf{e}^*\right)\right\|_2^2 . \tag{4.24}$$

where

$$\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathrm{S}}}\left(\mathbf{S}\right) \qquad \text{(encoding)}, \tag{4.25}$$

$$\mathbf{e}^* = \underset{\mathbf{e}\in\mathcal{C}}{\arg\min}\left\|\boldsymbol{\mu} - \mathbf{e}\right\|_2^2 \qquad \text{(quantization)}, \tag{4.26}$$

$$\hat{\mathbf{z}} = \boldsymbol{\mu} + \mathrm{sg}\left(\mathbf{e}^* - \boldsymbol{\mu}\right) \qquad \text{(reparameterization)}, \tag{4.27}$$

$$\hat{\boldsymbol{\sigma}}^2_{\mathrm{S}} \triangleq \hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\theta},\mathrm{S}}\left(\hat{\mathbf{z}}\right) \qquad \text{(decoding)}. \tag{4.28}$$

In the main training stage, the loss function for speech variance estimation is

$$\mathcal{L}_{\mathrm{Y}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathrm{Y}}\right) = \mathrm{D}_{\mathrm{IS}}(\mathbf{S} \,\|\, \hat{\boldsymbol{\sigma}}^2_{\mathrm{S}'}) + \beta\left\|\boldsymbol{\mu}' - \mathrm{sg}\left(\mathbf{e}^*\right)\right\|_2^2 , \tag{4.29}$$

where

$$\boldsymbol{\mu}' \triangleq \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathrm{Y}}}\left(\mathbf{S}\right) \qquad \text{(noisy encoding)}, \tag{4.30}$$

$$\mathbf{e}^{*\prime} = \underset{\mathbf{e}\in\mathcal{C}}{\arg\min}\left\|\boldsymbol{\mu}' - \mathbf{e}\right\|_2^2 \qquad \text{(quantization)}, \tag{4.31}$$

$$\hat{\mathbf{z}}' = \boldsymbol{\mu}' + \mathrm{sg}\left(\mathbf{e}^{*\prime} - \boldsymbol{\mu}'\right) \qquad \text{(reparameterization)}, \tag{4.32}$$

$$\hat{\boldsymbol{\sigma}}^2_{\mathrm{S}'} \triangleq \hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\theta},\mathrm{S}}\left(\hat{\mathbf{z}}'\right) \qquad \text{(decoding)}. \tag{4.33}$$

Spectral-fine-structure-aware speech variance estimation

The speech variance estimation framework introduced above is based on IS divergence, which rather focuses on the spectral envelope and tends to ignore the spectral fine structure [106]. To solve this problem, the proposed method simply adds the $\mathcal{L}_{\mathrm{F0}}$ to the speech variance estimation loss in both states, in other words,

$$\mathcal{L}_{\mathrm{S\ (with\ SFS\text{-}aware)}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathrm{S}}, \mathcal{C}\right) = \mathcal{L}_{\mathrm{S}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathrm{S}}, \mathcal{C}\right) + \mathrm{L}_{\mathrm{F0}}(\mathbf{S} \parallel \hat{\boldsymbol{\sigma}}^2_{\mathrm{S}}), \tag{4.34}$$

$$\mathcal{L}_{\mathrm{Y\ (with\ SFS\text{-}aware)}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathrm{S}}, \mathcal{C}\right) = \mathcal{L}_{\mathrm{Y}}\left(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathrm{Y}}, \mathcal{C}\right) + \mathrm{L}_{\mathrm{F0}}(\mathbf{S} \parallel \hat{\boldsymbol{\sigma}}^2_{\mathrm{S}'}), \tag{4.35}$$

where

$$\mathrm{L}_{\mathrm{F0}}(\mathbf{S} \parallel \hat{\boldsymbol{\sigma}}^2) = \sum_m \mathcal{L}_{\mathrm{F0}}(\mathbf{s}_m \parallel \hat{\boldsymbol{\sigma}}^2_m). \tag{4.36}$$

4.3.3 Noise Variance Estimation

The noise variance estimator is trained to reduce the IS divergence between the noise spectrogram and the predicted noise variance as follows

$$\mathcal{L}_{\mathrm{N}}\left(\boldsymbol{\theta}_{\mathrm{N}}\right) = \mathrm{D}_{\mathrm{IS}}(\mathbf{N} \parallel \hat{\boldsymbol{\sigma}}^2_{\mathrm{N},\boldsymbol{\theta}_{\mathrm{N}}}). \tag{4.37}$$

The noisy speech log-power spectrogram is empirically subtracted from the estimated speech variance to condition the noise variance estimator on the estimated speech variance. Although it is not entirely accurate, this results in a representation that resembles the noise log-power spectrogram better than the noisy log-power spectrogram, where speech information is suppressed.

4.4 Experiments

4.4.1 Dataset

The proposed method is evaluated using the open dataset from Valentini *et al.* [37]. This dataset has been used in several recent speech enhancement studies, included here as baselines. The clean training set comprises 28 speakers (14 males and 14 females) and the test set of two speakers (one male and one female) from the Voice Bank corpus [107]. The noisy training set

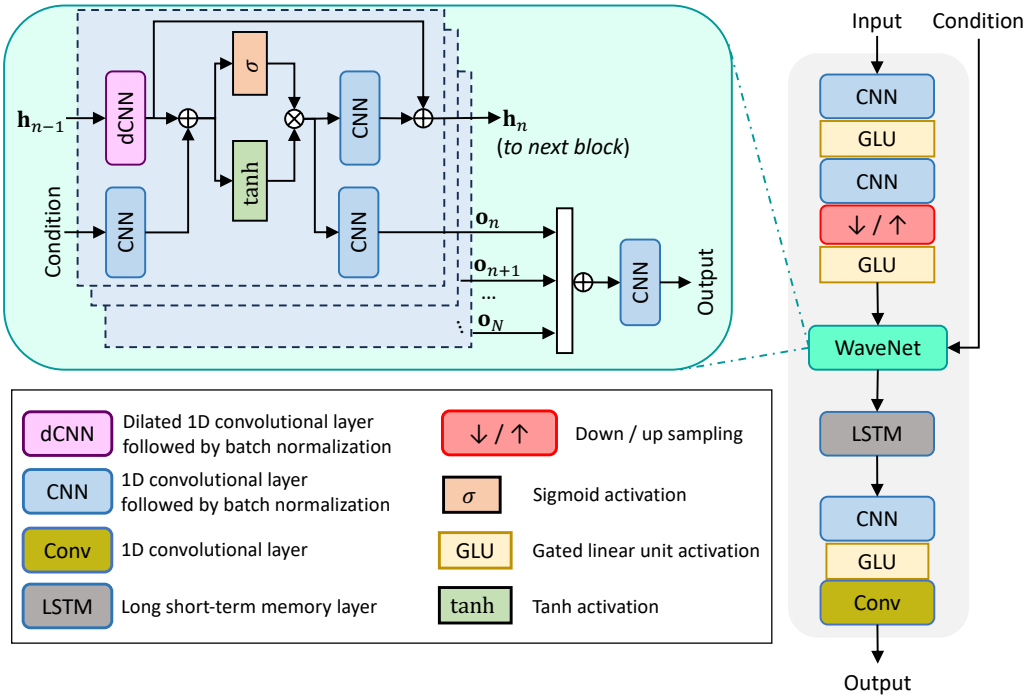

Figure 4.5: Architecture of WaveNet.

is constructed by mixing the clean training set with ten types of noise data at four signal-to-noise ratios (SNRs): 15 dB, 10 dB, 5 dB, and 0 dB. Eight of ten noise types are real recorded noise from the DEMAND dataset [108], while the other two (speech-shaped and babble) are artificially generated. The noisy test set is constructed by mixing the clean test set with five other types of noises from DEMAND dataset [108] at four SNRs: 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. The training and test sets have no speakers or noise types in common. All speech waveforms are resampled from 48 kHz to a 16 kHz sampling rate. To improve the data variance, the input speech is randomly scaled between -35 dB and -20dB.

4.4.2 Configurations for $\mathcal{L}_{\text{F0}}$

For $\mathcal{L}_{\text{F0}}$ computation, the F0 distribution over a set of $C = 241$ F0 candidates linearly located from 60 Hz to 300 Hz is used; in other words, the resolution is 1 Hz. To avoid numerical problems, mean-and-variance normalization is applied to the significance matrix before applying the Softmax function.

The $\mathcal{L}_{\text{F0}}$ contains two hyperparameters: the Softmax temperature ι and the entropy threshold θ_{h}. The principle of selecting these hyperparameters is that these parameters should be optimal for F0 detection using entropy (Eq. (4.5)) on the training dataset. Therefore, first, the Softmax temperature is selected so that the frames containing F0 should have a lower entropy. As the entropy $\text{H}(\mathbf{p})$ is bounded between 0 and $\log C$, the Softmax temperature ι is selected to minimize the loss of binary classification as follows:

$$\mathcal{L}_{\text{BCE}}(\iota) = -\sum_{\text{frame } \boldsymbol{\rho}} z(\boldsymbol{\rho}) \log\left(1 - \frac{H(\mathbf{p})}{\log C}\right) + (1 - z(\boldsymbol{\rho})) \log \frac{H(\mathbf{p})}{\log C}\,, \tag{4.38}$$

where

$$z(\boldsymbol{\rho}) = \begin{cases} 1\,, & \text{if the frame } \rho \text{ contains F0}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{4.39}$$

The ground truth for F0 detection is obtained using PYIN algorithm on the data.

Different $\mathcal{L}_{\text{BCE}}(\iota)$ under different ι are visualized in Figure 4.6(a), indicating the optimal $\iota \approx 0.45$. Then, the entropy threshold is selected to optimize the F0 detection results based on $\text{H}(\mathbf{p})$

$$\hat{z}(\boldsymbol{\rho}) = \begin{cases} 1\,, & \text{if } H(\mathbf{p}) < \theta_{\text{h}}\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{4.40}$$

In this scenario of binary classification where 1 is positive label and 0 is negative label, false positives, i.e., detecting F0 in a frame which does not containing F0, are more dangerous as it makes the model optimize a meaningless property. Therefore, precision, accuracy, and balance accuracy are used to analyze the detection rate at different θ_{h}. The results are visualized in Figure 4.6(b). While the optimal θ_{h} is around $0.65 \log C$, the metrics drop very fast as the threshold increases farther from this value, which makes this optimal value sensitive. Therefore, the implementation of the $\mathcal{L}_{\text{F0}}$ uses $\theta_h = 2 \approx 0.36 \log C$ to keep the stable precision while does not reduce too much the accuracy and balance accuracy.

4.4.3 Implementation

In the implementation, the STFT uses the hanning window function with a window length of 25 ms (400 samples) and a hop length of 6.25 ms (100 samples). The number of points for the fast Fourier transform is 512, which results in 257 frequency bins.

The WaveNet-based module shown in Figure 4.5 is the basic block to construct our models. The VQVAE is implemented with a U-Net hierarchical structure of two layers (bottom and top) similar to [109] for the speech variance estimator. The bottom VQVAE aims to capture information in a higher temporal resolution, followed by the top, which tries to model the features in a lower temporal resolution. The noise variance estimator and all the encoders and decoders in the speech variance estimator have the architecture of the WaveNet-based module without the long short-term memory (LSTM) layer. The inputs and outputs of the neural network are log-compressed to improve estimation performance [110].

The training procedure consists of two stages, shown in Figure 4.4, which are the pre-training and main-training stages. In the pre-training stage, the VQVAE is trained to estimate speech variance using the clean training set of 1000 epochs. The data-dependent codebook re-estimation [111] and EMA algorithm [112] are employed for codebook updating to obtain higher codebook perplexity. The noisy encoders, decoders, noise variance estimator, and phase correction network are trained for 1000 epochs in the main training stage. An Adam optimizer and a One-cycle Learning Rate scheduler with an initial learning rate of 5×10^{-4} and a maximum learning rate of 2×10^{-4} are used for all training stages.

4.4.4 Evaluation Metrics

The Wide-band Perceptual Evaluation of Speech Quality (PESQ-WB) [85,86] and Short-Time Objective Intelligibility (STOI) [89] metrics are used to evaluate the overall performance of the proposed method. The PESQ scores, which range from -0.5 (bad) to 4.5 (excellent), measure the speech quality by comparing the enhanced signal to the clean reference speech signal. The STOI metric is highly correlated to perceptual speech intelligibility. The STOI scores range between 0 (lowest intelligibility) and 1 (highest intelligi-

Table 4.1: Performance of the proposed method with different speech variance estimation model configurations on Valentini *et al.* dataset. All the results in this table are obtained without phase correction.

Speech variance models	PESQ-WB	STOI
Without $\mathcal{L}_{\text{F0}}$	2.815	0.941
With $\mathcal{L}_{\text{F0}}$	**2.817**	**0.942**

Table 4.2: Improvement of PESQ-WB and STOI for each speaker in the test set of Valentini *et al.* dataset when applying $\mathcal{L}_{\text{F0}}$. All the results in this table are obtained without phase correction.

Speaker	PESQ-WB Impr. (%)	STOI Impr. (%)
`p232` (male)	0.434	0.006
`p257` (female)	0.340	0.144
Overall	0.385	0.078

bility). For both metrics, a higher score indicates a better result.

4.4.5 Results and discussion

The effectiveness of the proposed spectral-fine-structure enhancement method is analyzed by comparing the model's performance with and without $\mathcal{L}_{\text{F0}}$. The results reported in Table 4.1 show that using $\mathcal{L}_{\text{F0}}$ can slightly improve speech enhancement performance. For further analysis, the improvement percentage of PESQ-WB and STOI for each speaker is evaluated on the test set when applying $\mathcal{L}_{\text{F0}}$, where the improvement percentage of a metric after an operation is defined as

$$\text{Impr.}(\%) = \frac{\text{Final} - \text{Initial}}{\text{Initial}} \times 100\%\,, \tag{4.41}$$

where Final and Initial are metric values of the model before and after applying that operation.

The results in Table 4.2 show that the PESQ-WB and STOI improve overall, which means the proposed fine-structure enhancement technique helps model speech variance better. The improvement percentage of PESQ-WB

is higher for the male speaker, which aligns with our original purpose when proposing $\mathcal{L}_{\text{F0}}$ to improve spectral fine structure when F0 is low. On the other hand, the improvement percentage of STOI is lower for the male speaker. The hypothesis for the cause of this result is the limited capacity of the speech variance estimation model. There is a trade-off when modeling the spectral envelope and spectral fine structure of highly fluctuated spectra from the male speaker. The spectral envelope is more associated with linguistic information and speech intelligibility than the spectral fine structure. Therefore, the fine structure enhancement does not improve STOI much for the male speaker. This trade-off is less severe for the female speaker with a higher F0 and less fluctuated spectral fine structure. Before using the proposed spectral enhancement, the estimated spectral fine structure was already better for higher F0 spectra, which resulted in a lower improvement in PESQ-WB. However, with the higher improvement of STOI for the female speaker, the $\mathcal{L}_{\text{F0}}$ forces the model to utilize its capacity in modeling speech variance.

Figure 4.7 illustrates the estimated speech variances and Wiener filters with and without using spectral-fine-structure enhancement for some samples in the Valentini *et al.* test dataset [37]. The figure visualizes the phenomena described above and confirms our hypothesis for the effectiveness of the proposed $\mathcal{L}_{\text{F0}}$. For the frames with low F0 in the first three samples, the use of $\mathcal{L}_{\text{F0}}$ improves the spectral fine structure of predicted speech variance significantly in the pre-training stage, which results in a Wiener filter estimation with closer detailed patterns to the clean speech spectrogram. Nevertheless, the formants and formant transitions are more apparent when not using $\mathcal{L}_{\text{F0}}$, especially with the weak-harmonic frames. For the frames with high F0, the effect of $\mathcal{L}_{\text{F0}}$ is not visually significant.

Ablation study: The Role of IS Divergence

This section analyzes the use of IS divergence in the model. While the IS divergence is essential for the proposed speech enhancement method based on certain assumptions, it overlooks the spectral fine structure [30]. As a result, we explore the impact of substituting the IS divergence with the log spectral distortion (LSD), a commonly used metric in the spectral domain, described by the equation:

$$\mathrm{d}_{\mathrm{LSD}}(x \parallel y) = \frac{1}{2}\left(\log\frac{x}{y}\right)^2 . \tag{4.42}$$

The first and third rows of Table 4.1 indicate that using LSD leads to a decrease in model performance compared to using IS divergence, despite the speech variance estimated by optimizing LSD displaying more similar detailed patterns. Using LSD alters the statistical assumption of speech and noise power spectrograms from the exponential distribution to the log-normal distribution. Under this new assumption, the distribution of the noisy speech also changes, and its parameter no longer possesses the additive property as in (4.12). Consequently, the Wiener filter is no longer applicable. Therefore, despite capturing more detailed patterns, LSD does not address the improvement of spectral fine structure due to the violation of assumptions, necessitating the incorporation of $\mathcal{L}_{\mathrm{F0}}$.

4.5 Summary

In summary, the chapter proposes a method for modeling the spectral modulation characteristics of speech and applying them to speech enhancement. Specifically, it introduces a method that uses the discrete F0 distribution to model the spectral structure characteristics of voiced speech, enabling us to measure the difference in spectral speech characteristics quantitatively, namely $\mathcal{L}_{\mathrm{F0}}$. Then, a speech enhancement method is proposed based on Wiener filtering, which applies the proposed $\mathcal{L}_{\mathrm{F0}}$ as a loss function for spectral-fine-structure-aware speech variance estimation. The results demonstrate that (1) the method effectively models the spectral modulation characteristics of speech and (2) applying the proposed method to speech enhancement improves the quality and intelligibility of the enhanced speech. These results satisfy the first objective of this research, supporting the answer to the first research question: Enhancing the speech characteristics in amplitude can improve speech enhancement performance.

In summary, the chapter proposed a method for modeling the spectral modulation characteristics of speech and applying them to speech enhancement. Specifically, it introduced a method that used the discrete F0 distribution to model the spectral structure characteristics of voiced speech,

enabling us to measure the difference in spectral speech characteristics quantitatively, namely $\mathcal{L}_{\mathrm{F0}}$. Then, a speech enhancement method was proposed based on Wiener filtering, which applied the proposed $\mathcal{L}_{\mathrm{F0}}$ as a loss function for spectral-fine-structure-aware speech variance estimation. The results demonstrated that (1) the method effectively modeled the spectral modulation characteristics of speech and (2) applying the proposed method to speech enhancement improved the quality and intelligibility of the enhanced speech. These results satisfied the first objective of this research, supporting the answer to the first research question, which is: Enhancing the speech characteristics in amplitude can improve speech enhancement performance

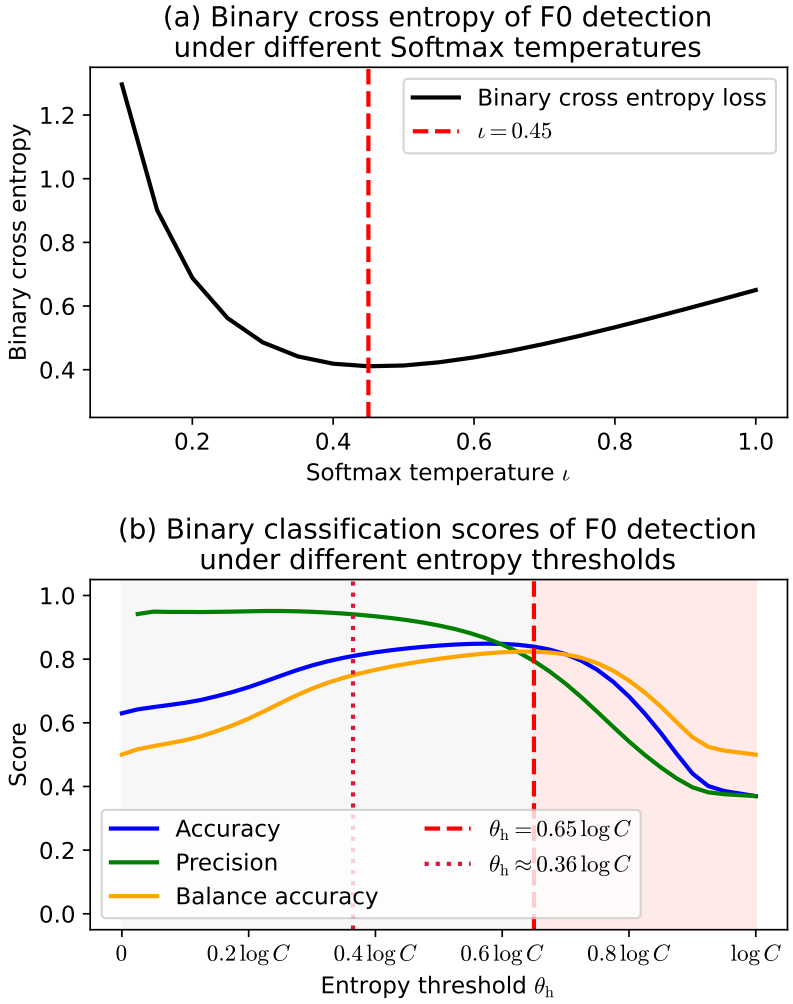

Figure 4.6: Data analysis results for selecting the Softmax temperature and entropy threshold of $\mathcal{L}_{\text{F0}}$: (a) Softmax temperature ι and (b) entropy threshold θ_{h}.

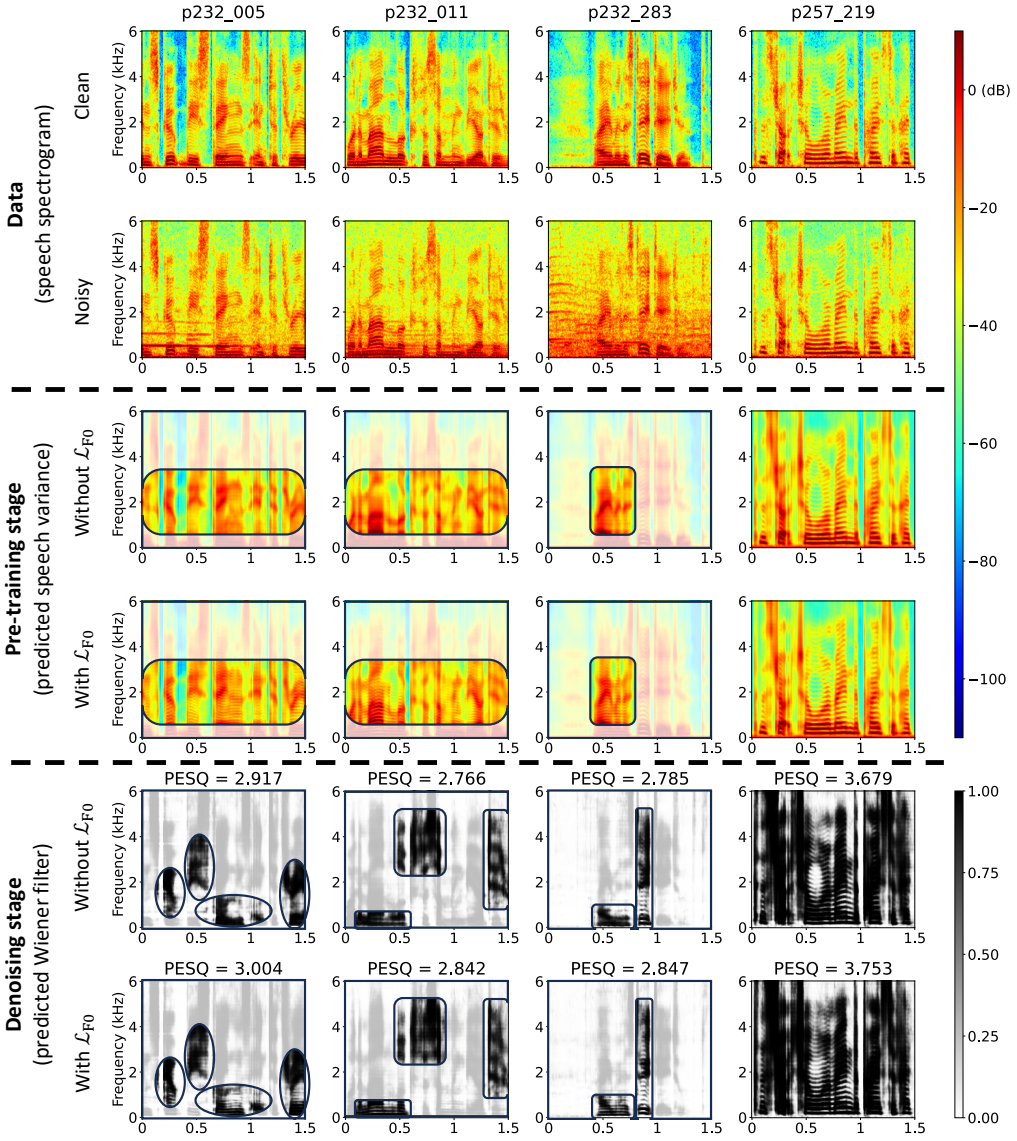

Figure 4.7: Samples of the effect of $\mathcal{L}_{\text{F0}}$ in the estimated variance of speech power spectrogram in pre-training and main training (denoising) stages on four different samples (four columns) from the Valentini test set. The top figures are the clean and noisy speech power spectrograms as references. The differences between the outputs of the proposed model with and without using $\mathcal{L}_{\text{F0}}$ are highlighted in the first three samples.

Chapter 5

Relationship between amplitude and IFD in the time-frequency representation

In the complex time-frequency representation, while the amplitude contains clear patterns to analyze, the phase is much more complicated. Using the approach of "explaining the unknown from the known," this chapter seeks to answer the second research question in this dissertation: 'Is there a connection between the phase and the amplitude.' Specifically, this chapter proposes the analytical derivative method, establishing an equation connecting the amplitude and the instantaneous frequency deviation (IFD), a phase feature based on modulation theory. Using single-tone frequency-modulated (FM) signals (ground-truth sinusoidal IFD) for evaluation, the results show that the IFD extracted by the proposed method is close to the ground-truth IFD (with small root-mean-squared error). Therefore, the equation connecting amplitude and IFD is correct, answering the second research question.

5.1 Problem formulation

The complex time-frequency representation $\tilde{x}(\omega, \tau)$ of a signal contains two components: amplitude $A(\omega, \tau) = |\tilde{x}(\omega, \tau)|$ and phase $\phi(\omega, \tau) = \angle\tilde{x}(\omega, \tau)$. If $\tilde{x}(\omega, \tau)$ is not only amplitude-modulated but also frequency-modulated along

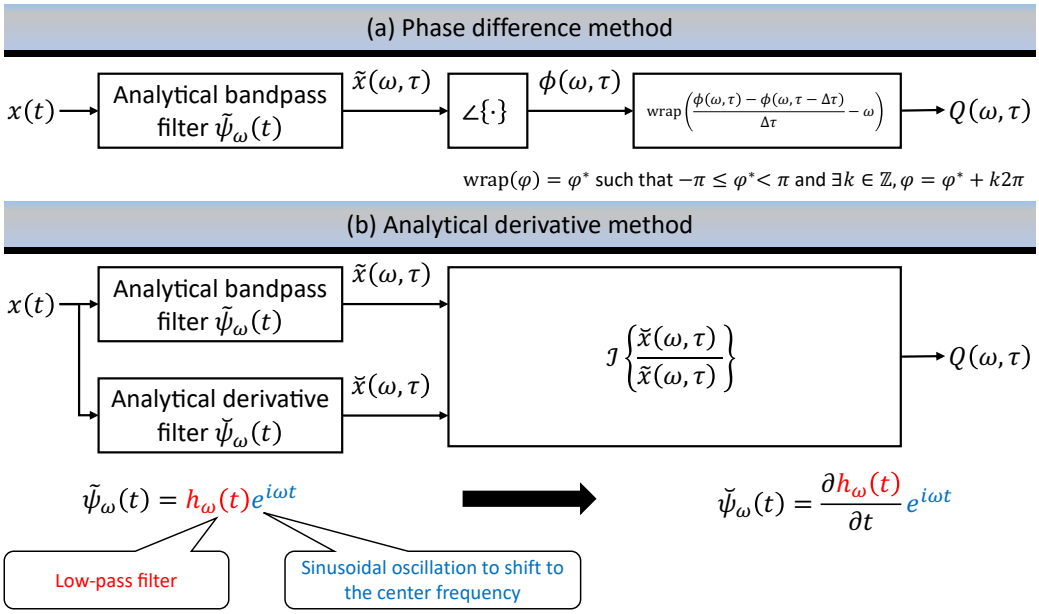

Figure 5.1: Block diagram of IFD extraction methods: (a) phase difference method (conventional) and (b) analytical derivative method (proposed).

the temporal axis with carrier frequency ω, then $\tilde{x}(\omega,\tau)$ must satisfy

$$\begin{aligned}\tilde{x}(\omega,\tau) &= A(\omega,\tau)\exp i\phi\,(\omega,\tau)\\ &= A(\omega,\tau)\exp\left\{i\int_0^\tau[\omega+Q(\omega,\eta)]\,d\eta\right\}\\ &= A(\omega,\tau)\exp\left[i\omega\tau+i\int_0^\tau Q(\omega,\eta)d\eta\right],\end{aligned}\tag{5.1}$$

where $Q(\omega,\tau)$ is the instantaneous frequency deviation (IFD), which is the deviation of the instantaneous frequency from the carrier frequency ω. As a results, $Q(\omega,\tau)$ can be derived straightforwardly by the following formula

$$Q(\omega,\tau)=\frac{\partial\phi(\omega,\tau)}{\partial\tau}-\omega\,.\tag{5.2}$$

In discrete system, Eq. (5.2) allows to approximate the IFD as shown in Figure 5.1 (a), namely phase difference method. While this method is simple, it cannot quantitatively explain the relationship between $A(\omega,\tau)$ and $Q(\omega,\tau)$. In addition, as the phase is circular-valued, the phase difference also suffers

the phase-wrapping issue. To solve this issue, the next section establishes another $Q(\omega, \tau)$ equation that connects with $A(\omega, \tau)$.

5.2 Analytical derivative method for IFD extraction

The analytical derivative method applies the technique in instantaneous frequency extraction by Murty and Yegnanarayana [113], which takes the logarithm (principal value) in both sides of Eq. (5.1), resulting in

$$\log \tilde{x}(\omega, \tau) = \log A(\omega, \tau) + i\omega\tau + i \int_0^\tau Q(\omega, \eta) d\eta \,. \tag{5.3}$$

Then, taking the time derivative on both sides of Eq. (5.3) results in

$$\frac{1}{\tilde{x}(\omega, \tau)} \cdot \frac{\partial \tilde{x}(\omega, \tau)}{\partial \tau} = \frac{1}{A(\omega, \tau)} \cdot \frac{\partial A(\omega, \tau)}{\partial \tau} + i\omega + iQ(\omega, \tau) \,. \tag{5.4}$$

Assuming that $A(\omega, \tau)$ is continuous and differentiable with respect to τ, $Q(\omega, \tau)$ can be computed as

$$\begin{aligned} Q(\omega, \tau) &= \mathcal{I} \left\{ \frac{1}{\tilde{x}(\omega, \tau)} \cdot \frac{\partial \tilde{x}(\omega, \tau)}{\partial \tau} - i\omega \right\} \\ &= \mathcal{I} \left\{ \frac{1}{\tilde{x}(\omega, \tau)} \cdot \left(\frac{\partial \tilde{x}(\omega, \tau)}{\partial \tau} - i\omega \tilde{x}(\omega, \tau) \right) \right\} \\ &= \mathcal{I} \left\{ \frac{\breve{x}(\omega, \tau)}{\tilde{x}(\omega, \tau)} \right\} \end{aligned} \,, \tag{5.5}$$

where $\mathcal{I}\{\cdot\}$ indicates the imaginary part of a complex number and

$$\breve{x}(\omega, \tau) = \frac{\partial \tilde{x}(\omega, \tau)}{\partial \tau} - i\omega \tilde{x}(\omega, \tau) \,. \tag{5.6}$$

The term $\tilde{x}(\omega, \tau)$ is a complex time-frequency representation of $x(t)$; in other words, $\tilde{x}(\omega, \tau)$ is the result of filtering $x(t)$ by an analytical band-pass

filter $\tilde{\psi}_\omega$ of which center frequency is ω (rad/s) (see Section 2.1.2). Therefore,

$$\begin{aligned}\tilde{x}(\omega,\tau) &= \left(x * \tilde{\psi}\right)(\tau) \\ &= \int_{-\infty}^{\infty} x(\tau - t)\,\tilde{\psi}_\omega(t)\,dt\,. \end{aligned} \tag{5.7}$$

As a result,

$$\begin{aligned}\breve{x}(\omega,\tau) &= \int_{-\infty}^{\infty} x(\tau - t)\left(\frac{d\tilde{\psi}_\omega(t)}{dt} - i\omega\tilde{\psi}_\omega(t)\right)dt \\ &= \left(x * \breve{\psi}_\omega\right)(\tau) \qquad\qquad\qquad , \end{aligned} \tag{5.8}$$

where

$$\breve{\psi}_\omega(t) = \frac{d\tilde{\psi}_\omega(t)}{dt} - i\omega\tilde{\psi}_\omega(t)\,. \tag{5.9}$$

As $\tilde{\psi}_\omega(t)$ is an analytical band-pass filter, $\tilde{\psi}_\omega(t)$ typically has the form of

$$\tilde{\psi}_\omega(t) = h_\omega(t)e^{i\omega t}\,, \tag{5.10}$$

where $h_\omega(t)$ is the low-pass filter and $e^{i\omega t}$ is the oscillation term that shifts the spectrum of $h_\omega(t)$ to the center frequency ω [114]. The form in Eq. (5.10) is commonly seen in several filterbanks, for examples,

- Gabor wavelet where $h_\omega(t)$ is a Gaussian distribution,
- Gammatone wavelet where $h_\omega(t)$ is a gamma distribution, and
- STFT where $h_\omega(t)$ is the window function, i.e., $h_\omega(t)$ does not depend on ω.

When $\tilde{\psi}_\omega(t)$ has the form above, Eq. (5.9) beceomes

$$\breve{\psi}_\omega(t) = \frac{dh_\omega(t)}{dt}\,e^{i\omega t}\,, \tag{5.11}$$

which can be derived analytically when $h_\omega(t)$ is given.

Eq. (5.5), Eq. (5.8), and Eq. (5.11) constructs an alternative method to extract the IFD $Q(\omega,\tau)$ without phase calculation, namely analytical derivative method, of which process diagram is illustrated in Figure 5.1 (b).

5.3 Validation simulation

This section describes a validation simulation to verify the correctness of the proposed analytical derivative method for IFD extraction in the previous section. As aforementioned, extracting IFD means extracting the modulating signal of an FM signal. Therefore, this evaluation requires an FM signal with known IFD (i.e., modulating signal) to compare the extracted IFD and the ground-truth IFD. In communication theory, the maximum modulating frequency and peak frequency deviation of the modulating signal are two parameters that characterize the properties of FM signals. Therefore, this section uses a single-tone FM signal to validate the proposed analytical method because (1) single-tone FM signal only varieties by the two parameters above and (2) the modulating signal is known.

5.3.1 Single-tone FM signal

A single-tone FM signal is a type of FM signal with constant modulating frequency; in other words, the modulating signal is sinusoidal. The mathematical representation of a single-tone FM signal can be expressed as

$$x(t) = \cos\left(2\pi f_{\mathrm{c}} t + \int_0^t Q(\tau) d\tau\right), \tag{5.12}$$

where $Q(t)$ is the sinusoidal modulating signal, i.e.,

$$Q(t) = 2\pi f_{\Delta} \cos\left(2\pi f_{\mathrm{m}} t\right). \tag{5.13}$$

The parameters defining $Q(t)$ includes

- f_{Δ} is the peak frequency deviation (in Hz), and
- f_{m} is the modulating frequency (in Hz).

Derived from f_{Δ} and f_{m}, two other parameters characterize an FM signal are

- Modulation index β measures how much the frequency of the carrier signal is allowed to vary due to the modulating signal, which has the following formula

$$\beta = \frac{f_{\Delta}}{f_{\mathrm{m}}}. \tag{5.14}$$

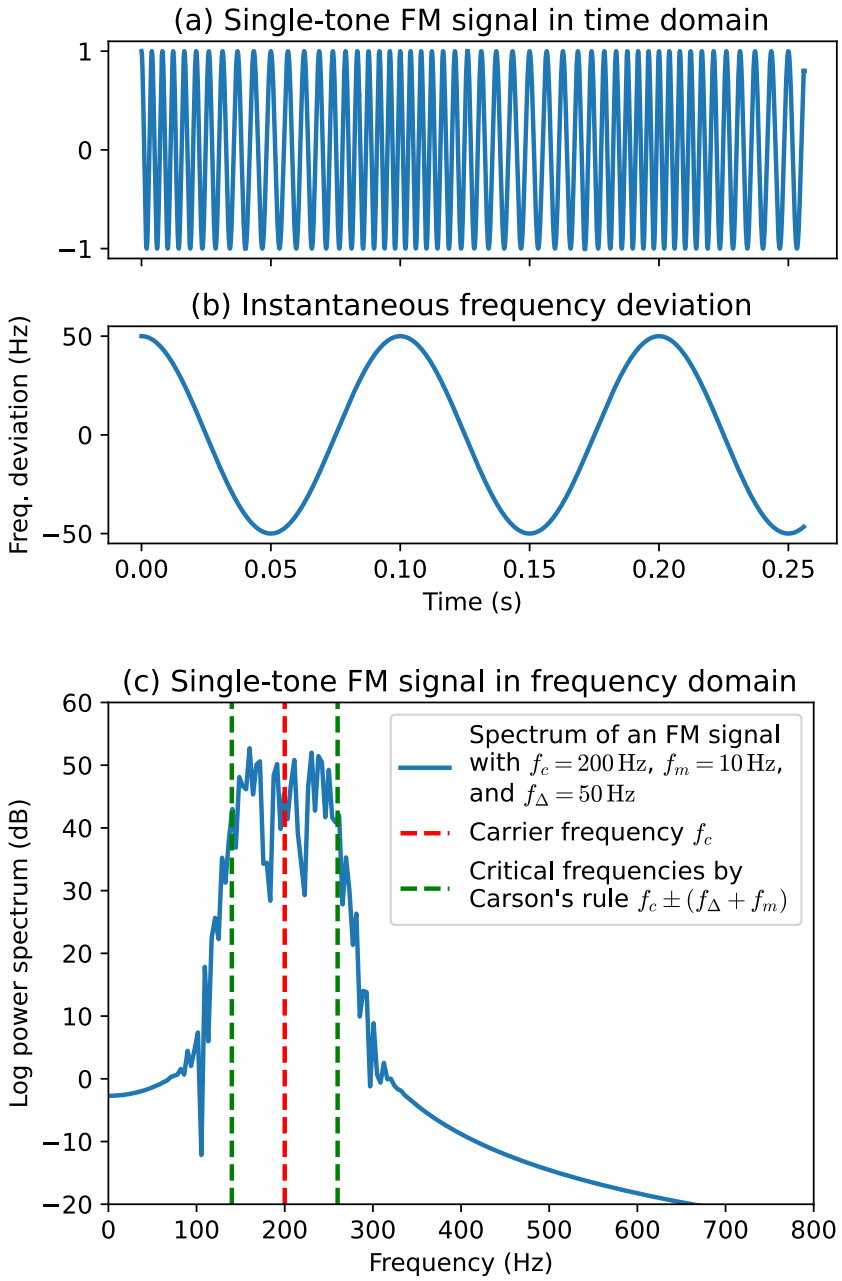

Figure 5.2: Example of a single-tone FM signal with $f_\mathrm{c} = 400\,\mathrm{Hz}$, $f_\mathrm{m} = 10\,\mathrm{Hz}$, and $f_\Delta = 50\,\mathrm{Hz}$: (a) the signal in the time domain, (b) the instantaneous frequency deviation of the signal, and (c) the power spectrum of the signal.

- Carson's bandwidth (in Hz) is an estimation of the bandwidth of a frequency-modulated (FM) signal defined by Carson's rule [115] as

$$B = 2\left(f_\Delta + f_\mathrm{m}\right) = 2f_\mathrm{m}\left(\beta + 1\right) . \tag{5.15}$$

Figure 5.2 illustrates a single-tone FM signal, its IFD (modulating signal), and its power spectrum. Most of the signal's power stays within Carson's bandwidth.

5.3.2 Simulation procedure and configurations

Procedure

Using the single-tone FM signal, the simulation procedure is as follows:

1. Input the following parameters: f_c, f_m and f_Δ,
2. Construct a single-tone FM signal $x(t)$ with modulating signal $Q(t)$ shown in Eq. (5.12) and Eq. (5.13),
3. Construct the analytical band-pass filter $\tilde{\psi}_\omega(t)$ with the center frequency equal to the carrier frequency, i.e., $\omega = 2\pi f_\mathrm{c}$,
4. Use phase difference and analytical derivative methods (Figure 5.1) to estimate $Q(t)$, and
5. Use root-mean-squared error (RMSE) to measure the estimation error.

Configurations

As Hanning-window STFT was used in the previous chapter for speech enhancement, this simulation also uses the Hanning window for band-pass filtering. In this case, the analytical band-pass filter becomes

$$\tilde{\psi}_\omega(t) = w(t)e^{i\omega t} , \tag{5.16}$$

where $w(t)$ indicates the Hanning window of length T

$$w(t) = \begin{cases} \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi t}{T}\right), & -\frac{T}{2} \le t \le \frac{T}{2} \\ 0\,, & \text{otherwise} \end{cases} . \tag{5.17}$$

The window length is used with the same configurations as the previous chapter, i.e., $T = 25\,\text{ms}$. In addition, the analytical derivative method requires the first derivative of the window function, i.e.,

$$\frac{d}{dt}w(t) = \begin{cases} -\frac{\pi}{T}\sin\left(\frac{2\pi t}{T}\right), & -\frac{T}{2} \leq t \leq \frac{T}{2} \\ 0\,, & \text{otherwise} \end{cases}. \tag{5.18}$$

The simulation examines differnt scenarios of f_m and f_Δ as follows:

- Modulating frequency f_m in the range of $(0\,\text{Hz}, 50\,\text{Hz}]$ with a resolution of $1\,\text{Hz}$, and
- Peak frequency deviation f_Δ in the range of $[0\,\text{Hz}, 70\,\text{Hz}]$ with a resolution of $1\,\text{Hz}$.

As carrier frequency f_c does not contribute much, f_c is set to a safe value based on Carson's rule [115], which is ten times greater than Carson's bandwidth, i.e., $f_\text{c} = 2400\,\text{Hz}$. All the FM signals are sampled at $16\,\text{kHz}$ sampling rate with a duration of around two seconds ($32,768$ samples).

5.3.3 Results

Figure 5.3 illustrates the validation results by pairwisely comparing the RMSE of the IFD extracted by the proposed analytical derivative method, the conventional phase difference method, and the ground-truth IFD provided by the modulating signal. The results show that the proposed analytical derivative method provides a nearly equivalent IFD estimation with the conventional phase difference method, where the error is less than $1\,\text{Hz}$. In other words, the two methods are nearly equivalent, indicating that the formula used in the proposed analytical method is correct.

In addition, the estimation errors of both methods increase in the direction of the increase in Carson's bandwidth. When either f_m or f_Δ increases while the other is constant, the estimation errors of both methods when compared to the ground-truth IFD increase (see Figure 5.4 and Figure 5.5). However, with the same Carson's bandwidth, the FM signals with higher f_m have larger estimation error. The reason for this phenomenon is that despite having the same bandwidth, the signal with higher f_m contains peaks that

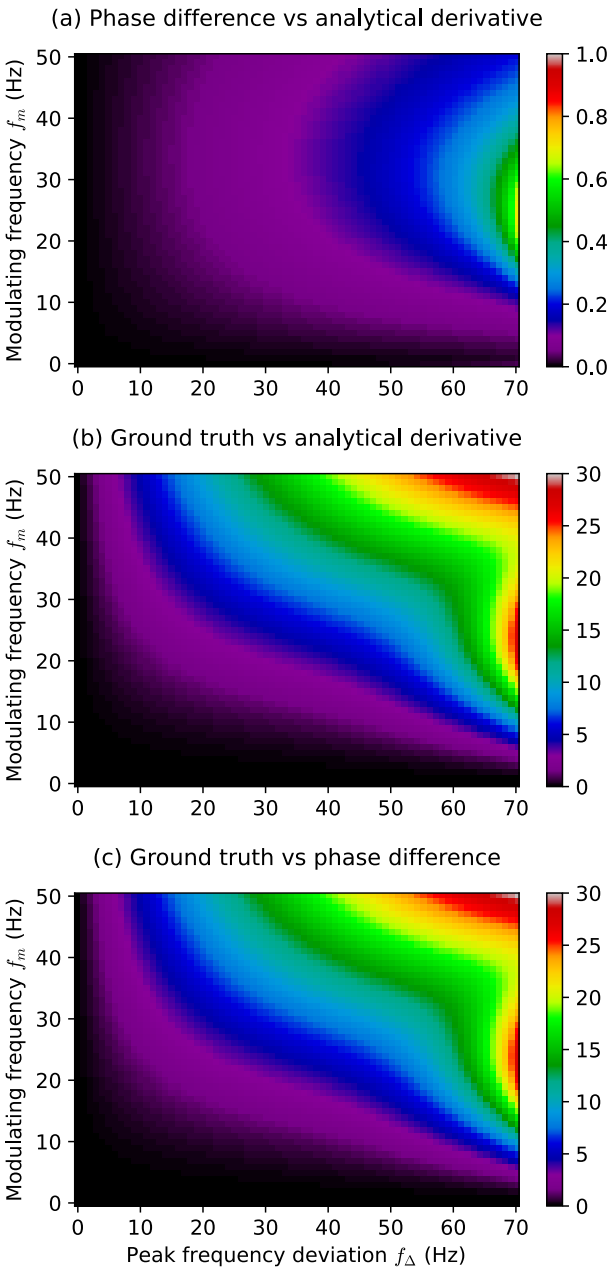

Figure 5.3: Root mean squared error of comparing the ground truth IFD, IFD extracted from the phase difference method, and IFD extracted from the analytical derivative method (proposed): (a) phase difference vs analytical derivative, (b) ground truth vs analytical derivative, and (c) ground truth vs phase difference.

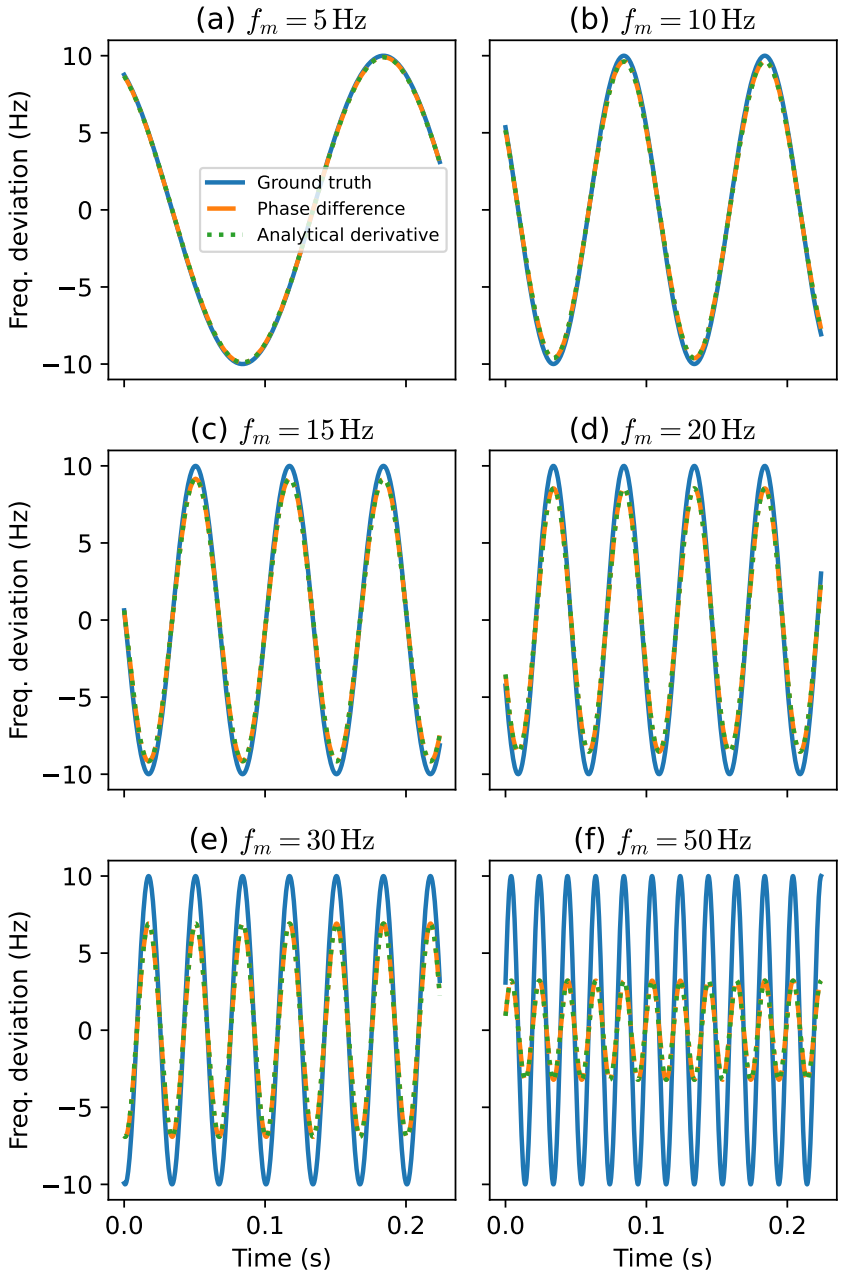

Figure 5.4: Examples of the estimated IFD using phase difference and analytical derivative methods concerning the change of modulating frequency f_m: (a) $f_\mathrm{m} = 5\,\mathrm{Hz}$, (b) $f_\mathrm{m} = 10\,\mathrm{Hz}$, (c) $f_\mathrm{m} = 15\,\mathrm{Hz}$, (d) $f_\mathrm{m} = 20\,\mathrm{Hz}$, (e) $f_\mathrm{m} = 30\,\mathrm{Hz}$, and (f) $f_\mathrm{m} = 50\,\mathrm{Hz}$.

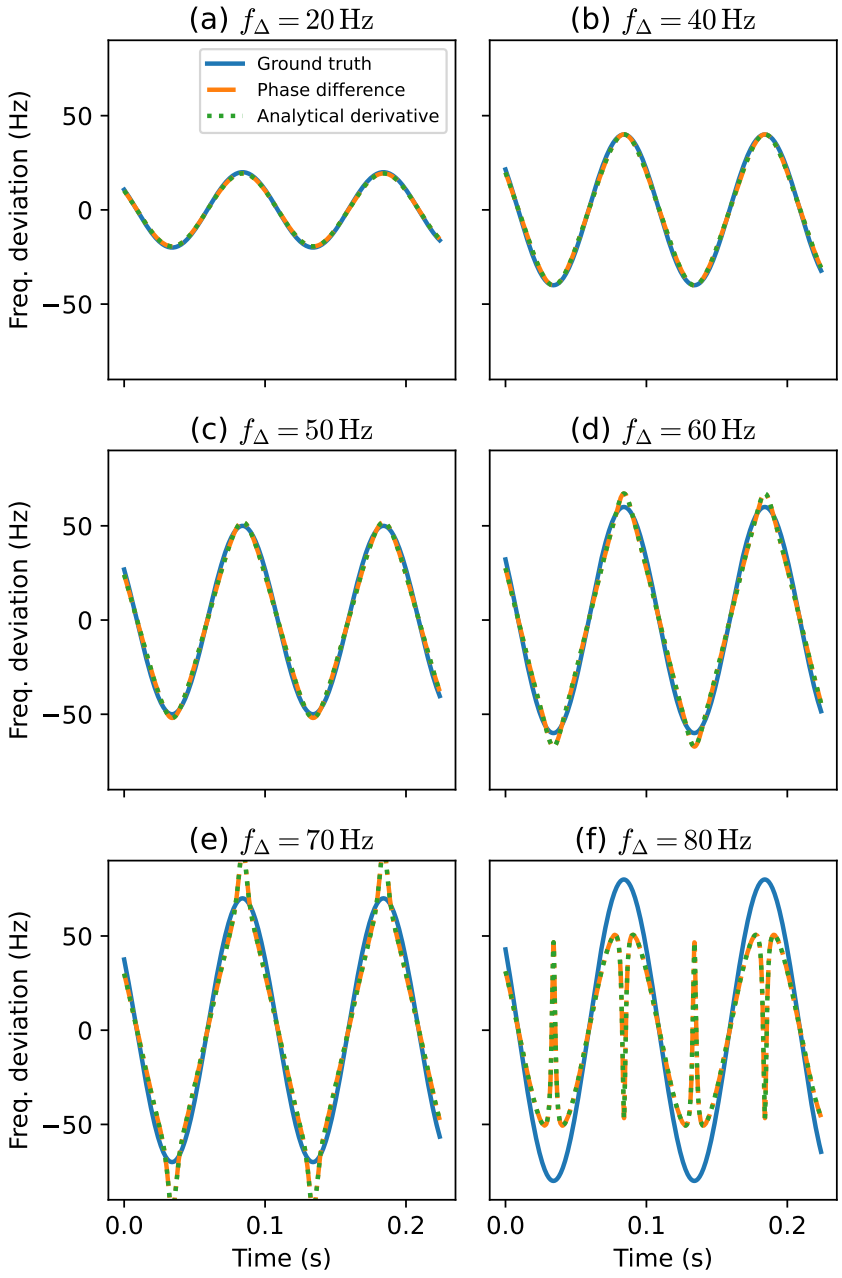

Figure 5.5: Examples of the estimated IFD using phase difference and analytical derivative methods concerning the change of peak frequency deviation f_Δ: (a) $f_\Delta = 20\,\mathrm{Hz}$, (b) $f_\Delta = 40\,\mathrm{Hz}$, (c) $f_\Delta = 50\,\mathrm{Hz}$, (d) $f_\Delta = 60\,\mathrm{Hz}$, (e) $f_\Delta = 70\,\mathrm{Hz}$, and (f) $f_\Delta = 80\,\mathrm{Hz}$.

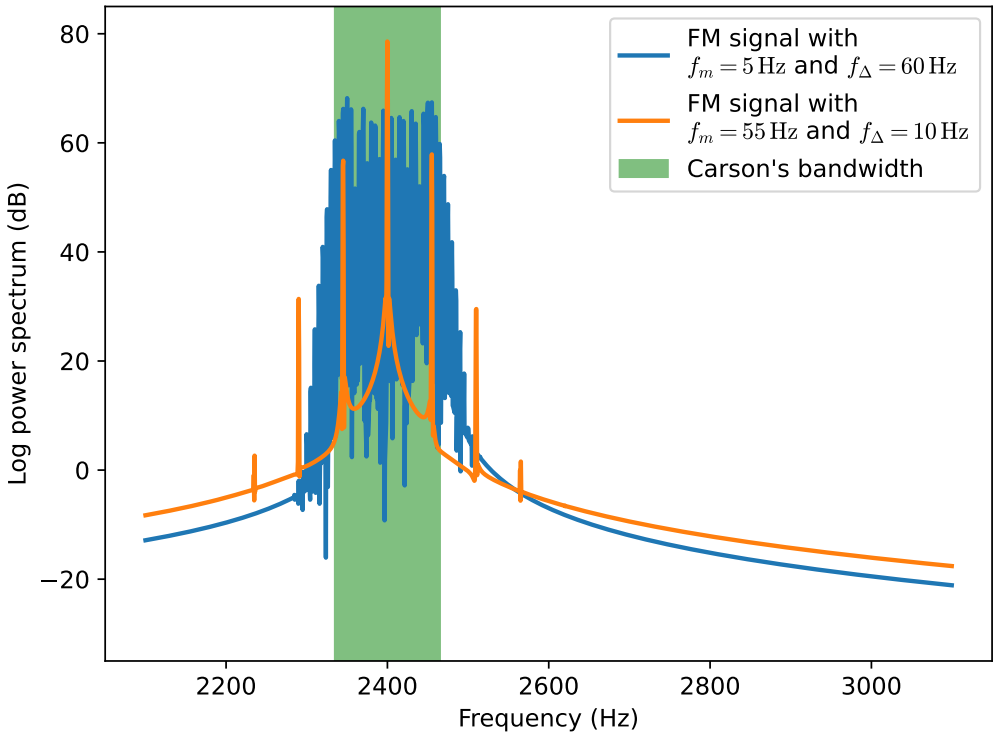

Figure 5.6: Examples of the spectrum of two single-tone FM signals with the same bandwidth but different modulation indices.

are farther away from the center/carrier frequency, causing more leaks to occur (see Figure 5.6).

5.4 Discussion

The proposed analytical derivative method and the conventional phase difference method show equivalent results from the validation experiment results. What are the benefits of using the analytical derivative method?

First, compared to the conventional method, the proposed analytical derivative method does not involve phase calculation and does not require circular wrapping. When computing the complex time-frequency representation phase, only the principal value is calculated, for example, an angle from $-\pi$ to π locating a point in the unit circle on a complex plane. However, when it comes to time-domain signal processing, the principal value and

the trajectory of the signal should be considered. The analytical derivative method overcomes the phase wrapping and provides a rational solution to phase analysis in the complex time-frequency domain.

Second, the Eq. (5.5) can be expanded as

$$Q(\omega,\tau) = \frac{|\breve{x}(\omega,\tau)|}{A(\omega,\tau)} \sin\left(\angle\breve{x}\left(\omega,\tau\right) - \angle\tilde{x}\left(\omega,\tau\right)\right), \tag{5.19}$$

introducing the inverse multiplicative relationship between $A(\omega,\tau)$ and $Q(\omega,\tau)$, which is the main purpose of this chapter. The second research objective is satisfied with this relationship, answering the second research question. Using this connection, Chapter 6 delves into enhancing the IFD of speech in the complex time-frequency representation.

5.5 Summary

In summary, that chapter proposed a novel method to extract instantaneous frequency deviation (IFD), namely the analytical derivative, and established an equation that connected the amplitude to the IFD. Using single-tone frequency-modulated (FM) signals, the proposed method was verified to work correctly. These findings confirmed the proposed equation's validity, satisfying the second research objective and answering the second research question.

Chapter 6

Speech enhancement by enhancing IFD

This chapter investigates the answer to the third research question in this dissertation: 'Is it possible to enhance the speech from the relationship between amplitude and phase.' In complex time-frequency representation, phase processing has always been challenging due to the lack of techniques to process circular data. Previously, Chapter 5 establishes a connection between the amplitude and instantaneous frequency deviation (IFD) - a representation of the phase based on modulation theory, revealing that perhaps employing circular data processing techniques is unnecessary. From such an idea, this chapter proposes a speech enhancement method by enhancing the IFD of each frame in a short-time Fourier transform (STFT) domain via a learnable affine transformation. The evaluation results of Valentini *et al.* [37] show that the proposed method significantly improves speech quality, answering the third research question.

6.1 Problem formulation

The IFD serves as a representation of phase within the complex time-frequency representation, exhibiting analogous patterns to amplitude (see Figure 3.5) and thus holding promise for speech enhancement [70]. Nevertheless, the original definition of IFD in Eq. (5.2) relies entirely on phase, which is susceptible to the wrapping issue. Consequently, IFD encounters the same challenge as

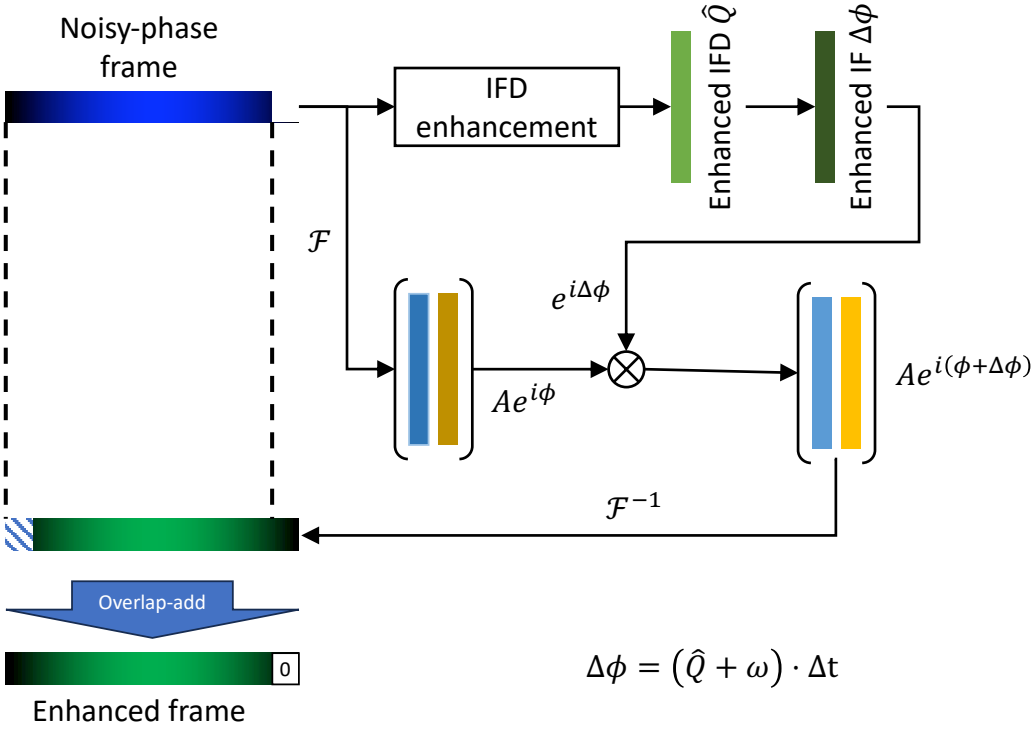

Figure 6.1: Process diagram of frame-wise IFD enhancement.

phase: the lack of a modification technique due to circular data restriction. Another complication arises when modifying IFD, namely the integration problem during the reconstruction of the complex time-frequency representation, which necessitates the phase as per the definition in Eq. (5.1). In summary, speech enhancement using IFD is hindered by two primary issues: modification and integration. The subsequent section outlines a proposed method aimed at addressing these challenges individually.

6.2 Framewise IFD enhancement with learnable affine transform

6.2.1 Frame-wise IFD enhancement: beyond integration

IFD, defined as the deviation of the instantaneous frequency from the center frequency, offers several benefits over analyzing phase variations directly. IFD is less affected by rapid frequency changes than the phase, providing a more intuitive understanding of signal changes over time. However, in speech enhancement, the output is the waveform, which requires reconstructing the complex time-frequency representation where the phase is crucial. From the definition in Eq. (5.1), the phase can be computed from IFD via time integration as follows

$$\phi(\omega, \tau) = \omega\tau + \int_0^\tau Q(\omega, \eta) d\eta \,, \tag{6.1}$$

This integration process contains several issues in practical computation, such as:

- Error accumulation: Any small errors in the estimation or measurement of instantaneous frequency deviation accumulate over time during integration, potentially leading to significant phase errors. The error accumulation is particularly problematic in long-duration signals.
- Computational complexity: Continuous integration requires efficient and precise computational methods, especially in real-time signal processing applications. The need for high-resolution and high-sampling-rate data to minimize errors adds to computational complexity.
- Initial phase uncertainty: Phase integration can sometimes start at some other point τ_0 instead of 0, i.e.,

  $$\phi(\omega, \tau) = \phi(\omega, \tau_0) + \omega\tau + \int_{\tau_0}^\tau Q(\omega, \eta) d\eta \,. \tag{6.2}$$

  While the solution is potential, the initial phase $\phi(\omega, \tau_0)$ at τ_0 must be known beforehand. Error in estimating the initial phase can propagate

through the integration process, affecting the accuracy of the resultant phase.

As a result, integration problems when reconstructing the phase from IFD can yield even more distortion than the noisy phase itself. Therefore, it is necessary to have a better phase reconstruction method to obtain the phase from the IFD so that the distortion is as small as possible.

This section introduces one effective solution for enhancing IFD when the time-frequency representation is obtained by short-time Fourier transform (STFT). In STFT, the resolution is uniform; in other words, the spectral information, including amplitude, phase, and IFD, at a specific time τ belongs to one specific frame in the time domain, and the conversion between time and frequency domain can be obtained using Fourier transform and inverse Fourier transform easily. Therefore, instead of integrating from another frame at τ_0 as in [70] to obtain $\phi(\omega, \tau)$, the proposed method uses the $\phi(\omega, \tau)$ itself as the initial phase and applies the estimated (enhanced) IFD to that frame. The detailed process diagram of this enhancement can be shown in Figure 6.1.

6.2.2 Learnable affine transform for IFD enhancement

Previously, Chapter 5 establishes Eq. (5.5), which describes IFD as a real-valued rational function that involves an inverse multiplicative relationship with the amplitude. On the other hand, most amplitude enhancement can be described using spectral gain functions such as Wiener filter [9], which applies a multiplication operator on the amplitude. Therefore, it is rational to think that the enhanced IFD can be obtained by applying a multiplication operator, i.e., a scaling factor, on the noisy IFD. The following equation can express this operator

$$\hat{Q}_{\text{enhanced}}(\omega, \tau) = Q_{\text{noisy}}(\omega, \tau)\alpha(\omega, \tau)\,, \tag{6.3}$$

where $\alpha(\omega, \tau)$ is the scaling factor.

However, as shown in Eq. (5.19), the IFD contains other factors besides the amplitude, which makes the scaling too ideal. Therefore, this section proposes a simple solution, that is to introduce a shifting factor which makes

the modification a learnable affine transform

$$\hat{Q}_{\text{enhanced}}(\omega, \tau) = Q_{\text{noisy}}(\omega, \tau)\alpha(\omega, \tau) + \gamma(\omega, \tau)\,, \tag{6.4}$$

where $\gamma(\omega, \tau)$ is the shifting factor.

To obtain the optimal $\alpha(\omega, \tau)$ and $\gamma(\omega, \tau)$ that leads to best speech quality, this proposed method uses scale-invariance signal-to-distortion ratio (SISDR) as the optimization criteria, which offers the benefit of being invariant to signal scaling, making it a robust and perceptually relevant metric for audio quality that aligns closely with human auditory perception [116]. The SISDR measures the difference between a ground-truth waveform $\mathbf{s}$ and an estimated waveform $\hat{\mathbf{s}}$ as follows

$$\text{SISDR}\left(\mathbf{s}, \hat{\mathbf{s}}\right) = 10 \log_{10} \frac{\|\mathbf{s}_{\text{target}}\|_2^2}{\|\mathbf{e}_{\text{noise}}\|_2^2}\,, \tag{6.5}$$

$$\mathbf{s}_{\text{target}} = \frac{\hat{\mathbf{s}}^\top \mathbf{s}}{\|\mathbf{s}\|_2^2} \cdot \mathbf{s}\,, \tag{6.6}$$

$$\mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\text{target}}\,. \tag{6.7}$$

6.3 Experiments

6.3.1 Dataset

The dataset used in this experiment is the same as the one in Section 4.4, which is Valentini *et al.* [37] dataset - a benchmark in recent speech enhancement studies. The clean training set includes 28 speakers (14 males and 14 females), and the test set features two speakers (one male and one female) from the Voice Bank corpus [107]. The noisy test set is formed by combining the clean test set with five different noise types from the DEMAND dataset [108], including `bus`, `cafe`, `living`, `office`, and `psquare`, at four SNRs: 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. No common speakers or noise types exist between the training and test sets. All speech waveforms are resampled to a 16 kHz sampling rate, and the input speech is randomly scaled between -35 dB and -20 dB to enhance data variance.

6.3.2 Implementation

In the implementation, the STFT uses the hanning window function with a window length of 25 ms (400 samples) and a hop length of 6.25 ms (100 samples). The number of points for the fast Fourier transform is 512, which results in 257 frequency bins.

The speech is first amplitude enhanced using the proposed method in Chapter 4. Then, the amplitude enhancement is inputted into the IFD enhancement module. The deep neural network to estimate scaling and shifting factors to enhance IFD follows the WaveNet-based module with the LSTM layer (see Figure 4.5), of which all the convolutional layers are complex-valued [78]. The network's input is complex, and the spectrogram and the output are the scaling and shifting factors, which are real-valued. The values of scaling factors are clamped between 0 and 2, while the values of shifting factors are clamped between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. The network is trained for 1000 epochs with an Adam optimizer and a One-cycle Learning Rate scheduler with an initial learning rate of 5×10^{-4} and a maximum learning rate of 2×10^{-4}.

6.3.3 Evaluation Metrics

Similar to Section 4.4, the intrusive metrics Wide-band Perceptual Evaluation of Speech Quality (PESQ-WB) [85, 86] and Short-Time Objective Intelligibility (STOI) [89] metrics are used to evaluate the overall performance of the proposed method. The PESQ-WB scores, which range from -0.5 (bad) to 4.5 (excellent), measure the speech quality by comparing the enhanced signal to the clean reference speech signal. The STOI metric is highly correlated to perceptual speech intelligibility. The STOI scores range between 0 (lowest intelligibility) and 1 (highest intelligibility).

Besides the intrusive metrics, this evaluation also employs DNSMOS P.835 [117] - a non-intrusive objective metric designed to evaluate the perceptual quality of denoised speech, specifically developed for the Deep Noise Suppression (DNS) Challenge [118]. The metric follows ITU-T Recommendation P.835 [83, text], which provides a standardized approach for separately assessing three aspects of speech quality via three scores in the MOS scale: signal distortion (SIG), background intrusiveness (BAK), and overall quality

Table 6.1: Results of proposed and state-of-the-art methods trained on Valentini *et al.* dataset.

Method	PESQ-WB	STOI	DNSMOS P.835		
			SIG	BAK	OVRL
Noisy	1.97	0.91	3.33	3.12	2.69
SEGAN [20]	2.16	0.93	–	–	–
MMSE-GAN [119]	2.53	0.93	–	–	–
Wave U-Net [120]	2.40	–	–	–	–
MetricGAN [22]	2.86	0.92	–	–	–
DCT-UNet [121]	2.70	–	–	–	–
μ-law SGAN [122]	2.86	0.94	–	–	–
DCCRN [73]	2.68	–	–	–	–
DCCRN+ [76]	2.84	–	–	–	–
Proposed method (noisy IFD)	2.82	0.94	3.41	3.97	3.09
Proposed method (enhanced IFD)	**2.87**	**0.94**	**3.44**	**3.99**	**3.13**

(OVRL). DNSMOS P.835 employs a neural network-based model to predict these scores from denoised audio samples, offering an efficient and scalable alternative to subjective listening tests. By providing reliable and consistent evaluations, DNSMOS P.835 facilitates developing and benchmarking noise suppression algorithms, ensuring that improvements are aligned with human auditory perception.

6.3.4 Results

To ensure a fair comparison, only the performance of our proposed model against other baseline models was trained using the same Valentini *et al.* dataset [37]. From the results reported in Table 6.1, the proposed method outperforms many baselines and produces comparative results against other strong baselines, notably the state-of-the-art Deep Complex Convolution Recurrent Network (DCCRN) method [73] from the 2020 Deep Noise Suppression Challenge (DNS2020). Also, the proposed method improves the DCCRN+ model in Interspeech 2021 [76].

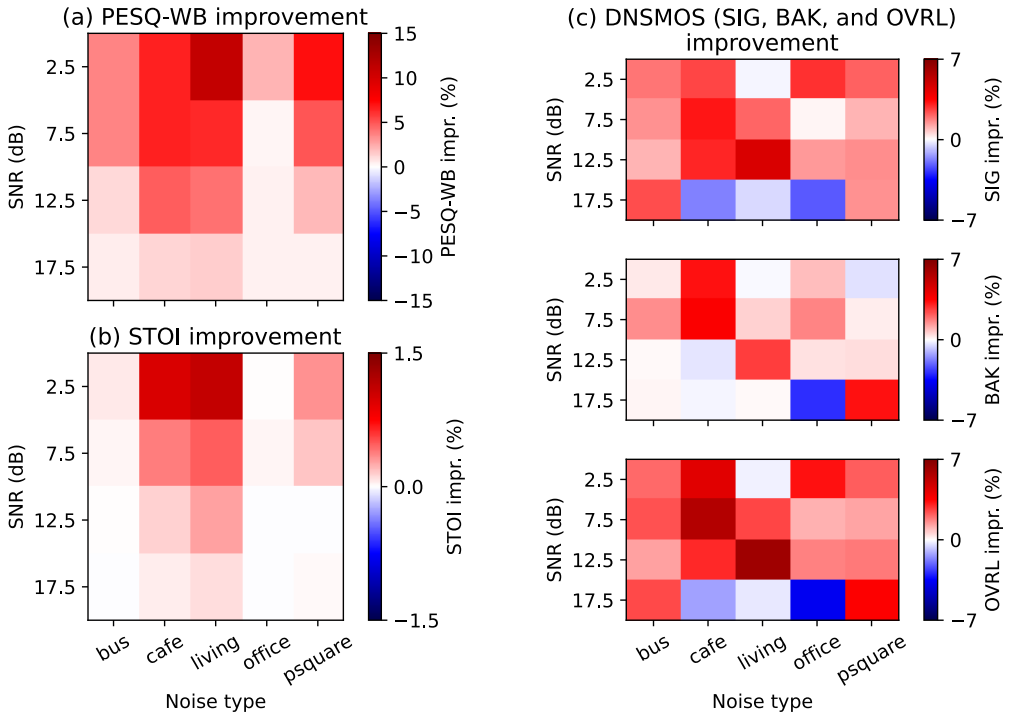

Figure 6.2: Effect of in IFD enhancement module in model performance on Valentini *et al.*'s test set under different metrics including (a) PESQ-WB, (b) STOI, and (c) DNSMOS (a composite of three values: SIG, BAK, and OVRL). SNRs and noise types aggregate the scores.

6.3.5 The effectiveness of IFD enhancement

To elucidate the effectiveness of IFD enhancement, the evaluation metrics are measured with and without applying the IFD enhancement module, and the improvement percentage is computed for each metric using Eq. (4.41). This evaluation is grouped by noise types and SNRs in the Valentini *et al.* test dataset [37]. The evaluation results are visualized in Figure 6.2. The results show that, on average, incorporating phase correction always improves PESQ-WB and STOI, and the effectiveness is more significant in low SNR conditions, improving the PESQ-WB to 15% and STOI to 1.5%. Also, the SIG, BAK, and OVRL improve in most cases and sometimes reduce in high SNR conditions. In low SNR conditions, the performance improves the most in `living` and `cafe` noise types. Most noise signals in the `living` type contain music or singing voices, while noise signals in the `cafe` types are

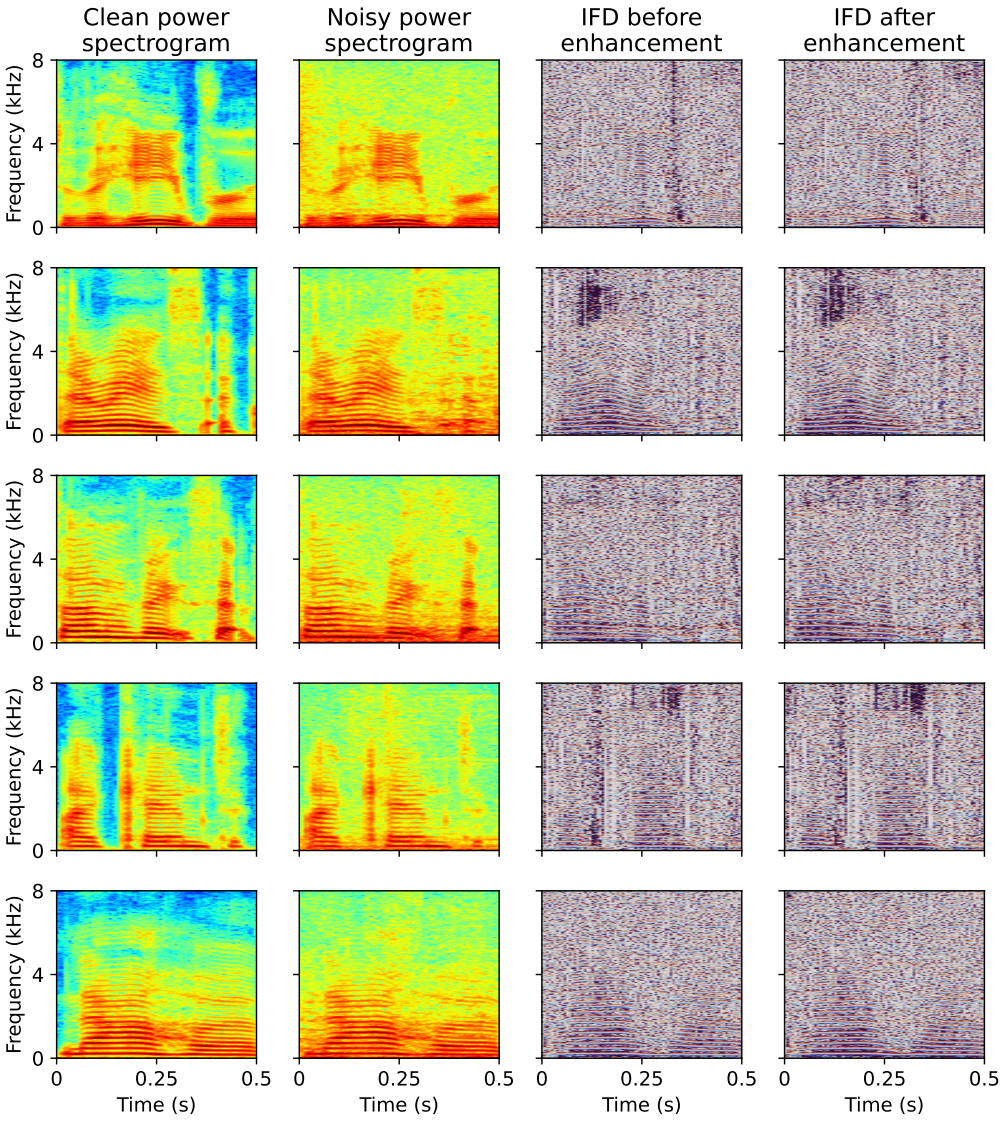

Figure 6.3: Samples of the effect of the IFD enhancement module on five different samples (five rows) from the Valentini test set. The columns from left to right are the clean-speech power spectrogram, noisy-speech power spectrogram, IFD before enhancement, and IFD after enhancement. In the enhanced IFD, the areas between harmonics are emphasized, resulting in the suppression of speech power in these areas.

babble noise recorded in the crowded environment [37]. Such types of noise have formants in their spectra, where the power is concentrated at some frequencies. In contrast, the spectra of `bus` and `office` noises spread broader along the frequency axis, distributing the SNR uniformly in all frequency bins. In other words, the IFD enhancement effectively suppresses noises in which power spectral density is concentrated. To support this hypothesis, Figure 6.3 illustrates the effect of IFD enhancement on some samples, showing that in the enhanced IFD, the areas between harmonics are emphasized, resulting in suppression of speech power in these areas.

6.4 Summary

In summary, using the relation found in Chapter 5, the chapter proposed a speech enhancement method by enhancing the instantaneous frequency deviation (IFD) that introduced the solution for the two problems when using IFD for speech enhancement: how to modify and how to integrate. Specifically, a learnable affine transform was proposed for modification, and frame-wise IFD enhancement was proposed for integration. Using different speech quality and intelligibility metrics, including PESQ-WB, STOI, and DNSMOS P.835, the results showed that enhancing IFD using the proposed method significantly improved the speech quality of noisy-phase speech. These results satisfied the third objective of this research, supporting the answer to the third research question: utilizing the relationship between amplitude and phase via IFD could significantly enhance speech quality.

Chapter 7

Conclusion

7.1 Summary

The main objective of this research is to investigate the effectiveness of utilizing modulation characteristics of speech for enhancement, which contains three sub-objectives:

1. *Model the amplitude modulation characteristics for speech enhancement*

   In Chapter 4, a method to model the spectral modulation characteristics of speech in amplitude using the categorical distribution of fundamental frequency is proposed and applied for speech enhancement. The results show that the improvement in amplitude modulation characteristics leads to an improvement in speech enhancement performance. These results answer the first research question: enhancing the speech characteristics in amplitude improves speech enhancement performance.

2. *Derive the relationship between amplitude and the instantaneous frequency modulation*

   In Chapter 5, a method to extract instantaneous frequency deviation (IFD) is proposed, namely analytical derivative, and an equation connecting the amplitude to the IFD is established. Using single-tone frequency-modulated signals, the proposed method is verified to work correctly, which confirms the proposed equation's validity. These results answer the second research question: there is a relationship be-

tween the amplitude and phase via the IFD, where the IFD is related to the time differentiation of the phase and has an inverse multiplicative relationship with the amplitude.

3. *Enhance speech using the derived relationship*

   In Chapter 6, based on the relationship found in Chapter 5, a method to enhance speech via IFD is proposed to modify IFD by a learnable affine transform at frame-wise level. The results show that the proposed method improves speech enhancement, especially quality. These results support the answer to the third research question: the relationship between amplitude and phase helps improve speech enhancement performance.

All the results confirm that utilizing speech's modulation characteristics can improve speech enhancement performance, satisfying the research objective.

7.2 Contributions

The research on utilizing modulation characteristics for speech enhancement has significant practical applications. The improved speech quality resulting from this research can enhance user experience in mobile calls, VoIP services, and video conferencing by clarifying conversations and reducing misunderstandings caused by background noise. In assistive technologies, hearing aids and cochlear implants can benefit from these advanced enhancement techniques, providing users with a better auditory experience in noisy environments. Additionally, voice-activated systems, such as virtual assistants and automated customer service, can achieve higher accuracy and reliability by integrating these speech enhancement methods, leading to more effective and user-friendly interactions.

This research also contributes to advancements in various other fields. In audio signal processing, insights into modulation characteristics and the relationship between amplitude and instantaneous frequency modulation can inspire new noise reduction and audio signal manipulation methods. In machine learning and artificial intelligence, the proposed techniques and models can be adapted to improve speech recognition systems, making them more robust against noisy inputs. Furthermore, in neuroscience, understanding

how modulation characteristics affect speech perception can aid in developing better auditory models, contributing to the study of human hearing and cognitive processing of sounds. This research thus provides a foundational framework that can be built upon in multiple scientific and engineering domains.

7.3 Remaining works

This study only focuses on enhancing speech under additive noise by utilizing the modulation concept to incorporate the knowledge of speech into the speech enhancement model. However, several remaining areas of work can further expand and deepen the impact of this research. One significant area for future research is the exploration of speech enhancement techniques in non-additive noise environments, such as reverberant or highly dynamic acoustic settings. While additive noise is common, real-world scenarios often involve more complex noise types that interact with speech non-linearly. Extending the modulation-based approach to handle these complex noise conditions would enhance the robustness and applicability of the proposed methods.

Another critical step is developing real-time processing capabilities and implementing the proposed methods in practical devices. Achieving low-latency, high-efficiency speech enhancement suitable for real-time applications such as live communications, hearing aids, and smart devices presents technical and engineering challenges. Future research could optimize the computational aspects of the proposed methods to ensure they are feasible for real-time use on embedded systems and mobile platforms.

Further research is needed to ensure the proposed speech enhancement techniques are robust across different languages and speaker variations. Speech characteristics can vary widely among different languages and individual speakers, affecting the effectiveness of enhancement methods. Extensive testing and adaptation of the modulation-based approaches to accommodate diverse linguistic and phonetic contexts would be essential for broadening the applicability of these techniques.

While objective metrics are crucial, subjective listening tests are equally important to validate the perceived improvements in speech quality. Future

work should include comprehensive subjective evaluations involving diverse listener groups to assess the practical benefits of the enhancement methods. Such evaluations can provide insights into user preferences and highlight areas for further refinement.

In conclusion, while this study has laid a strong foundation by demonstrating the effectiveness of utilizing modulation characteristics for speech enhancement under additive noise conditions, there are numerous opportunities for expanding this work. Future research should explore complex noise environments, optimize for real-time applications, ensure cross-linguistic robustness, and validate with subjective tests. These directions hold the potential to advance the field of speech enhancement further, making the techniques more versatile and impactful across various real-world applications.

Publications

Journal papers

[1] H. Nguyen, T. V. Ho, M. Akagi and M. Unoki, "Phase-Aware Speech Enhancement With Complex Wiener Filter," *IEEE Access*, vol. 11, pp. 141573–141584, 2023.

[2] H. Nguyen and M. Unoki, "Improvement in Bone-Conducted Speech Restoration Using Linear Prediction and Long Short-Term Memory Model," *Journal of Signal Processing*, vol. 24, no. 4, pp. 175–178, 2020.

International conference papers

[1] H. Nguyen, M. L. Nguyen, M. Unoki, "Speaker Verification Using Distance based on Principal Component Analysis for Household Scenario Adaptation," in *Proc. RIVF*, pp. 441–446 Hanoi, 2023.

[2] H. Nguyen, K. Li, and M. Unoki, "Automatic Mean Opinion Score Estimation with Temporal Modulation Features on Gammatone Filterbank for Speech Assessment," in *Proc. Interspeech*, Korea, 2022.

[3] T. V. Ho, H. Nguyen, M. Akagi, and M. Unoki, "Vector-quantized Variational Autoencoder for Phase-aware Speech Enhancement," in *Proc. Interspeech*, pp. 176–180, Korea, 2022.

[4] H. Nguyen and M. Unoki, "Bone-conducted Speech Enhancement Using Vector-quantized Variational Autoencoder and Gammachirp Filterbank Cepstral Coefficients," in *Proc. EUSIPCO*, pp. 21–25, Serbia, 2022.

[5] K. Li, H. Nguyen, Masashi Unoki, "Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Using Temporal Modulation Features on Gammatone Filterbank," in *Proc. DCASE*, France, 2022.

Domestic conference papers

[1] H. Nguyen and M. Unoki, "Study on Bone-conducted Speech Enhancement Using Vector-quantized Variational Autoencoder and Gammachirp Filterbank Cepstral Coefficients," in *IEICE Technical Report*, pp. 109–114, Okinawa, 2022.

Bibliography

[1] W. Jiang, Z. Liu, K. Yu, and F. Wen, "Speech enhancement with neural homomorphic synthesis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 376–380, 2022.

[2] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," tech. rep., Nicolet Scientific Corp., Northvale, NJ., 1974.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[4] J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins, "A mixed-source model for speech compression and synthesis," *The Journal of the Acoustical Society of America*, vol. 64, no. 6, pp. 1577–1581, 1978.

[5] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[6] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*, p. paper 1134, 1998.

[7] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. IV–4164–IV–4164, 2002.

[8] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[9] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[11] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[13] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–253–I–256, 2002.

[14] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[15] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[16] S. Jensen, P. Hansen, S. Hansen, and J. Sorensen, "Reduction of broadband noise in speech by truncated qsvd," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 439–448, 1995.

[17] B. De Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Transactions on Signal Processing*, vol. 41, no. 9, pp. 2826–2838, 1993.

[18] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[20] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proceedings of Interspeech*, 2017.

[21] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, 2019.

[22] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 2031–2041, PMLR, 2019.

[23] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 659–666, 2021.

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

[25] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*, pp. 181–197. Springer US, 2005.

[26] C. Hummersone, T. Stokes, and T. Brookes, *On the Ideal Ratio Mask as the Goal of Computational Auditory Scene Analysis*, pp. 349–368. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[27] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech 2013*, pp. 436–440, 2013.

[28] C. Jitong and D. Wang, *DNN Based Mask Estimation for Supervised Speech Separation*, pp. 207–235. Springer International Publishing, 2018.

[29] G. Fant, "Acoustic theory of speech production," *Mouton*, 1960.

[30] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, vol. 12. Springer Berlin Heidelberg, 1976.

[31] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[32] M. M. Sondhi, J. B. Allen, and L. R. Rabiner, "Efficient processing of spectral envelopes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 330–338, 1981.

[33] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 598–608, 1980.

[34] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[35] R. Plomp, "The role of modulation in hearing," in *HEARING — Physiological Bases and Psychophysics*, (Berlin, Heidelberg), pp. 270–276, Springer Berlin Heidelberg, 1983.

[36] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[37] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. Interspeech 2016*, pp. 352–356, 2016.

[38] P. C. Loizou, *Speech Enhancement*. CRC Press, 2007.

[39] P. Hansen, P. Hansen, S. Hansen, and J. Sørensen, "Ulv-based signal subspace method for speech enhancement," in *International Workshop on Acoustic Echo and Noise Control, IWAENC'97*, pp. 9–12, Imperial Collega, 1997. nternational Workshop on Acoustic Echo and Noise Control, IWAENC ; Conference date: 01-01-1997.

[40] P. Hansen, *Signal subspace methods for speech enhancement*. PhD thesis, Technical University of Denmark, 1998.

[41] P. S. K. Hansen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Experimental comparison of signal subspace based noise reduction methods," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference - Volume 01*, ICASSP '99, (USA), p. 101–104, IEEE Computer Society, 1999.

[42] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 104–106, 2003.

[43] C. H. You, S. N. Koh, and S. Rahardja, "An invertible frequency eigendomain transformation for masking-based subspace speech enhancement," *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 461–464, 2005.

[44] O. M. Ssenwanger, *Elements of Statistical Analysis*. Elsevier, 1986.

[45] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 270–273, 2004.

[46] D. Malah and R. Cox, "A generalized comb filtering technique for speech enhancement," in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 160–163, 1982.

[47] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1559–1569, 2018.

[48] Y. Wakabayashi, "Speech enhancement using harmonic-structure-based phase reconstruction," *Acoustical Science and Technology*, vol. 40, no. 3, pp. 162–169, 2019.

[49] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, and L. E. White, *Neuroscience*. Sinauer Associates, 3rd ed., 2004.

[50] L. Robles and N. P. Cooper, "Electrically evoked cochlear microphonic potentials in the guinea pig," *Science*, vol. 187, no. 4179, pp. 363–365, 1975.

[51] H. J. von Beckinghausen, "Electrical responses of cochlear hair cells to vibratory stimulation," *Journal of the Acoustical Society of America*, vol. 23, no. 1, pp. 71–75, 1951.

[52] R. D. Patterson, I. Nimmo-Smith, D. J. Holdsworth, and M. P. Rice, "An efficient auditory filterbank based on the gammatone function," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 442–455, 1987.

[53] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2222–2232, 2006.

[54] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrowband carriers," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[55] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2906–2919, 1997.

[56] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 140–147, 1997.

[57] J. G. Lyons and K. K. Paliwal, "Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement," in *Proc. Interspeech 2008*, pp. 387–390, 2008.

[58] X. Lu, M. Unoki, and M. Akagi, "Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," *Acoustical Science and Technology*, vol. 29, no. 6, pp. 351–361, 2008.

[59] T. Ngo, R. Kubo, and M. Akagi, "Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function," *Speech Communication*, vol. 135, pp. 11–24, 2021.

[60] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, no. 3, pp. 261–279, 1999.

[61] N. Nower, Y. Liu, and M. Unoki, "Restoration scheme of instantaneous amplitude and phase using kalman filter with efficient linear prediction for speech enhancement," *Speech Communication*, vol. 70, pp. 13–27, 6 2015.

[62] Y. Liu, N. Nower, S. Morita, and M. Unoki, "Speech enhancement of instantaneous amplitude and phase for applications in noisy reverberant environments," *Speech Communication*, vol. 84, pp. 1–14, 2016.

[63] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings*, (Berlin, Heidelberg), p. 79–86, Springer-Verlag, 2020.

[64] P. Roach, R. Stibbard, J. Osborne, S. Arnfield, and J. Setter, "Transcription of prosodic and paralinguistic features of emotional speech," *Journal of the International Phonetic Association*, vol. 28, pp. 83–94, 1998.

[65] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 716–720, Institute of Electrical and Electronics Engineers Inc., 2018.

[66] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2018.

[67] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 676–680, 2021.

[68] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2014.

[69] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, pp. 465–494, 2011.

[70] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 63–76, 2019.

[71] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising," in *Proc. Interspeech 2020*, pp. 2467–2471, 2020.

[72] X. Li and R. Horaud, "Online Monaural Speech Enhancement Using Delayed Subband LSTM," in *Proc. Interspeech 2020*, pp. 2462–2466, 2020.

[73] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

[74] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6633–6637, 2021.

[75] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech 2020*, pp. 2477–2481, 2020.

[76] S. Lv, Y. Hu, S. Zhang, and L. Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," in *Proc. Interspeech 2021*, pp. 2816–2820, 2021.

[77] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, pp. 483–492, 2015.

[78] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *International Conference on Learning Representations*, 2018.

[79] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[80] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "Speech enhancement with phase sensitive mask estimation using a novel hybrid neural network," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 136–150, 2021.

[81] ITU-T, "P.800: Methods for subjective determination of transmission quality," tech. rep., International Telecommunication Union, 1996.

[82] ITU-T, "P.807: Subjective test methodology for assessing speech intelligibility," tech. rep., International Telecommunication Union, 2016.

[83] ITU-T, "P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," tech. rep., International Telecommunication Union, 2003.

[84] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, pp. 125–128 vol.1, 1993.

[85] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "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, pp. 749–752 vol.2, 2001.

[86] ITU-T, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," tech. rep., International Telecommunication Union, 2001. The Recommendation was out of date and deleted on 5 January 2024. Please refer to P.863 and its subsequent amendments.

[87] Y. Gaoxiong and Z. Wei, "The perceptual objective listening quality assessment algorithm in telecommunication: Introduction of itu-t new metrics polqa," in *2012 1st IEEE International Conference on Communications in China (ICCC)*, pp. 351–355, 2012.

[88] I. T. Union, "Perceptual objective listening quality analysis," Tech. Rep. P.863, ITU-T, 2011.

[89] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy

speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, 2010.

[90] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[91] C. Févotte, R. Gribonval, and E. Vincent, "Bss eval toolbox user guide - revision 2.0," tech. rep., IRISA, 2005.

[92] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[93] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "Visqol: The virtual speech quality objective listener," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pp. 1–4, 2012.

[94] N. Minematsu, "A modulation-demodulation model of speech communication," in *Proc. Speech Prosody 2010*, p. paper 913, 2010.

[95] K. Honda, *Physiological Processes of Speech Production*, pp. 7–26. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[96] W. A. Yost, *Fundamentals of Hearing: An Introduction*. Brill, 2013.

[97] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speech—a syllable-centric perspective," *Journal of Phonetics*, vol. 31, no. 3, pp. 465–485, 2003. Temporal Integration in the Perception of Speech.

[98] P. Maragos, T. Quatieri, and J. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pp. 421–424 vol.1, 1991.

[99] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, pp. 3795–3806, 06 1996.

[100] A. S. Leonov, I. S. Makarov, and V. N. Sorokin, "Frequency modulations in the speech signal," *Acoustical Physics*, vol. 55, 2009.

[101] A. Tabas and K. von Kriegstein, "Neural modelling of the encoding of fast frequency modulation," *PLOS Computational Biology*, vol. 17, pp. 1–30, 03 2021.

[102] R. Gransier, S. Peeters, and J. Wouters, "The importance of temporal-fine structure to perceive time-compressed speech with and without the restoration of the syllabic rhythm," *Scientific Reports*, vol. 13, 2023.

[103] A. P. Stark and K. K. Paliwal, "Speech analysis using instantaneous frequency deviation," in *Proc. Interspeech 2008*, pp. 2602–2605, 2008.

[104] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "Hgcn: Harmonic gated compensation network for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 371–375, 2022.

[105] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017.

[106] J. D. Markel and A. H. G. Jr., "Linear prediction of speech," *Springer-Verlag*, 1972.

[107] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp. 1–4, 2013.

[108] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *Proceedings of Meetings on Acoustics*, vol. 19, p. 35081, 2013.

[109] T. V. Ho and M. Akagi, "Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder," in

Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, pp. 140–144, 2020.

[110] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, p. 14802, 2021.

[111] A. Lancucki, J. Chorowski, G. Sanchez, R. Marxer, N. Chen, H. J. G. A. Dolfing, S. Khurana, T. Alumae, and A. Laurent, "Robust training of vector quantized bottleneck models," *Proceedings of the International Joint Conference on Neural Networks*, 2020.

[112] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6309–6318, 2017.

[113] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, 2008.

[114] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & systems (2nd ed.)*. USA: Prentice-Hall, Inc., 1996.

[115] J. Carson, "Notes on the theory of modulation," *Proceedings of the Institute of Radio Engineers*, vol. 10, pp. 57–64, 1922.

[116] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.

[117] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890, 2022.

[118] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9271–9275, 2022.

[119] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039–5043, 2018.

[120] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *CoRR*, vol. abs/1806.03185, 2018.

[121] C. Geng and L. Wang, "End-to-end speech enhancement based on discrete cosine transform," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 379–383, 2020.

[122] H. Li, Y. Xu, D. Ke, and K. Su, "μ-law sgan for generating spectra with more details in speech enhancement," *Neural Networks*, vol. 136, pp. 17–27, 2021.