Doctoral Dissertation

Short-time Fourier Transform Phase Reconstruction Using Deep Neural Networks and Phase-based Features

September 2023

Doctoral Program in Advanced Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

NGUYEN Binh Thien

Doctoral Dissertation Reviewed
by Ritsumeikan University

Short-time Fourier Transform Phase Reconstruction Using Deep Neural Networks and Phase-based Features

(深層ニューラルネットワークと位相ベース特徴を用いた短時間フーリエ変換位相再構成)

September 2023

2023 年 9 月

Doctoral Program in Advanced Information Science and Engineering

Graduate School of Information Science and Engineering

Ritsumeikan University

立命館大学大学院情報理工学研究科

情報理工学専攻博士課程後期課程

NGUYEN Binh Thien

グエン ビン ティエン

Supervisor : Professor NISHIURA Takanobu

研究指導教員： 西浦 敬信 教授

Dissertation submitted to Graduate School of Information Science and Engineering,
Ritsumeikan University
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Thesis Committee: Takanobu Nishiura
Yoichi Yamashita
Gang Xu

Abstract

Short-time Fourier transform (STFT) is one of the most powerful techniques for speech signal processing. It decomposes a signal into a complex spectrogram, which can be further separated into the amplitude and phase spectrograms. In contrast to the amplitude spectrogram that directly exhibits a meaningful structure, the phase spectrogram only displays a fuzzy pattern due to the wrapping issue. This leads to difficulty in interpreting or extracting useful information from the phase spectrum. However, numerous studies have highlighted the significance of the phase in signal processing, particularly its role in generating high-quality time-domain signals. Motivated by this, the objective of this thesis is to address two key aspects of phase processing: exploring the hidden information within the phase and developing techniques for phase reconstruction.

Extracting useful information from the phase becomes problematic due to the wrapping issue, which obscures its underlying structure. To tackle this issue, one solution is to transform the phase into alternative representations. This thesis studies several phase-based features, including instantaneous frequency (IF), group delay (GD), IF deviation, relative phase shift, and phase distortion, which have been conventionally used. Additionally, two novel phase-based features, namely derivative of IF and inter-frequency phase difference (IFPD), are introduced and their properties are investigated.

The phase reconstruction is challenging due to the phase sensitivity to the waveform shift and wrapping issue. To mitigate these problems, two-stage approaches indirectly estimate the phase through phase derivatives, i.e., IF and GD. In the first stage, the IF and GD are estimated from the amplitude using deep neural networks (DNNs), and then in the second stage, the phase is reconstructed by maintaining the IF/GD information. Conventional methods for the second stage do not consider the importance of high-amplitude time–frequency bins, e.g., the least squares-based method, or lack a solid model, e.g., the average-based method. To address these

limitations, this thesis proposes improvements to the second stage of two-stage algorithms by using the *von Mises* distribution-based maximum likelihood and weighted least squares. This thesis also provides theoretical discussions for the phase reconstruction, including investigations of the properties of the GD and roles of the IF/GD information in the inverse STFT. Subjective and objective experiments are conducted to compare the performances of our proposed and conventional methods. The results confirm that the proposed method using the IFPD performs better than other methods for all metrics.

Many phase reconstruction algorithms, including the two-stage algorithms, require iteration of future frame information to estimate the current-frame phase, which may only be feasible offline. In various applications, such as incremental text-to-speech, there is a demand for online phase reconstruction. Among the phase reconstruction approaches, the DNN-based methods show promise as they can be easily adapted for online applications by using a causal model. However, conventional DNN-based methods do not take into account the distinct properties of the phase at different time–frequency bins, which may lead to limitations in training the DNNs. To address this, this thesis proposes loss functions for phase reconstruction that incorporate frequency-specific and amplitude weights to distinguish the importance of phase elements based on their properties. The IFPD is also used to improve the phase connections along the frequency. To improve the generalization, this thesis augments the data by randomly shifting the signals in the time domain for each epoch during training. Experimental results show the superior performance of the proposed methods compared to conventional DNN-based and non-DNN online phase reconstruction methods.

Acknowledgment

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Takanobu Nishiura. The completion of this Ph.D. would not have been possible without his assistance and dedicated involvement throughout the process. I would like to thank Prof. Yoichi Yamashita and Prof. Gang Xu for acting as my thesis committee.

I would like to say a special thank you to Asst. Prof. Yukoh Wakabayashi. His invaluable feedback and encouragement greatly influenced how I conducted my research, and the results presented in this thesis would be impossible without his guidance. I would like to extend my sincere appreciation and gratitude to Lecturer Kenta Iwai and Asst. Prof. Yuting Geng for their exceptional contributions to my research. Their active involvement in the research meetings and insightful comments have greatly influenced and improved the quality of this work. I am grateful to all my lab members for their kind help and support that have made my study and life in Japan a wonderful time. I also greatly appreciate the time and effort they spent participating in my experiments.

Finally, I would like to express my special thanks to the Japanese Government (MEXT) Scholarship for giving me the opportunity to study doctoral degree in Japan. I would not have been able to pursue my study without generous supports from the scholarship.

Contents

List of Figures

List of Tables

List of Denotations

Abbreviations and Acronyms

DFT	Discrete Fourier transform
DIF	Derivative of instantaneous frequency
GD	Group delay
IF	Instantaneous frequency
IFD	Instantaneous frequency deviation
IFPD	Inter-frequency phase difference
ISTFT	Inverse short-time Fourier transform
LS	Least squares
ML	Maximum likelihood
MLC	Maximum likelihood using coordinate descent (two-stage phase reconstruction algorithm)
MLN	Maximum likelihood using Newton's method (two-stage phase reconstruction algorithm)
PD	Phase distortion
PESQ	Perceptual evaluation of speech quality
PGHI	phase gradient heap integration
RPS	Relative phase shift
RTISI	real-time spectrogram inversion algorithm
RTPGHI	real-time phase gradient heap integration
SPSI	single-pass spectrogram inversion
STFT	Short-time Fourier transform
STOI	Short-time objective intelligibility
TF	Time–frequency

Chapter 1

Introduction

1.1 Background

Short-time Fourier transform (STFT) is one of the most powerful techniques for speech signal processing. The Fourier transform decomposes a signal into sinusoids, producing complex-valued spectral coefficients represented in terms of their amplitudes and phases. The term "short-time" means we divide a longer time signal into shorter segments and then compute the Fourier transform on each segment. In contrast to the amplitude spectrogram that directly exhibits a meaningful structure, the phase spectrogram, only displays a fuzzy pattern. That is due to the phase wrapping issue that the phase value is limited in its principal values. This leads to difficulty in interpreting or extracting useful information from the phase spectrum. Furthermore, previous studies also demonstrated the unimportance of the phase. Helmholtz [1] concluded from the experiments that human ears are insensitive to the phase. Wang and Lim [2] reported that improving the accuracy of the spectral phase is not critical for the speech enhancement if the phase estimate is used to reconstruct speech by combining it with an independently estimated magnitude or to reconstruct speech using the phase-only signal reconstruction algorithm. Vary [3] argued that “If the magnitude is estimated correctly with a sufficient spectral resolution, no speech degradation is to be perceived as long as the local signal-to-noise ratios (SNRs) are at least about 6 dB." As a consequence, the amplitude spectrum has been studied extensively while not many efforts were dedicated to the phase spectrum for a long time in the past.

However, there were also studies opposing these assumptions about the uselessness of the

phase. Bilsen [4] showed that the phase changes influence the pitch perceptibility. Plomp and Steeneken [5] presented the effect of the phase on the timbre of complex tones. Paliwal *et al.* demonstrated the usefulness of the phase spectrum in speech signal processing [6] and human speech perception [7]. Recently, phase processing, especially phase reconstruction, has gained considerable attention [8–17] in the short-time Fourier transform (STFT)-based audio processing area. The information extracted from the phase can be combined with the amplitude information as inputs for the DNN models to improve their performances [18–21]. As a target to estimate, the phase reconstructed using the amplitude and observed noisy/mixed phase has been demonstrated to be useful in many applications including source separation [22–24] and speech enhancement [25–27]. In other contexts, when the amplitude spectrogram is artificially constructed (e.g., time-scale modification [28], speech synthesis [29–31], and audio restoration [32]), the observed phase does not exist, and the phase reconstruction has to be done using only the amplitude information.

Most former phase reconstruction approaches rely on the consistency property of the STFT, which originates from the redundancy of the information caused by the overlap of analysis windows. The approach proposed by Griffin and Lim [28] is the most well-known, which iteratively updates the phase estimate using the STFT and inverse STFT (ISTFT) while holding the amplitude information. Alternatively, [33] explicitly defines an inconsistency criterion and minimizes it with simplifications. Although yielding relatively good results, consistency-based approaches have several drawbacks; the whole amplitude spectrogram is required for each iteration, the convergence can be slow, and the reconstructed signals may contain artifacts such as echo or reverberation. Other phase reconstruction approaches based on signal modeling, including harmonic modeling, have been reported to achieve higher performance with a lower complexity in comparison with consistency-based approaches in various applications [24–27, 32]. More recently, iterative algorithms use alternating direction method of multipliers [34] and direction map [35] to improve the reconstruction quality and convergence rate. In a non-iterative manner, [36] utilizes the direct relationship between the logarithm of the amplitude and partial derivatives of the phase of the un-sampled STFT with respect to the Gaussian window. Additional features, such as instantaneous frequency (IF) [37] and group delay (GD) [38], have also been used to assist the phase reconstruction [39,40]. Other approaches [22,23,41,42] model the phase by using deep neural networks (DNNs) to further benefit from the prior knowledge of

the target signals.

One difficulty with DNN-based phase reconstruction is the wrapping issue. As the phase is wrapped in the range of $(-\pi, \pi]$, the conventional loss functions for regression, e.g., mean squared errors, become inefficient as they do not handle the periodicity. A solution for the wrapping issue is to use the *von Mises* distribution, which is a circular distribution. [43] and [44] are among the first studies to model the phase using the *von Mises* distribution for deriving a joint estimator of the amplitude and phase. Later, by using the same distribution, [41] proposed a cosine loss function for DNN-based phase estimation. Other approaches to deal with the wrapping issue are to cast the phase-regression problem into a classification problem of the quantized version of the phase [22, 23] or estimate the real and imaginary parts of the complex spectrogram instead of its amplitude and phase [45, 46]. However, there are also other problems for modeling the phase using DNNs, including the phase sensitivity to the waveform shift, i.e., only a small shift in the time domain can lead to a significant change in the phase spectrogram, especially at high frequencies. Another problem is sign indetermination [47], i.e., the STFTs of two signals $x(n)$ and $-x(n)$ have the same amplitudes but different phases. In other words, a given amplitude spectrogram may be consistent with both phase spectrograms $\mathbf{\Phi}$ and $\mathbf{\Phi} + \mathbf{1}\pi$, where $\mathbf{1}$ is an all-one matrix. The $\mathbf{\Phi}$ and $\mathbf{\Phi} + \mathbf{1}\pi$ usually yield very different values for most phase reconstruction loss functions; they are even opposite for the cosine loss function proposed in [41].

Rather than directly handling the phase, an alternative approach is to turn it into other representations using techniques such as differentiating or utilizing the phase distribution. Several alternative representations have been proposed and illustrated to display the useful information of the signal. They include the GD [38], IF [37], phase variance [48], phase distortion (PD) [49], relative phase shift (RPS) [50], and baseband phase difference [51]. To extract more information, several modifications of the IF and the GD have been proposed such as GD deviation [52], modified GD [38], IF deviation (IFD) [53], and IF density [54]. Each representation has its own strengths in different applications.

1.2 Objective

The aim of this thesis is to exploit the usefulness of phase-based features and develop DNN-based phase reconstruction algorithms.

In particular, this thesis reviews conventional phase-based features and proposes two new phase-based features, i.e., derivative of instantaneous frequency (DIF) and inter-frequency phase difference (IFPD). The DIF uses the derivative operation to reduce the wrapping issue, thereby revealing the useful structure of the phase. Meanwhile, the IFPD represents the phase relationships along the frequency, which is useful for phase reconstruction.

In addition to the phase features, this thesis develops phase reconstruction algorithms in two settings: offline and online. For the offline setting, this thesis focuses on the two-stage algorithms. In the first stage, the phase features are reconstructed from the amplitude. This thesis proposes several approaches for the second stage to reconstruct the phase from the phase features. These approaches include the weighted least squares method and maximum likelihood (ML) using Newton's method and coordinate descent. Several theoretical discussions/comparisons of the methods are also included.

For the online setting of phase reconstruction, the phase is directly reconstructed from the amplitude using a causal DNN model. This thesis introduces several improvements to the training process to enhance the results, including using a data augmentation scheme and a weighted loss function. Specifically, this thesis augments the training data by randomly shifting the signals in the time domain before calculating the STFT for each training epoch, based on the sensitivity property of the phase. For the loss function, the phase is modeled by the *von Mises* mixture model to reduce the sign indetermination problem, the amplitude is utilized as weights to separate the importance of phase elements at various time–frequency (TF) bins, and the phase features are used to improve the relationship of phase elements.

1.3 Organization

The remaining of this thesis is organized as follows.

Chapter 2 is dedicated to investigating the phase-based features. This chapter begins by establishing the notation and formulation. Subsequently, it provides a review on the conventional

phase-based features, such as IF, GD, IFD, RPS, and PD. In addition, two novel phase-based features, the DIF and IFPD, are introduced and their properties are investigated.

In Chapter 3, the focus is on the two-stage phase reconstruction algorithms in an offline setting. The chapter reviews conventional methods used in the second stage, which are the least squares (LS)-based and circular average-based methods. Following that, theoretical discussions are presented, including the GD analysis/normalization and an examination of the IF and GD information in the phase reconstruction. On the basis of these analyses, this thesis introduces the proposed methods for the second stage. The performance of the proposed methods are verified through subjective and objective experiments.

The online phase reconstruction is presented in Chapter 4. Upon examining the conventional method for DNN-based phase reconstruction, this thesis proposes new loss functions and the data augmentation methods to enhance the training process. Experiments are then conducted to confirm the efficacy of the proposed methods compared to the conventional methods.

Finally, Chapter 5 concludes this thesis and outlines future work.

Chapter 2

Phase-based features

This chapter presents alternative representations of the phase. Before going to the details of the phase representations, we define some notations and formulations. Let $x(n)$ be a signal in the time domain, where n is the time sample index. The STFT of $x(n)$ is defined as

$$X_{k,\ell} = \sum_{n=0}^{N-1} w(n)x(n+\ell R)e^{-j2\pi kn/N}, \tag{2.1}$$

where $\ell \in \{0, \ldots, L-1\}$ is the frame index, $k \in \{0, \ldots, K-1\}$ is the frequency bin index, and R and N are the window shift and number of Discrete Fourier transform (DFT) points, respectively. The window length is denoted as M. The STFT phase and amplitude are then denoted as $\Phi_{k,\ell} = \angle X_{k,\ell}$ and $|X_{k,\ell}|$, respectively, where $\angle$ is the angle operator. The wrapping function mapping a value into the principal range of $(-\pi, \pi]$ is denoted as $\mathcal{P}(\cdot)$.

In the following, conventional phase-based features, i.e., the IF, GD, IFD, RPS, and PD, are reviewed in Section 2.1, and the proposed features, DIF and IFPD, are described in Section 2.2. Fig. 2.1 shows an example of these phase-based features of a speech sample.

2.1 Conventional phase-based features

2.1.1 Instantaneous frequency and group delay

Among the features, the dual representations playing a major role in phase-based feature extraction are the IF and the GD. IF is defined as the derivative of the phase with respect to time,

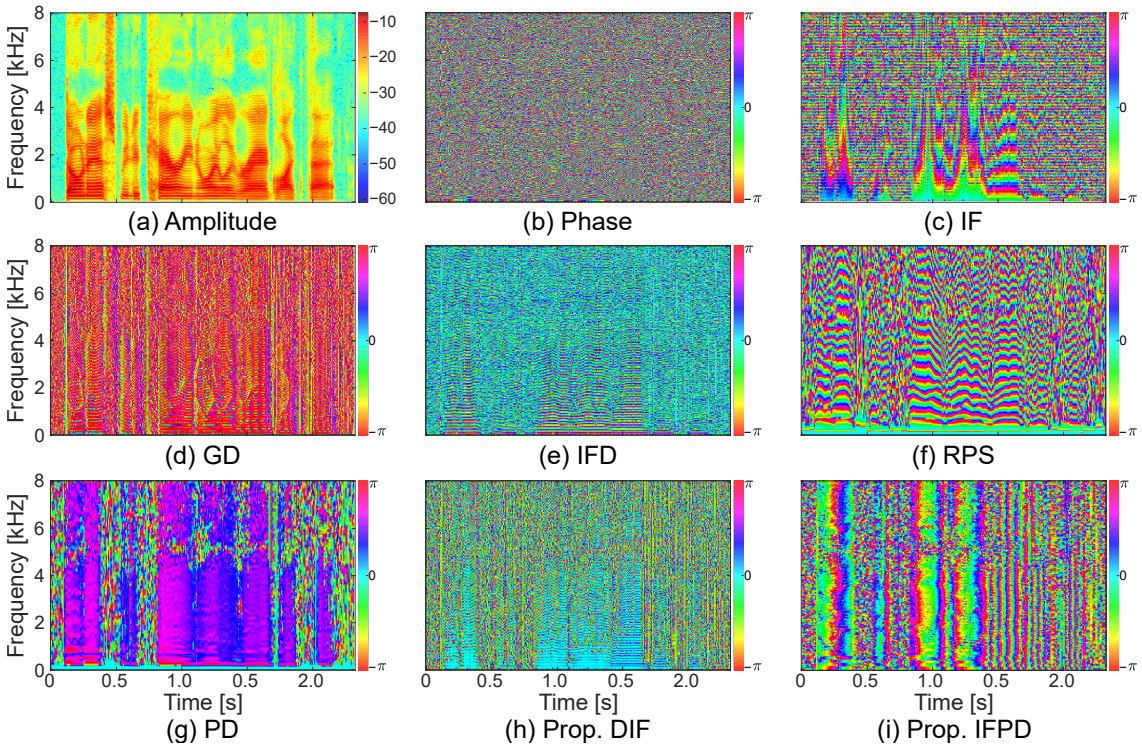


Figure 2.1: Example of phase-based features. STFT is calculated using Hann window with 32-ms length, 8-ms shift, and 512-point DFT. Sampling rate is 16 kHz.

which can be estimated by the phase difference as

$$V_{k,\ell} = \mathcal{P}(\Phi_{k,\ell+1} - \Phi_{k,\ell}). \tag{2.2}$$

Similarly, GD, which is a negative frequency derivative of the phase, can be calculated as

$$U_{k,\ell} = \mathcal{P}(\Phi_{k,\ell} - \Phi_{k+1,\ell}). \tag{2.3}$$

By differentiating the phase, the effect of the wrapping issue can be reduced, thereby revealing the harmonic structure of the speech signal, as illustrated in Fig. 2.1(c) and (d). The IF and the GD have been reported to be useful in formant extraction [55,56], speaker identification [57,58], source separation [59], speech segmentation [60], and so forth.

2.1.2 IF deviation

By utilizing the relationship between the IF and the frequency bins, Stark *et al.* [53] proposed a modification of the IF, called IF deviation, as

$$\text{IFD}_{k,n} = \mathcal{P}(V_{k,n} - \Omega_k), \tag{2.4}$$

where Ω_k represents the normalized frequency at bin k. The proposed formula above is for the STFT with the window shift of 1 sample. For the window shift of R samples, it becomes

$$\text{IFD}_{k,\ell} = \mathcal{P}\left(\frac{V_{k,\ell} - R\Omega_k}{R}\right). \tag{2.5}$$

Fig. 2.1(e) shows the IFD spectrogram. We can see that the harmonic frequencies are shown as the thin bright lines, which better exhibit the harmonic structure of the signal compared to the IF spectrogram.

2.1.3 Relative phase shift

By assuming a harmonic model on the speech signal, [50] proposed a representation for harmonic phase information called RPS as

$$\text{RPS}_\ell^h = \Phi_\ell^h - h\Phi_\ell^1, \tag{2.6}$$

where h is the index of the hth harmonic component and Φ_ℓ^h is its phase at frame ℓ. In [61], the phase of the hth harmonic component can be decomposed into a sum of the source shape, the linear phase, and the phase of the vocal tract filter. The RPS calculation discards the linear phase terms in the harmonic phase. The remaining terms can be assumed to change smoothly along time, thereby reducing the phase wrapping issue and displaying useful information as illustrated in Fig. 2.1(f).

2.1.4 Phase distortion

The phase distortion [49] is a variation of the relative phase shift. The relative phase shift calculation still contains the harmonic number h. However, h depends on the harmonic structure, which relies on the fundamental frequency and harmonicity property of the model. In addition, the variance of the RPS will increase as h increases. To solve these problems, the phase distortion

is defined as the difference between two consecutive relative phase shifts as

$$\begin{aligned} \mathrm{PD}_\ell^h &= \mathrm{RPS}_\ell^{h+1} - \mathrm{RPS}_\ell^h \\ &= \Phi_\ell^{h+1} - \Phi_\ell^h - \Phi_\ell^1. \end{aligned} \tag{2.7}$$

The PD measures the desynchronization between sinusoidal components of the voice source as shown in Fig. 2.1(g).

2.2 Proposed phase-based features

2.2.1 Derivative of instantaneous frequency

In order to track the change of the IF along the frequency, this thesis defines the DIF as the derivative of the IF with respect to frequency. In other words, the DIF is the second order mixed partial derivative of the phase; therefore, it is also the negative time derivative of the GD. We can calculate the DIF from either the IF or the GD as

$$\begin{aligned} \mathrm{DIF}_{k,\ell} &= \mathcal{P}\left(V_{k+1,\ell} - V_{k,\ell}\right) \\ &= \mathcal{P}\left(U_{k,\ell} - U_{k,\ell+1}\right). \end{aligned} \tag{2.8}$$

Figure 2.1(h) depicts the DIF spectrogram. We can see that the DIF shows a clear harmonic structure corresponding to the amplitude spectrogram. Because the phase contains information about the frequency, by differentiating, we can reduce the influence of the wrapping issue, thereby revealing information about the periodic segments, i.e., the voiced speech. However, for the same reason, the DIF cannot show much information about the non-periodic segments like the unvoiced speech. The unvoiced-speech segment in Fig. 2.1(h) (from 0.4 to 0.5 seconds) looks similar to the non-speech segment (from 0.0 to 0.2 seconds), even though we can easily distinguish them in the amplitude spectrogram.

Figure 2.2 compares the distributions of the DIF and the raw phase in the speech and non-speech segments. In both segments, the raw phase has a uniform distribution with the values spread from $-\pi$ to π. The values of the DIF concentrate around zero. In the non-speech segment, the DIF distribution peaks at a positive value, while in the speech segment, it has a high peak at zero. It demonstrates that the DIF has different distributions in speech and non-speech segments, although the raw-phase distributions at those segments seem to be indistinguishable. This property can be useful for the voice activity detection application.

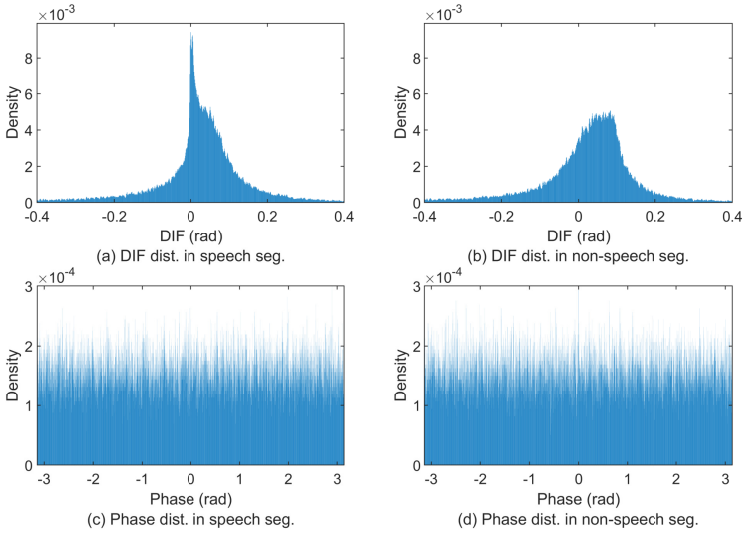


Figure 2.2: Comparison of the estimated distributions of the DIF and the raw phase in speech/non-speech segments.

2.2.2 Inter-frequency phase difference

As a phase difference between two consecutive TF bins, the GD only represents local relationships of the phase. However, all TF bins in the same frame are interdependent because each of them depends on all the underlying components of the signal. For those reasons, to better represent the phase relationships along frequency, this thesis generalizes the calculation of the GD to the phase difference between two frequency bins with the frequency hop of i bins. The IFPD is defined as

$$U_{k,\ell}^{(i)} = \mathcal{P}(\Phi_{k,\ell} - \Phi_{k+i,\ell}). \tag{2.9}$$

For $i = 1$, $U_{k,\ell}^{(i)}$ is identical to $U_{k,\ell}$. It is worth noting that the subtraction operation in the GD calculation is an estimation of the phase derivative, while for the IFPD, it is merely the phase difference between two frequency bins. For the frequency hops smaller than the main-lobe width of the window function, the IFPD has similar properties to the GD in that its value at the strong components is close to zero. For larger frequency hops, the IFPD may capture the phase difference between two harmonic components if the hop is close to the multiple of the fundamental frequency. This property relates to the RPS and PD described above, which also reflect the phase relationships between harmonic components. However, the calculation of these

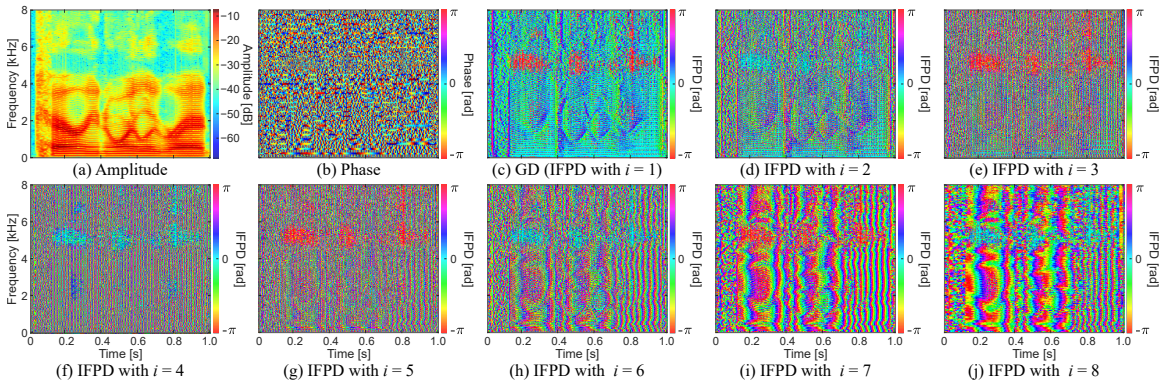


Figure 2.3: Example of IFPD of speech signal with various frequency hops i. STFT is calculated using Hamming window with 32-ms length, 4-ms shift, and 512-point DFT. Sampling rate is 16 kHz.

features relies on the estimation of the fundamental frequency and harmonic model, while the IFPD is literally based on the DFT. An example of the IFPD of a speech signal with the frequency hop varying from 1 to 8 bins is shown in Fig. 2.3.

To illustrate how the IFPD captures the phase difference between harmonic components, we analyze the example of the IFPD in Fig. 2.3. Let us represent two harmonic components by sinusoids as

$$x_1(n) = A_1 \cos(2\pi k_1 n/N + \phi_1), \tag{2.10}$$

$$x_2(n) = A_2 \cos(2\pi k_2 n/N + \phi_2), \tag{2.11}$$

where A_1 and A_2 are the amplitudes, ϕ_1 and ϕ_2 are the initial phases, and k_1 and k_2 are the frequencies of the sinusoids. The phase difference between the two components at frame ℓ is

$$\Delta\varphi_\ell = \mathcal{P}\left(\frac{2\pi\Delta k}{N} R\ell + \Delta\varphi_0\right), \tag{2.12}$$

where $\Delta\varphi_0 = \phi_1 - \phi_2$ is the phase difference at the time origin, and $\Delta k = k_1 - k_2$ is the frequency difference. Since the signal in Fig. 2.3 has the fundamental frequency close to 4 bins, the IFPD with the frequency hop of 4 bins shows the phase difference between two consecutive harmonic components. With $\Delta k = -4$, $R = 64$, and $N = 512$, from (2.12), we have $\Delta\varphi_\ell = \mathcal{P}\left(-\ell\pi + \Delta\varphi_0\right)$. Along the time frame, as ℓ increases, $\Delta\varphi_\ell$ switches between $\Delta\varphi_0$ and $\Delta\varphi_0 + \pi$, resulting in the vertical stripes in the IFPD spectrogram shown in Fig. 2.3(f). Considering two more distant harmonic components with the frequency hop of

8 bins, corresponding to $\Delta k = -8$, the phase difference is nearly constant along the time frame as $\Delta\varphi_\ell = \mathcal{P}(-2\ell\pi + \Delta\varphi_0) = \Delta\varphi_0$. Although $\Delta\varphi_\ell$ is a constant, the IFPD spectrogram shown in Fig. 2.3(j) changes slowly along the time in accordance with the variations of the fundamental frequency. The IFPD spectrograms with other frequency hops in Fig. 2.3 also exhibit similar patterns, although not as clearly as those with frequency bin hops that are multiple of the fundamental frequency.

Chapter 3

Two-stage phase reconstruction

3.1 Introduction

Instead of directly reconstructing the phase, two-stage approaches were proposed [11–13] for reconstructing the phase through the phase derivatives, i.e., the IF and GD. Although the phase changes quickly along the time and frequency, its change rate between neighboring elements is more stable. The IF and GD extract that change rate through the derivative operation, thereby reducing the sensitivity and wrapping issues and revealing the underlying structure of the phase. Experimental results [11] indicated the efficacy of such a two-stage approach over directly reconstructing the phase.

The first stage is almost the same for all the two-stage phase reconstruction algorithms in that the IF and GD are estimated from the amplitude using DNNs. Not only are the IF and GD more structured than the phase, they are also not affected by the sign-indetermination problem because the ambiguity of 1π becomes zero after the derivative operation. Therefore, the IF and GD are reconstructed much more easily than the phase itself. In the second stage, the phase is reconstructed from the IF and GD using various methods including the LS [11] and circular average [12].

In this chapter, this thesis reviews conventional methods and presents proposed methods for two-stage phase reconstruction algorithms. Specifically, this thesis improves upon the LS-based method [11] by introducing amplitude weights, which reflect the importance of each TF bin, to the error function and applying a tridiagonal system algorithm to reduce the calculation

time. To tackle the wrapping issue, this thesis also proposes a ML-based approach using the *von Mises* distribution. The ML problem is solved using Newton's method and a coordinate-descent strategy. Comparisons among the approaches for the second stage are then discussed from theoretical aspects. This thesis also presents several new analyses for two-stage phase reconstruction algorithms. Specifically, the properties of the GD are examined to propose a GD normalization method for facilitating the training process of the GD-estimation model in the first stage. Another analysis is the investigation of a narrow-band signal to interpret the effects of the IF and GD information on the reconstructed waveforms using the ISTFT. From the discussions, this thesis uses the IFPD to improve two-stage phase reconstruction algorithms. Subjective and objective experiments are conducted to verify the performance of the proposed methods.

The remainder of this chapter is organized as follows. This thesis reviews related work in Section 3.2, and analyses the phase properties for two-stage phase reconstruction algorithms in Section 3.3. In Section 3.4, this thesis describes the proposed methods and present the comparisons of them with conventional methods. In Section 3.5, this thesis discusses the experiments on the efficacy of our proposed methods and present the results. Finally, this thesis concludes the chapter in Section 3.6.

3.2 Conventional two-stage phase reconstruction

We first define the notations and formulations used in the two-stage phase reconstruction algorithms. The vector notations of phase spectrum, IF, and GD at frame ℓ are denoted as $\boldsymbol{\phi}_\ell = (\Phi_{0,\ell}, \cdots, \Phi_{K-1,\ell})^\mathsf{T}$, $\boldsymbol{v}_\ell$, and $\boldsymbol{u}_\ell$, respectively, where $(\cdot)^\mathsf{T}$ is a matrix transposition operator. Thus, we have

$$\boldsymbol{v}_\ell = \boldsymbol{\phi}_{\ell+1} - \boldsymbol{\phi}_\ell, \tag{3.1}$$

$$\boldsymbol{u}_\ell = \boldsymbol{D}\boldsymbol{\phi}_\ell, \tag{3.2}$$

where $\boldsymbol{D}$ is a $(K-1)\times K$ upper bidiagonal matrix defined as

$$(\boldsymbol{D})_{i,j} = \begin{cases} 1, & \text{if } i = j \\ -1, & \text{if } i+1 = j \\ 0, & \text{otherwise} \end{cases}. \tag{3.3}$$

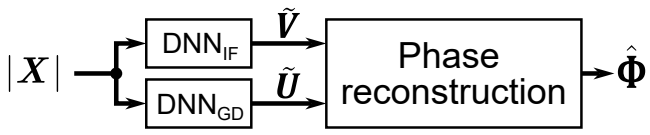


Figure 3.1: Diagram of two-stage phase reconstruction algorithms. First stage consists of two DNNs for estimating IF $\tilde{\boldsymbol{V}}$ and GD $\tilde{\boldsymbol{U}}$. Second stage reconstructs phase $\hat{\boldsymbol{\Phi}}$ from $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{U}}$.

The matrix notations for the amplitude, phase, IF, and GD spectrograms are $|\boldsymbol{X}|$, $\boldsymbol{\Phi}$, $\boldsymbol{V}$, and $\boldsymbol{U}$, respectively. In two-stage phase reconstruction algorithms, the notations $\tilde{\cdot}$ and $\hat{\cdot}$ denote the estimates in the first and second stages, respectively, and $\cdot^*$ denotes the normalization (which is described in Sections 3.2.1 and 3.3.1).

Two-stage phase reconstruction algorithms are aimed at estimating the phase $\boldsymbol{\Phi}$ from a given amplitude $|\boldsymbol{X}|$ indirectly through the IF $\boldsymbol{V}$ and GD $\boldsymbol{U}$, as illustrated in Fig. 3.1.

3.2.1 First stage: IF/GD estimation using DNN

The first stage is similar for two-stage phase reconstruction algorithms, in which the IF $\tilde{\boldsymbol{V}}$ and GD $\tilde{\boldsymbol{U}}$ are reconstructed from the amplitude using *von Mises* distribution-based DNNs [41]. The *von Mises* distribution is also known as the circular normal distribution, which can be used to model circular data like the phase. Its probability density function is defined as

$$f(x|\mu,\kappa) = \frac{e^{\kappa\cos(x-\mu)}}{2\pi I_0(\kappa)}, \tag{3.4}$$

where x is a circular variable, μ is a measure of location, κ is a measure of concentration, and $I_0(\kappa)$ is a modified Bessel function of order 0. μ and $1/\kappa$ are analogous to the mean and variance of the normal distribution. The negative logarithm of (3.4) is given as

$$-\log f(x|\mu,\kappa) = -\kappa\cos(x-\mu) + \mathcal{C}, \tag{3.5}$$

where $\mathcal{C}$ is a constant to x. By modeling the IF/GD by the *von Mises* distribution with the assumption that κ is constant for all the data points, the error function for the DNNs reconstructing the IF/GD is defined as

$$\mathcal{L}_{\text{DNN}}(\boldsymbol{y}_\ell, \tilde{\boldsymbol{y}}_\ell) = -\sum_k \cos(Y_{k,\ell} - \tilde{Y}_{k,\ell}), \tag{3.6}$$

where $\boldsymbol{y}_\ell$ and $\tilde{\boldsymbol{y}}_\ell$ are the original and estimated values of the output, which is either the IF or GD. To improve the DNN training process, the IF and GD are normalized so that their distributions have peaks at zero. [18] proposed an IF normalization method as

$$V^*_{k,\ell} = \mathcal{P}\left(V_{k,\ell} - 2\pi kR/N\right), \tag{3.7}$$

which removes the between-frame phase shift of $2\pi kR/N$ from the IF. [18] also proposed a GD normalization scheme that subtracts π from all of its elements, which is based on the observation that the GD histogram has a peak near π. [12] demonstrated that the IF and GD normalizations in [18] are useful for the DNN training process of two-stage phase reconstruction algorithms.

3.2.2 Second stage: phase estimation from IF and GD

In the second stage, the phase $\hat{\boldsymbol{\Phi}}$ is reconstructed from the estimated IF $\tilde{\boldsymbol{V}}$ and GD $\tilde{\boldsymbol{U}}$. Two conventional methods for the second stage, LS-based [11] and weighted circular average-based [12], will be described in the following.

Least squares

Inspired by the LS-based method for 2D-phase unwrapping in [62], [11] proposed recursively reconstructing the phase from the IF and GD for each frame by minimizing the quadratic error function defined as

$$\mathcal{L}_{\text{LS}}(\boldsymbol{\phi}_\ell) = \|\boldsymbol{\phi}_\ell - \boldsymbol{\phi}_{\ell-1} - \tilde{\boldsymbol{v}}_{\ell-1}\|^2 + \|\boldsymbol{D}\boldsymbol{\phi}_\ell - \tilde{\boldsymbol{u}}_\ell\|^2. \tag{3.8}$$

When estimating $\boldsymbol{\phi}_\ell$, $\boldsymbol{\phi}_{\ell-1}$ is replaced with the wrapped version of its previously estimated value, i.e., $\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} = \mathcal{P}(\hat{\boldsymbol{\phi}}_{\ell-1})$. Since $\tilde{\boldsymbol{v}}_\ell$ and $\tilde{\boldsymbol{u}}_\ell$ are also wrapped, which may lead to a detrimental effect on the LS solution, [11] proposed modifying $\tilde{\boldsymbol{u}}_\ell$ in (3.8) as

$$\tilde{\boldsymbol{u}}_\ell \leftarrow \boldsymbol{D}(\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1}) + \mathcal{P}\left(\tilde{\boldsymbol{u}}_\ell - \boldsymbol{D}(\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1})\right). \tag{3.9}$$

(3.9) adds 2π jumps to $\tilde{\boldsymbol{u}}_\ell$ to make it more consistent with the GD calculated from the previously estimated phase and IF, i.e., $\boldsymbol{D}(\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1})$. The solution for minimizing (3.8) is

$$\hat{\boldsymbol{\phi}}_\ell = (\boldsymbol{I}_K + \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}(\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1} + \boldsymbol{D}^{\mathsf{T}}\tilde{\boldsymbol{u}}_\ell), \tag{3.10}$$

where $\boldsymbol{I}_K$ is a $K \times K$ identity matrix.

Circular average

By incorporating the amplitude information, [12] proposed a simple weighted circular average-based method that estimates the phase for each TF bin as

$$\hat{\Phi}_{k,\ell} = \angle \sum_{q=1}^{Q} W_{k,\ell}^{(q)} \exp\left(j\varphi_{k,\ell}^{(q)}\right), \tag{3.11}$$

where $\varphi_{k,\ell}^{(q)}$ is an estimate of $\Phi_{k,\ell}$ computed from the IF, GD, and the qth previously estimated phase element near $\Phi_{k,\ell}$. $W_{k,\ell}^{(q)}$ is the amplitude weight, and Q is the number of the neighbors involved. [12] also empirically determined that $Q = 3$ yields the best result, i.e., $\hat{\Phi}_{k,\ell}$ is calculated from $\hat{\Phi}_{k-1,\ell}$, $\hat{\Phi}_{k,\ell-1}$, and $\hat{\Phi}_{k+1,\ell-1}$.

3.3 Phase analysis for two-stage phase reconstruction algorithms

In this section, this thesis provides theoretical discussions for two-stage phase reconstruction algorithms. Section 3.3.1 analyzes the properties of the phase and GD calculated with two types of the window functions and introduces our analytic GD normalization formula. Section 3.3.2 investigates the effects of phase modifications on the ISTFT of a sinusoidal wave to interpret how the IF and GD are useful for phase reconstruction.

3.3.1 GD analysis and normalization

For normalizing the GD, [18] proposed subtracting the peak π of the GD histogram. However, this peak varies with the window length and number of DFT points. Instead of following the method in [18], this thesis analyses the GD values and introduce an analytic formula for GD normalization as follows.

When calculating the STFT, we usually multiply each signal frame by a window function before calculating the DFT. This is equivalent to a convolution in the frequency domain, as

$$\boldsymbol{x}_\ell = \boldsymbol{s}_\ell * \boldsymbol{w}, \tag{3.12}$$

where $\boldsymbol{x}_\ell$, $\boldsymbol{s}_\ell$, and $\boldsymbol{w}$ are the DFT spectra of the windowed signal, target signal at frames ℓ, and

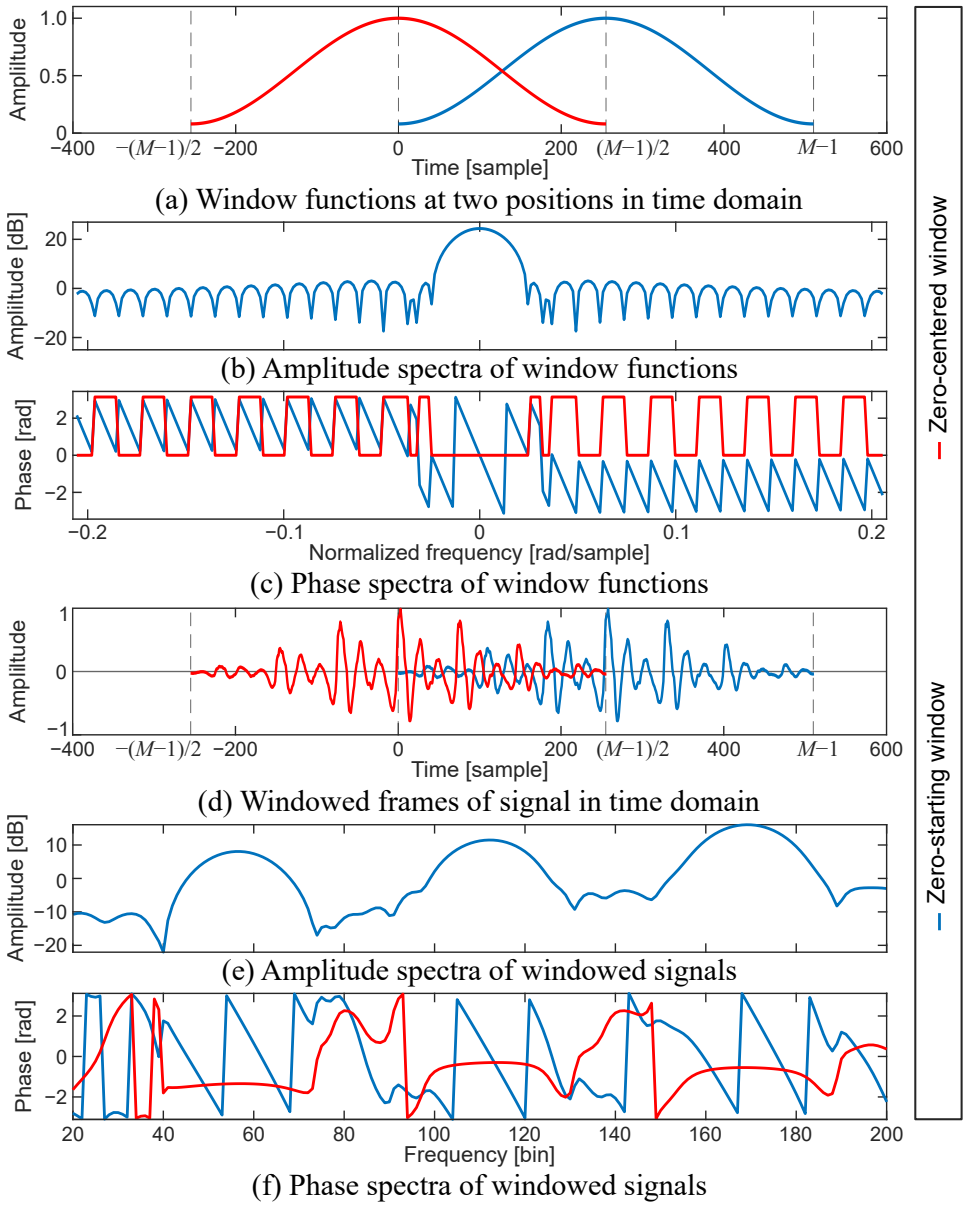


Figure 3.2: Example of Hamming window at two positions and windowed frame of speech signal in time and frequency domains. Window length and number of DFT points are $M = N = 511$.

window function, respectively. Equivalently, elements of $\boldsymbol{x}_\ell$ can be calculated as

$$x_\ell(k) = \sum_{m=0}^{N-1} s_\ell(m) w(k-m). \tag{3.13}$$

From (3.12), we can see that the phase of $\boldsymbol{x}_\ell$ is affected by $\boldsymbol{w}$. The phase of $\boldsymbol{w}$ relies not only on the type of the window function but also on the position (in other words, the time origin) of the window function in the DFT formula. Fig. 3.2(a) to (c) shows an example of the Hamming window at two positions in the time and frequency domains, i.e., that starting from

zero (typically used in STFT implementations, as shown with blue lines) and that centered at zero (as shown with red lines). When the window function is symmetric and centered at zero (in other words, the time origin of the DFT formula is at the center of the frame), $\boldsymbol{w}$ is a real-valued vector. From (3.13), we see that each element $x_\ell(k)$ is a linear combination of all the elements of $\boldsymbol{s}_\ell$. However, the contributions of the elements of $\boldsymbol{s}_\ell$ at frequency bins far from bin k are scaled down by the low side lobes of $\boldsymbol{w}$, as shown in Fig. 3.2(b). As a result, the phase of $\boldsymbol{x}_\ell$ at each frequency bin is mostly dominated by the nearest strong component of $\boldsymbol{s}_\ell$. By shifting the zero-centered window to the right side by $(M-1)/2$ samples, we obtain a zero-starting window, i.e., the time origin of the DFT formula is at the beginning of the frame. In this case, the amplitude is the same, but the phase at each frequency bin k will be shifted by $-\frac{2\pi k}{N}\frac{M-1}{2}$ compared with the case of a zero-centered window, as illustrated in Fig. 3.2(b) and (c).

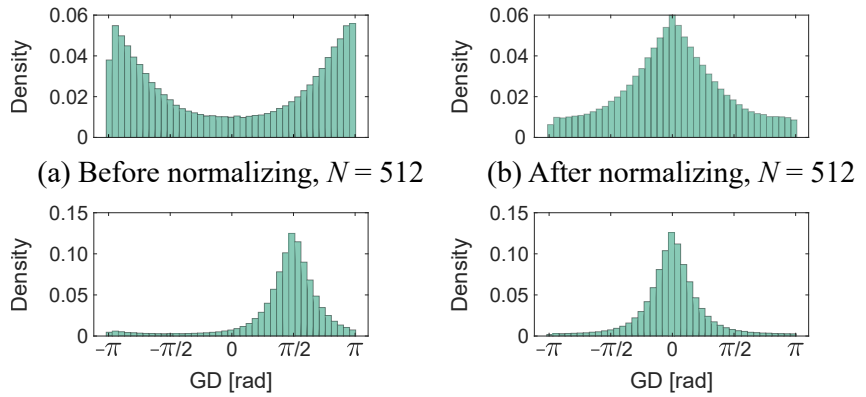
Fig. 3.2(d) to (f) shows an example of a windowed frame of a speech sample calculated using the zero-centered and zero-starting windows. For the zero-centered window, we see that phase elements around a strong component (an amplitude peak), e.g., around the frequency bin of 56, are dominated by that component, hence nearly constant. As a result, the GD, which is a negative frequency-derivative of the phase, is approximately zero. For the zero-starting window, the phase at those frequency bins becomes an oblique line due to the linear phase shift, and the GD will be a non-zero constant. These properties are similar for weak components of the signal, although the affected area around a weak component is narrower than that around a strong component.

From the above discussion, we can see that the GD values calculated using the zero-centered window naturally concentrate around zero. The linear phase shift introduced by the commonly used zero-starting window shifts the peak of the GD histogram to a non-zero constant. We hence propose normalizing the GD by compensating for that linear phase shift using either the following methods.

- Circular-shift the windowed signal to the left by $(M-1)/2$ samples when calculating the DFT.

- Compensate for the phase shift by

$$\Phi^*_{k,\ell} = \mathcal{P}\left(\Phi_{k,\ell} + \frac{2\pi k}{N}\frac{(M-1)}{2}\right). \tag{3.14}$$



(a) Before normalizing, $N = 512$ (b) After normalizing, $N = 512$

(c) Before normalizing, $N = 1024$ (d) After normalizing, $N = 1024$

Figure 3.3: Examples of GD histograms of speech sample before and after normalization. STFT is calculated with $M = 512$ and various N.

- Directly modify the GD by

$$U^*_{k,\ell} = \mathcal{P}\left(U_{k,\ell} - \frac{\pi(M-1)}{N}\right). \tag{3.15}$$

It is worth noting that $(M-1)/2$ is not necessarily an integer; hence, M can be either even or odd. (3.15) is similar to the GD normalization formula proposed by [18] in terms of subtracting a number from the GD. However, when $M = N$, instead of subtracting π as with the method in [18], we can see from (3.15) that the subtrahend is $\pi(M-1)/N$. The advantage of (3.15) is that it can be applied to other settings of M and N without requiring calculating the GD histogram of the training data. Fig. 3.3 shows examples of GD histograms of a speech sample before and after normalizing using (3.15).

3.3.2 IF and GD information in phase reconstruction

Two-stage phase reconstruction algorithms indeed reconstruct the phase by maintaining the phase relationships between TF bins along time and frequency through the IF and GD, respectively. Since the ISTFT has the form of a sum of complex numbers, if the phase relationships between those complex numbers are maintained, i.e., the phase differences between TF bins remain unchanged, the amplitude of the reconstructed signal will be consistent even if the phase is shifted. To investigate the role of the phase relationships in the ISTFT, we modify the phase spectra calculated from a sinusoidal wave and observe the effects on the reconstructed waveform as follows.

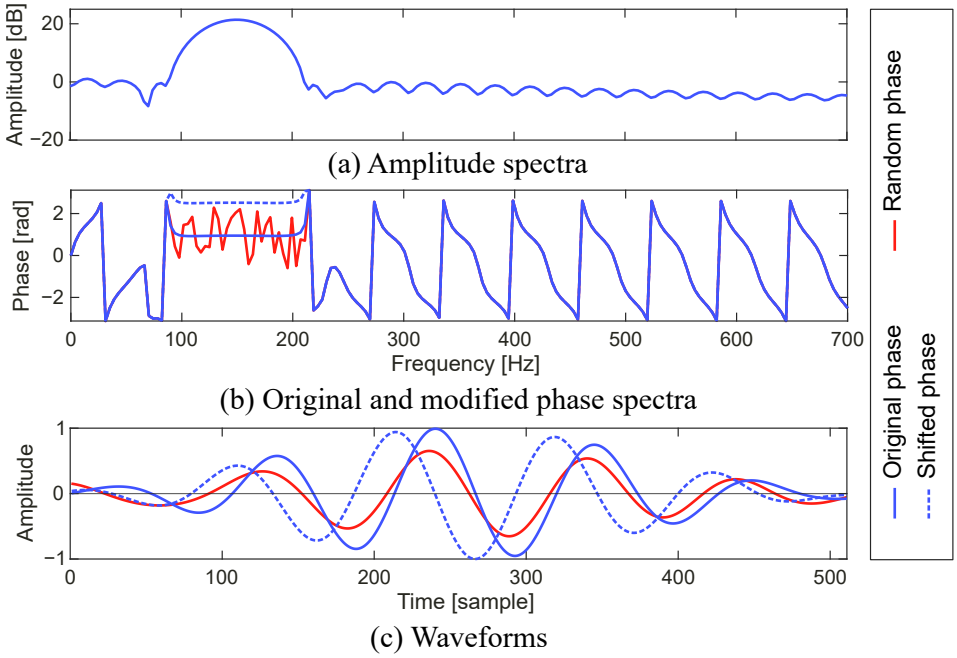


Figure 3.4: Example of effects of phase modifications on IDFT. DFT of 32-ms frame of 150-Hz sinusoidal wave is calculated with Hamming window. Sampling frequency is 16 kHz. Original phase spectrum (blue solid line) is modified by adding random numbers (red solid line) or constant of $\pi/2$ (blue dashed line). Note that only area around 150 Hz is modified because other areas do not have much of effect on IDFT.

Fig. 3.4 shows the DFT spectra and waveforms of the sinusoidal wave, in which the phase spectrum is modified in a frequency range by using two methods: 1) adding random numbers to each element, i.e., the phase relationships along the frequency are broken, and 2) adding a constant of $\pi/2$ to all the elements, i.e., the phase relationships along the frequency are maintained. Each phase spectrum is then combined with the amplitude spectrum to reconstruct the waveform using the inverse DFT (IDFT). We can see from Fig. 3.4(c) that the reconstructed waveform for the maintained phase relationships has almost the same amplitude as the original waveform, although it is shifted in the time domain. In contrast, the waveform of the randomly modified phase has a lower amplitude due to the misalignment of the complex spectral bins in the IDFT calculation.

Fig. 3.5 shows an example of two consecutive frames in the same signal used above. The phase spectra are shifted by adding a constant, which is $\pi/2$ in this example, to all the elements. The IDFTs of the original and modified versions of the first and second frames are then combined

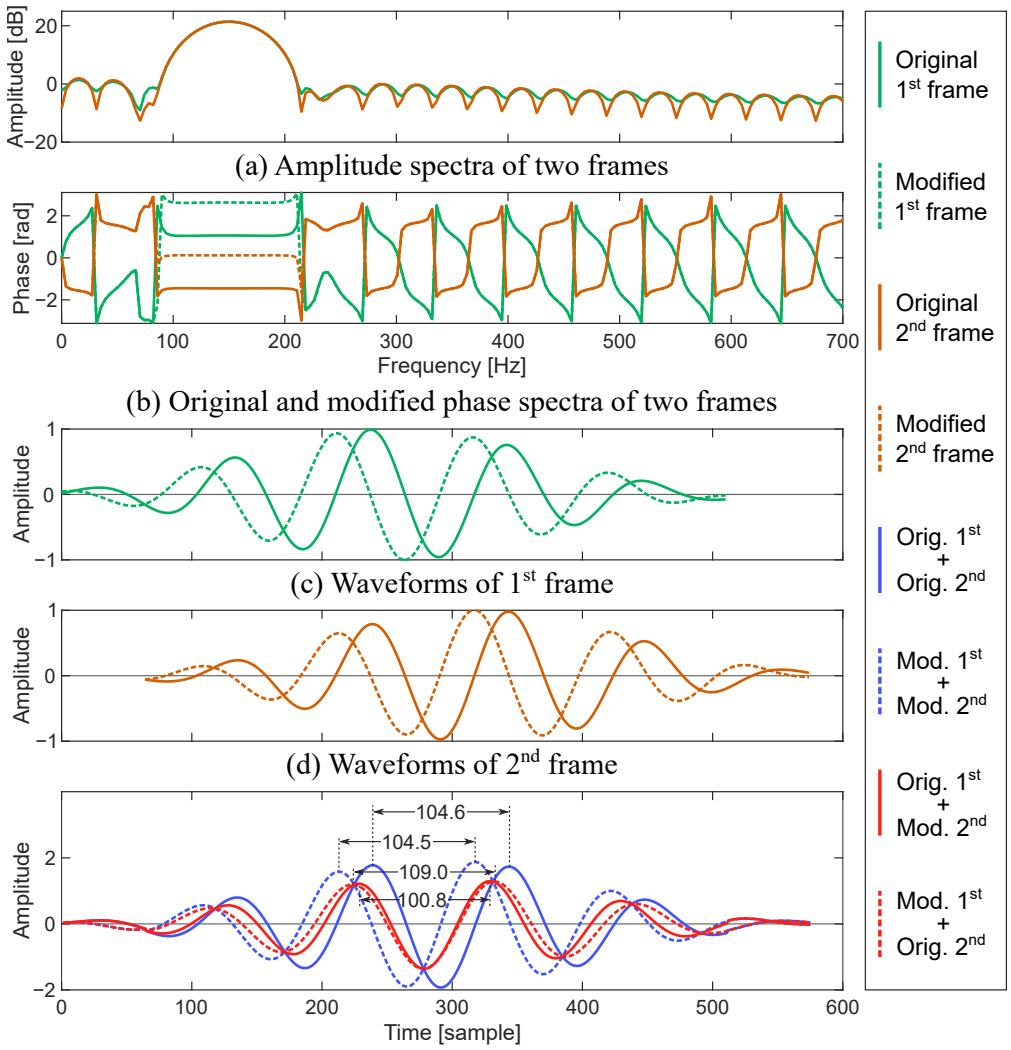


Figure 3.5: Example of effects of phase modifications on overlap-add. Signal and DFT setting are same as in Fig. 3.4, in which two consecutive frames with hop of $4\,\mathrm{ms}$ are shown. Phase spectra are modified by adding $\pi/2$ to all elements.

for each pair using the overlap-add method. We can see from this figure that, if the phase spectra of both frames are shifted the same way, i.e., the IF information is maintained, their waveforms are aligned well for the overlap-add to yield the same amplitude as the original signal. In other situations, when the phase spectrum of only one frame is modified, the misalignments between the reconstructed waveforms of the two frames decrease the amplitude of the overlap-added signals. Those misalignments also cause a frequency modulation to the overlap-added signals, as illustrated by the period changes in Fig. 3.5(e).

From the above discussion, we can see that the distortions in the IF and GD information result in the degradation in both the amplitude and frequency of the signal. When the phase relationships between TF bins are maintained, even if the phase is shifted, the reconstructed signal still has the same amplitude as the original signal. The shifted phase only introduces a shift of the signal in the time domain, which, based on our observation, makes no difference in perception of sound quality.

3.4 Proposed phase reconstruction methods

All two-stage phase reconstruction algorithms have the same first stage that estimates the IF and GD from the amplitude using DNNs. In this section, this thesis proposes two approaches for reconstructing the phase from the IF and GD in the second stage, i.e., weighted LS-based (Section 3.4.1) and ML-based using *von Mises* distribution (Section 3.4.2). The ML-based methods are also divided into two optimization approaches, i.e., using Newton's method and coordinate descent. This thesis then presents the comparison of the proposed and conventional methods in theoretical aspects (Section 3.4.3). The application of the IFPD to the two stage phase reconstruction algorithm is described in Section 3.4.4.

3.4.1 Weighted least squares

The contribution of each TF bin to the ISTFT depends highly on its amplitude. We also observed that errors of the IF and GD reconstructed in the first stage are low at high-amplitude positions. Therefore, we improved the conventional LS-based method [11] by adding amplitude weights to the error function (3.8) to emphasize the importance of the high-amplitude TF bins. The weighted error function is defined as

$$\begin{aligned}\mathcal{L}_{\text{WLS}}(\boldsymbol{\phi}_\ell) =& \|\sqrt{\breve{\boldsymbol{W}}^v_{\ell-1}}(\boldsymbol{\phi}_\ell - \boldsymbol{\phi}_{\ell-1} - \tilde{\boldsymbol{v}}_{\ell-1})\|^2 \\ &+ \|\sqrt{\breve{\boldsymbol{W}}^u_{\ell}}(\boldsymbol{D}\boldsymbol{\phi}_\ell - \tilde{\boldsymbol{u}}_{\ell})\|^2,\end{aligned} \tag{3.16}$$

where $\breve{\boldsymbol{W}}^v_\ell$ and $\breve{\boldsymbol{W}}^u_\ell$ are diagonal weight matrices of the IF $\tilde{\boldsymbol{v}}_\ell$ and GD $\tilde{\boldsymbol{u}}_\ell$, respectively. The kth element on the main diagonal of both $\breve{\boldsymbol{W}}^v_\ell$ and $\breve{\boldsymbol{W}}^u_\ell$ are empirically set to $|X_{k,\ell}|^p$. The power of p ($p \geq 1$) is used to further separate the low- and high-amplitude TF bins. We also use the GD modification method (3.9) to address the wrapping issue.

Algorithm 1 Pseudo-code of weighted LS-based method for reconstructing phase from IF and GD.

Input: Amplitude $|\boldsymbol{X}|$, estimated IF $\tilde{\boldsymbol{V}}$ and GD $\tilde{\boldsymbol{U}}$

Output: Phase spectrogram $\hat{\boldsymbol{\Phi}}$

$\hat{\Phi}_{0,0} \leftarrow 0$

$\hat{\Phi}_{k,0} \leftarrow \hat{\Phi}_{k-1,0} - \tilde{U}_{k-1,0}$, for $k \in \{1, \ldots, K-1\}$

for $\ell \in \{1, \ldots, L-1\}$ **do**

$\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} \leftarrow \mathcal{P}(\hat{\boldsymbol{\phi}}_{\ell-1})$

Update $\tilde{\boldsymbol{u}}_\ell$ as in (3.9)

Calculate $\hat{\boldsymbol{\phi}}_\ell$ as in (3.17)

The analytic solution for minimizing (3.16) is

$$\begin{aligned}\hat{\boldsymbol{\phi}}_\ell =&(\breve{\boldsymbol{W}}^v_{\ell-1} + \boldsymbol{D}^\mathsf{T}\breve{\boldsymbol{W}}^u_\ell \boldsymbol{D})^{-1} \\ &\cdot (\breve{\boldsymbol{W}}^v_{\ell-1}(\hat{\boldsymbol{\phi}}^{\mathcal{P}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1}) + \boldsymbol{D}^\mathsf{T}\breve{\boldsymbol{W}}^u_\ell \tilde{\boldsymbol{u}}_\ell).\end{aligned} \tag{3.17}$$

The derivation of this solution is explained in Appendix. Most of the calculation time of (3.17) is spent in calculating the inverse of a $K \times K$ matrix. However, we can see that (3.17) has a form of $\hat{\boldsymbol{\phi}}_\ell = \boldsymbol{A}^{-1}\boldsymbol{b}$, where $\boldsymbol{A}$ is a matrix and $\boldsymbol{b}$ is a vector. Moreover, the matrix $\boldsymbol{A} = \breve{\boldsymbol{W}}^v_{\ell-1} + \boldsymbol{D}^\mathsf{T}\breve{\boldsymbol{W}}^u_\ell \boldsymbol{D}$ is a symmetric tridiagonal matrix. For that reason, we apply the tridiagonal system algorithm [63] to calculate $\boldsymbol{A}^{-1}\boldsymbol{b}$ with a complexity of $O(K)$ (instead of the $O(K^3)$ required by the Gaussian elimination for a non-tridiagonal matrix $\boldsymbol{A}$), thus significantly reducing the calculation time. The pseudo-code for this method is given in Algorithm 1.

3.4.2 Maximum likelihood using *von Mises* distribution

The LS-based methods for the second stage are greatly affected by the wrapping issue with the periodic variables such as the phase, IF, and GD. To address this issue, we propose an ML-based approach using a circular distribution, i.e., the *von Mises* distribution. In addition, the use of the *von Mises* distribution for the second stage makes it consistent with the first stage since, in the first stage, the IF/GD are also modeled using the same distribution.

We define a model as

$$\tilde{y} = \boldsymbol{d}^\mathsf{T}\boldsymbol{\psi} + \varepsilon, \tag{3.18}$$

where $\tilde{y}$ is the element of either $\tilde{\boldsymbol{V}}$ or $\tilde{\boldsymbol{U}}$, $\boldsymbol{\psi}$ is the flattened vector of the phase spectrogram $\boldsymbol{\Phi}$, $\boldsymbol{d}$ is a corresponding vector consisting of 0, 1, and -1 [similar to the matrix $\boldsymbol{D}$ in (3.3)], and ε is the residual of the model. Unlike the first stage that models the distribution of the IF/GD conditioned on the amplitude to train the DNN parameters, we define a *von Mises* distribution over $\tilde{y}$ conditioned on $\boldsymbol{d}$, and the phase $\boldsymbol{\psi}$ becomes the parameter of the model to be fitted. In other words, $p(\tilde{y}|\boldsymbol{d};\boldsymbol{\psi})$ is equal to a *von Mises* distribution, the measure of location of which is $\hat{y} = \boldsymbol{d}^{\mathsf{T}}\boldsymbol{\psi}$.

From (3.4), by taking the negative logarithm of $p(\tilde{y}|\boldsymbol{d};\boldsymbol{\psi})$ of all the elements of $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{U}}$ with the assumption of a constant concentration κ, we derive an error function for the whole phase spectrogram as

$$\begin{aligned}\mathcal{L}_{\mathrm{ML}}(\boldsymbol{\Phi}) = -\sum_{k,l}\Big(&W^u_{k,\ell}\cos\left(\tilde{U}_{k,\ell} - \hat{U}_{k,\ell}\right)\\ &+ W^v_{k,\ell}\cos\left(\tilde{V}_{k,\ell} - \hat{V}_{k,\ell}\right)\Big),\end{aligned} \tag{3.19}$$

where $W^u_{k,\ell}$ and $W^v_{k,\ell}$ are the weights of the GD and IF at TF bin (k,ℓ), respectively, which are empirically selected as $W^u_{k,\ell} = W^v_{k,\ell} = |X_{k,\ell}|$. Thanks to the cosine functions, (3.19) is not affected by the wrapping issue of $\tilde{U}_{k,\ell}$ and $\tilde{V}_{k,\ell}$ as with the LS-based methods.

The partial derivative of (3.19) is given by

$$\frac{\partial \mathcal{L}_{\mathrm{ML}}}{\partial \Phi_{k,\ell}} = \sin(\Phi_{k,\ell})C_{k,\ell} - \cos(\Phi_{k,\ell})S_{k,\ell}, \tag{3.20}$$

where

$$\begin{aligned}C_{k,\ell} = {}& W^v_{k,\ell}\cos(\Phi_{k,\ell+1} - \tilde{V}_{k,\ell}) + W^v_{k,\ell-1}\cos(\Phi_{k,\ell-1} + \tilde{V}_{k,\ell-1})\\ &+ W^u_{k,\ell}\cos(\Phi_{k+1,\ell} + \tilde{U}_{k,\ell}) + W^u_{k-1,\ell}\cos(\Phi_{k-1,\ell} - \tilde{U}_{k-1,\ell}),\end{aligned} \tag{3.21}$$

and $S_{k,\ell}$ is defined the same as $C_{k,\ell}$, in which all the cosine functions are replaced with sine functions. Note that, at the boundaries of $k = 0$, $k = K-1$, $\ell = 0$, and $\ell = L-1$, we remove the terms containing the indices of $k-1$, $k+1$, $\ell-1$, and $\ell+1$, respectively, from $C_{k,\ell}$ and $S_{k,\ell}$.

It is impossible to find the analytic solution of the equation setting the gradient vector of $\mathcal{L}_{\mathrm{ML}}(\boldsymbol{\Phi})$ to zero. On the basis of several properties of the error function, we propose the following two optimization approaches.

Using Newton's method

The first approach is to break (3.19) into frames so that the Hessian matrix becomes tridiagonal. Considering only the terms containing the phase at frame ℓ in $\mathcal{L}_{\text{ML}}(\mathbf{\Phi})$, the frame-wise error function is given by

$$\begin{aligned}\mathcal{L}_{\text{MLF}}(\boldsymbol{\phi}_\ell) = -\sum_k \Big(& W^u_{k,\ell} \cos\left(\tilde{U}_{k,\ell} - \hat{U}_{k,\ell}\right) \\ & + W^v_{k,\ell-1} \cos\left(\tilde{V}_{k,\ell-1} - \hat{V}_{k,\ell-1}\right) \\ & + W^v_{k,\ell} \cos\left(\tilde{V}_{k,\ell} - \hat{V}_{k,\ell}\right)\Big).\end{aligned} \tag{3.22}$$

The gradient vector $\nabla_{\boldsymbol{\phi}_\ell}\mathcal{L}_{\text{MLF}}(\boldsymbol{\phi}_\ell)$ can be calculated with the kth element identical to (3.20). We can see from (3.20) that the partial derivative with respect to $\Phi_{k,\ell}$ contains only two phase elements of the same frame, i.e., $\Phi_{k+1,\ell}$ and $\Phi_{k-1,\ell}$. Consequently, the Hessian matrix of $\mathcal{L}_{\text{MLF}}(\boldsymbol{\phi}_\ell)$, denoted as $\boldsymbol{H}$, is a symmetric tridiagonal matrix, the element on the main diagonal of which is given by

$$\frac{\partial^2 \mathcal{L}_{\text{MLF}}}{\partial \Phi^2_{k,\ell}} = \cos\left(\Phi_{k,\ell}\right) C_{k,\ell} + \sin\left(\Phi_{k,\ell}\right) S_{k,\ell}, \tag{3.23}$$

and the element on the first diagonal above (or below) is

$$\frac{\partial^2 \mathcal{L}_{\text{MLF}}}{\partial \Phi_{k,\ell} \partial \Phi_{k+1,\ell}} = -W^u_{k,\ell} \cos\left(\Phi_{k,\ell} - \Phi_{k+1,\ell} - \tilde{U}_{k,\ell}\right). \tag{3.24}$$

The tridiagonality of the Hessian matrix motivates us to use Newton's method to update the phase estimate. However, there is a problem that $\boldsymbol{H}$ is often not positive definite as $\mathcal{L}_{\text{MLF}}(\boldsymbol{\phi}_\ell)$ is periodic. To solve this problem, we apply a regularization strategy, as in a previous study [64], to update the phase estimate as

$$\hat{\boldsymbol{\phi}}_\ell \leftarrow \hat{\boldsymbol{\phi}}_\ell - (\boldsymbol{H} + \gamma \boldsymbol{I}_K)^{-1} \nabla_{\boldsymbol{\phi}_\ell} \mathcal{L}_{\text{MLF}}(\hat{\boldsymbol{\phi}}_\ell), \tag{3.25}$$

where γ is a damping factor. $\gamma = 0$ is equivalent to no regularization. When γ is large, $\boldsymbol{H}$ is dominated by $\gamma \boldsymbol{I}_K$, and (3.25) approximates the standard gradient descent with the updating rate of $1/\gamma$. Ideally, γ is adaptive so that it is large enough to offset the negative eigenvalues of $\boldsymbol{H}$. We calculate γ from the smallest eigenvalue λ of $\boldsymbol{H}$ for each update as

$$\gamma = \begin{cases} -\beta\lambda, & \text{if } \lambda < 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.26}$$

Algorithm 2 Pseudo-code of ML-based method using Newton's method for reconstructing phase from IF and GD.

Input: Amplitude spectrogram $|\boldsymbol{X}|$, estimated IF $\tilde{\boldsymbol{V}}$ and GD $\tilde{\boldsymbol{U}}$, number of iterations N_1 and N_2

Output: Phase spectrogram $\hat{\boldsymbol{\Phi}}$

$\hat{\Phi}_{0,0} \leftarrow 0$

$\hat{\Phi}_{k,0} \leftarrow \hat{\Phi}_{k-1,0} - \tilde{U}_{k-1,0}$, for $k \in \{1, \dots, K-1\}$

for $\ell \in \{1, \dots, L-1\}$ **do**

  $\hat{\boldsymbol{\phi}}_\ell \leftarrow \hat{\boldsymbol{\phi}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1}$

  for $i \in \{1, \dots, N_1\}$ **do**

    Update $\hat{\boldsymbol{\phi}}_\ell$ as in (3.25) removing terms containing $\Phi_{k,\ell+1}$ from $C_{k,\ell}$ and $S_{k,\ell}$

for $i \in \{1, \dots, N_2\}, \ell \in \{0, \dots, L-1\}$ **do**

  Update $\hat{\boldsymbol{\phi}}_\ell$ as in (3.25)

where β is a scaling constant. We can efficiently estimate only the smallest eigenvalue of the tridiagonal matrix $\boldsymbol{H}$ as in [65]. As $\boldsymbol{H} + \gamma \boldsymbol{I}_K$ is also tridiagonal, the complexity of (3.25) can be reduced from $O(K^3)$ to $O(K)$ by using the tridiagonal system algorithm, similar to what was mentioned in Section 3.4.1.

In the update (3.25) for $\boldsymbol{\phi}_\ell$, the phase at the next and previous frames for calculating $C_{k,\ell}$ and $S_{k,\ell}$ are replaced with their estimates, i.e., $\hat{\boldsymbol{\phi}}_{\ell+1}$ and $\hat{\boldsymbol{\phi}}_{\ell-1}$. However, those estimates are not available at the beginning. We found that a random initialization may lead to slow convergence and poor results. Therefore, for the first several iterations, we recursively reconstruct the phase $\boldsymbol{\phi}_\ell$ by using only $\hat{\boldsymbol{\phi}}_{\ell-1}$. In other words, we remove the terms containing $\Phi_{k,\ell+1}$ from $C_{k,\ell}$ and $S_{k,\ell}$ when calculating (3.25). This is equivalent to removing the terms of $W^v_{k,\ell} \cos\left(\tilde{V}_{k,\ell} - \hat{V}_{k,\ell}\right)$ from the error function (3.22). The full version of (3.25) is then used to smooth the phase estimate, i.e., $\boldsymbol{\phi}_\ell$ is updated using both $\hat{\boldsymbol{\phi}}_{\ell-1}$ and $\hat{\boldsymbol{\phi}}_{\ell+1}$. The pseudo-code for this method is given in Algorithm 2.

Using coordinate descent

Another approach for minimizing (3.19) is based on its separability property. From (2.2) and (2.3), we can see that each phase element $\Phi_{k,\ell}$ is only present in at most four terms in the sum

Algorithm 3 Pseudo-code of ML-based method using coordinate descent for reconstructing phase from IF and GD.

Input: Amplitude spectrogram $|\boldsymbol{X}|$, estimated IF $\tilde{\boldsymbol{V}}$ and GD $\tilde{\boldsymbol{U}}$, number of iterations N_1 and N_2

Output: Phase spectrogram $\hat{\boldsymbol{\Phi}}$

$\hat{\Phi}_{0,0} \leftarrow 0$

$\hat{\Phi}_{k,0} \leftarrow \hat{\Phi}_{k-1,0} - \tilde{U}_{k-1,0}$, for $k \in \{1, \ldots, K-1\}$

for $\ell \in \{1, \ldots, L-1\}$ **do**

$\hat{\boldsymbol{\phi}}_\ell \leftarrow \hat{\boldsymbol{\phi}}_{\ell-1} + \tilde{\boldsymbol{v}}_{\ell-1}$

for $i \in \{1, \ldots, N_1\}, k \in \{0, \ldots, K-1\}$ **do**

Update $\hat{\Phi}_{k,\ell}$ as in (3.27) removing terms containing $\Phi_{k,\ell+1}$ from $C_{k,\ell}$ and $S_{k,\ell}$

for $i \in \{1, \ldots, N_2\}, \ell \in \{0, \ldots, L-1\}, k \in \{0, \ldots, K-1\}$ **do**

Update $\hat{\Phi}_{k,\ell}$ as in (3.27)

of $\mathcal{L}_{\text{ML}}$, i.e., the terms containing $\hat{V}_{k,\ell}$, $\hat{V}_{k,\ell-1}$, $\hat{U}_{k,\ell}$, and $\hat{U}_{k-1,\ell}$. In other words, varying $\Phi_{k,\ell}$ will change only those terms and will not have much of an effect on the optimal states of other phase elements in $\mathcal{L}_{\text{ML}}$. Therefore, we use a coordinate-descent strategy [66] that sequentially minimizes $\mathcal{L}_{\text{ML}}$ for each $\Phi_{k,\ell}$ where all other phase elements are fixed.

As $S_{k,\ell}$ and $C_{k,\ell}$ are independent from $\Phi_{k,\ell}$, we can easily set the first derivative $\partial\mathcal{L}_{\text{ML}}/\partial\Phi_{k,\ell}$ to zero and check the second derivative to find the minimum. The solution is

$$\hat{\Phi}_{k,\ell} = \begin{cases} \arctan\left(S_{k,\ell}/C_{k,\ell}\right), & \text{if } f'' > 0 \\ \arctan\left(S_{k,\ell}/C_{k,\ell}\right) + \pi, & \text{otherwise} \end{cases}, \tag{3.27}$$

where $f'' = \partial^2\mathcal{L}_{\text{ML}}/\partial\Phi_{k,\ell}^2$, which is identical to (3.23).

(3.27) is sequentially calculated throughout the whole spectrogram. Because the update of $\hat{\Phi}_{k,\ell}$ affects the optimal states of other phase elements around it, we need several iterations to make all the elements converge. Similar to the approach using Newton's method, for the first several iterations, we remove from $C_{k,\ell}$ and $S_{k,\ell}$ the terms containing $\Phi_{k,\ell+1}$ when calculating the solution of (3.27). The pseudo-code for this coordinate-descent approach is illustrated in Algorithm 3.

3.4.3 Comparison of methods for second stage

In this subsection, this thesis presents several theoretical comparisons among the methods for the second stage of two-stage phase reconstruction algorithms, i.e., the conventional LS-based method, conventional circular average-based method, the proposed weighted LS-based method and ML-based method with the two optimization schemes.

Least squares [11] and weighted least squares

The proposed weighted LS-based method differs from the conventional LS-based method only in terms of the weights in the error function, which results in changes in the calculation of the solution, especially the matrix inversion. In the solution (3.10) of the conventional LS-based method, we can calculate the inverse of the matrix $(\boldsymbol{I}_K + \boldsymbol{D}^\mathsf{T}\boldsymbol{D})$ in advance as it is constant. For our weighted LS-based method, the matrix $(\boldsymbol{W}_{\ell-1}^v + \boldsymbol{D}^\mathsf{T}\boldsymbol{W}_\ell^u\boldsymbol{D})$ in (3.17) depends on the weights; hence, its inverse must be computed for each frame. However, thanks to the tridiagonality property of the matrix, (3.17) can be calculated with a complexity of $O(K)$, which is the same as (3.10).

Least squares and maximum likelihood

As the LS-based methods can also be interpreted as the ML, LS- and ML-based methods differ in the distributions used, i.e., Gaussian and *von Mises* distributions. The *von Mises* distribution seems to be more efficient as it handles the wrapping issue and is the same distribution used in the first stage. The LS-based method is greatly affected by the wrapping issue. Although (3.9) is used to mitigate this issue, it may not be reliable when the errors of $\tilde{U}$ and/or $\tilde{V}$ are high. However, the advantage of the LS-based methods is that they yield a unique solution, while the ML-based methods require iterative methods for the optimization.

Circular average [12] and maximum likelihood

Like our *von Mises* distribution-based ML-based method, the circular average-based method is not affected by the wrapping issue. However, it consists of a single pass of recursively calculating each phase element using the average operation. This makes the phase estimate at each TF bin heavily dependent on the IF, GD, and other previously estimated phase elements

nearby. Therefore, the circular average-based method may not be efficient when the estimated IF and GD have high errors, or when the underlying components of the signal are not stable. In contrast, our ML-based methods define a solid optimization problem, which can be solved using various optimization techniques. Although for the first several iterations, our ML-based methods also reconstruct the phase recursively, the later iterations help to compromise the IF and GD errors at all TF bins, thus smoothing the phase estimate. Regarding the calculation speed, the circular average-based method has the same complexity as one iteration of our ML-based methods, which is $O(K)$.

Maximum likelihood using Newton's method and coordinate descent

Our ML-based methods using Newton's method (MLN) and using coordinate descent (MLC) are two different strategies to solve the same optimization problem: MLN breaks the error function into frames, while MLC breaks it into elements. The update scope of MLN seems to be more advanced than that of MLC as it modifies more elements at the same time. However, with the use of amplitude weights, the high-amplitude TF bins may restrict the change of the low-amplitude ones when they are updated simultaneously. An advantage of MLC is that it does not require tuning the parameters such as β in MLN. Regarding the calculation speed, although one iteration of both MLN and MLC has the complexity of $O(K)$, MLN is slower due to the eigenvalue estimation.

3.4.4 IFPD for two-stage phase reconstruction

In two-stage phase reconstruction algorithms, the GD is used to maintain the phase relationship along frequency. However, as a difference between two consecutive phase elements, the GD can only preserve the local relationship. To enhance the relationship between two distant phase elements along the frequency, the IFPD can be utilized. In the first stage, the IFPD with various hops is reconstructed from the amplitude using DNNs, similar to the IF and GD. In the second stage, we penalize the IFPD errors in addition to the IF and GD errors in the loss function for reconstructing the phase. Because the IFPD is also wrapped, which may aggravate the wrapping issue in the LS-based method, the ML-based method is used. With the IFPD, the error function

(3.19) becomes

$$\mathcal{L}_{\text{ML\_IFPD}}(\boldsymbol{\phi}_\ell) = -\sum_{k,\ell}\Big(\sum_{i\in\boldsymbol{S}} W_{k,\ell}^{u^{(i)}}\cos(\tilde{U}_{k,\ell}^{(i)} - \hat{U}_{k,\ell}^{(i)}) + W_{k,\ell}^{v}\cos(\tilde{V}_{k,\ell} - \hat{V}_{k,\ell})\Big), \tag{3.28}$$

where $\boldsymbol{S}$ is the set of the frequency hops used for calculating the IFPD, including $i = 1$ for the GD. The weight for $U_{k,\ell}^{(i)}$ is defined as

$$W_{k,\ell}^{u^{(i)}} = \alpha^{(i)}|X_{k,\ell}|, \tag{3.29}$$

where the scalar $\alpha^{(i)}$ is used to adjust the contribution of $U_{k,\ell}^{(i)}$ in the error function.

Because the use of the IFPD makes the Hessian matrix no longer tridiagonal, which increases the computational complexity of Newton's method, (3.28) is minimized using the coordinate-descent strategy. The solution of $\hat{\Phi}_{k,\ell}$ for minimizing $\mathcal{L}_{\text{ML\_IFPD}}$ is the same as (3.27), except that we include terms containing $\hat{U}_{k,\ell}^{(i)}$ to the calculation of $C_{k,\ell}$ and $S_{k,\ell}$. We found that errors of the IFPD estimated in the first stage, $\tilde{U}_{k,\ell}^{(i)}$, become higher as the frequency hop i increases. This means that the reconstructed phase will get these errors if it is completely fitted with those IFPD estimates. Therefore, we only minimize the error function (3.28) with the IFPD for the first several iterations. After that, (3.19) is used to smooth the phase estimates with only the IF and GD.

In summary, the algorithm for reconstructing the phase with the IFPD is the same as Algorithm 3 with two modifications, i.e., the IFPD is required as an input, and, in the inner for-loop of the first for-loop, $\hat{\Phi}_{k,\ell}$ is updated to minimize $\mathcal{L}_{\text{ML\_IFPD}}$ instead of $\mathcal{L}_{\text{ML}}$. As discussed above, the update $\hat{\Phi}_{k,\ell}$ in the last for-loop in Algorithm 3 is still for minimizing $\mathcal{L}_{\text{ML}}$.

3.5 Experiments and results

3.5.1 Experimental setup

Experiments are conducted to evaluate the performances of two-stage phase reconstruction algorithms. All such algorithms share the same IF and GD estimated in the first stage. The methods used for the second stage include the conventional LS-based method [11] (`LS`), conventional circular average-based method [12] (`AVG`), the proposed weighted LS-based method (`WLS`), and

the proposed ML-based methods with 30 iterations using Newton's method (`MLN`; $N_1 = 10$ and $N_2 = 20$) and coordinate descent (`MLC`; $N_1 = 5$ and $N_2 = 25$). The proposed algorithm using the IFPD (`ML+IFPD`) was also evaluated with 30 iterations ($N_1 = 5$ and $N_2 = 25$). In addition, this thesis included conventional non-two-stage phase reconstruction algorithms for comparison. These are the Griffin-Lim method [28] with 100 iterations (`GL`), the phase gradient heap integration method [36] (`PGHI`), and the iterative method using alternating direction method of multipliers [34] with 100 iterations (`ADMMGLA`).

The data used for training were from the training set of the TIMIT dataset [67]. The sampling rate is $16\,\mathrm{kHz}$. The tests were performed on 300 samples (150 males and 150 females) randomly selected from the test set of the TIMIT dataset.

In the implementation, the STFT was calculated using a Hamming window with a 32-ms length, 4-ms shift, and 512-point DFT. To reconstruct the IF, GD, and IFPD in the first stage, we used fully connected feedforward DNNs with 4 hidden layers, each layer containing 1024 gated tanh units [68], and the last layer containing linear units. This DNN architecture is similar to those in [11, 12, 69]. In addition, the authors of [69] claimed in their work that, based on their experiments, there was no difference between the gated layers and LSTM (long short-term memory) layers in terms of the quality of the reconstructed speech. For these reasons, we decided to use this DNN architecture. It is worth noting that a separate DNN is used to estimate each of the IF, GD, and IFPD. The input of the DNN was joint vectors consisting of the log amplitude at the current and ± 2 frames and was normalized to zero mean and unit variance. The output of the DNN was one frame of the phase feature (IF, GD, or IFPD), which was also normalized using (3.7) for the IF and (3.14) for the GD and IFPD. These models were trained using the Adam optimizer for 400 epochs. The parameters of the methods in the second stage were determined by fine-tuning, which are described as follows. The power p in `WLS` was set to 10. The weight β in `MLN` was set to 2.4. For `ML+IFPD`, we used a set of six frequency hops of $\boldsymbol{S} = \{1, 2, 3, 4, 5, 6\}$ with the corresponding set of weights $\alpha^{(i)}$ of $\{1.0, 0.4, 0.3, 0.2, 0.1, 0.1\}$. The Linear Algebra Package (LAPACK) [70] was used for the tridiagonal system algorithm and eigenvalue estimation.

For the objective metrics, we measured the perceptual evaluation of speech quality (PESQ) [71] and short-time objective intelligibility (STOI) [72] of the reconstructed signals. The higher those scores, the higher the quality of the signal. We also calculated the consistency measure [33]

as

$$C(\hat{\boldsymbol{X}}) = 10\log_{10}\frac{\|\hat{\boldsymbol{X}} - \text{STFT}(\text{ISTFT}(\hat{\boldsymbol{X}}))\|^2}{\|\hat{\boldsymbol{X}}\|^2}, \tag{3.30}$$

where $(\hat{\boldsymbol{X}})_{k,\ell} = |X_{k,\ell}|e^{j\hat{\Phi}_{k,\ell}}$. The consistency measure indicates how much the phase spectrogram is consistent with the amplitude spectrogram, which is expected to be low.

To further compare the subjective performances among the two-stage algorithms, we conducted the BS.1116 test [73] using webMUSHRA [74], which is a web-based listening test framework. In each BS.1116 trial, the subject is presented with three stimuli labeled `A`, `B`, and `C`. `A` is always the reference (original signal), while `B` and `C` are randomly assigned by the hidden reference and reconstructed signal. The subject is asked to assess the impairments (if any) on `B` and `C` compared to `A` using a continuous 5-grade scale with anchors defined as

- (5.0) Imperceptible,
- (4.0) Perceptible, but not annoying,
- (3.0) Slightly annoying,
- (2.0) Annoying,
- (1.0) Very annoying.

Because one of `B` and `C` is identical to `A`, there must be at least one point of 5.0. As a general rule, if a subject rates the hidden reference with a score of less than 5.0 for more than 15% of the test, all the results of this subject will not be considered. The samples presented to each subject are randomly selected from the test set, in which the number of samples depends on the subject's demand (maximum 15 samples per subject, corresponding to 105 trials for 7 methods). The subjects participating in the test were all students ranging from 20 to 30 in age. Apart from the results of 3 subjects excluded by the test rules, 245 samples (which may be duplicated) were tested by 20 subjects in total.

3.5.2 Results

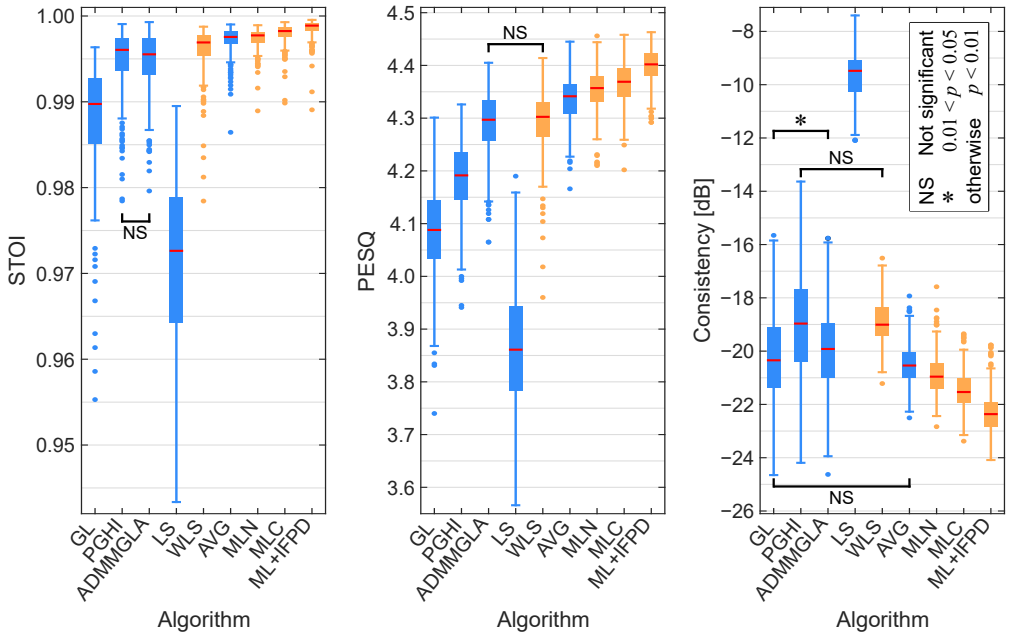
Table 3.1 lists the errors of the DNNs in the first stage, in which an error is defined as

$$\epsilon(\boldsymbol{Y}, \tilde{\boldsymbol{Y}}) = 1 - \frac{1}{KL}\sum_{k,\ell}\cos(Y_{k,\ell} - \tilde{Y}_{k,\ell}). \tag{3.31}$$

Table 3.1: Errors of DNNs in first stage

	IF	GD	IFPD				
			$i=2$	$i=3$	$i=4$	$i=5$	$i=6$
Training	0.133	0.231	0.290	0.432	0.530	0.729	0.717
Testing	0.148	0.245	0.307	0.450	0.575	0.804	0.811

Note: Error range is $[0, 2]$



(a) STOI (b) PESQ (c) Consistency

Figure 3.6: Objective scores of phase reconstruction algorithms, where blue and red respectively indicate conventional and proposed methods.

$\epsilon(\boldsymbol{Y}, \tilde{\boldsymbol{Y}})$ is similar to $\mathcal{L}_{\text{DNN}}$ in (3.6), however, (3.31) is for the whole spectrogram, while (3.6) is for each frame. The error range is $[0, 2]$. We can see from Table 3.1 that the higher the frequency hop i, the higher the errors of the DNNs for reconstructing the IFPD. The reason is most likely because, when the frequency hop is large, the connections between TF bins are weak due to the low side lobes of the window function. In such a case, the IFPD becomes less structured, hence, more difficult to estimate.

Fig. 3.6 shows the STOI, PESQ, and consistency measure of the reconstructed signals of the phase reconstruction algorithms. The results were analyzed with the paired sample t-test, which

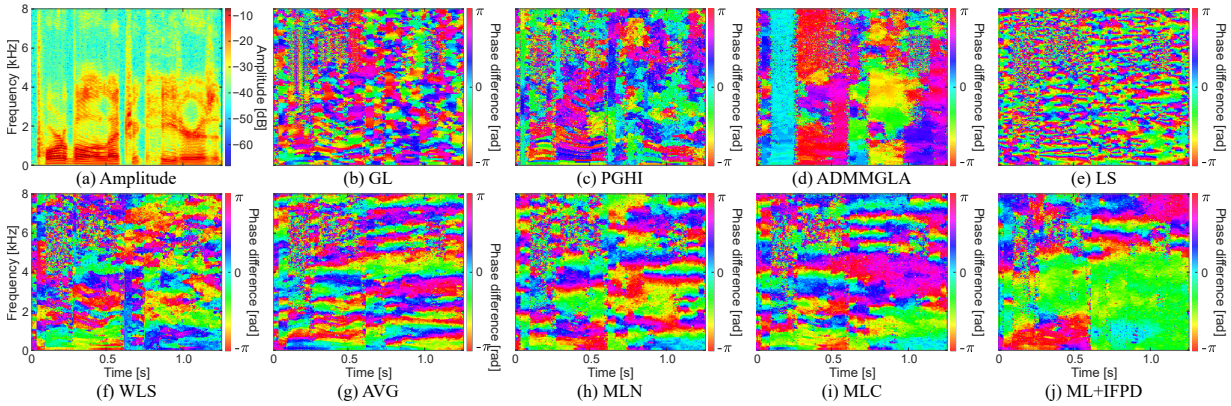


Figure 3.7: Examples of (a) log-amplitude, and (b)–(j) phase differences between original phase $\mathbf{\Phi}$ and estimated phase $\hat{\mathbf{\Phi}}$.

shows that the differences between the scores are statistically significant with a few exceptions. It can be seen from Fig. 3.6 that the two-stage methods, except the LS-based methods, performed better than the conventional non-two-stage methods, in which the `ML+IFPD` yielded the highest results. By using the IFPD in addition to the IF and GD for only several first iterations, `ML+IFPD` improved the results of `MLC` for all metrics. Although both `MLN` and `MLC` minimize the same error function with the same number of iterations, `MLC` achieved better results than `MLN`. A possible reason is that the update of `MLN` is an approximation while `MLC` directly solves the equation setting the derivative to zero. Although the solution in `MLC` is local, the separability of the error function motivates it. We can also see from Fig. 3.6 that, by adding the amplitude weights, `WLS` significantly improved the results of its baseline method `LS`. However, `WLS` was still worse than `AVG` and our ML-based methods. This is most likely due to the LS-based methods being affected by the wrapping issue.

Fig. 3.7 shows an example of the phase differences between the original and estimated phases, i.e., $\mathcal{P}(\mathbf{\Phi} - \hat{\mathbf{\Phi}})$. We may expect that the phase difference spectrogram has large regions of the same color, at which $\mathbf{\Phi}$ and $\hat{\mathbf{\Phi}}$ change at the same rates. In other words, the phase relationships in those regions are preserved, even if the absolute values of the phase are changed. In addition, we only focus on the high-amplitude regions since the phases of low-amplitude TF bins have little effect on the ISTFT. We can see from Fig. 3.7(b) and (e) that the same-color regions in the phase-difference spectrograms of `GL` and `LS` are small. At the boundaries of those regions, the phase relationships are distorted, resulting in impairments in the amplitude and

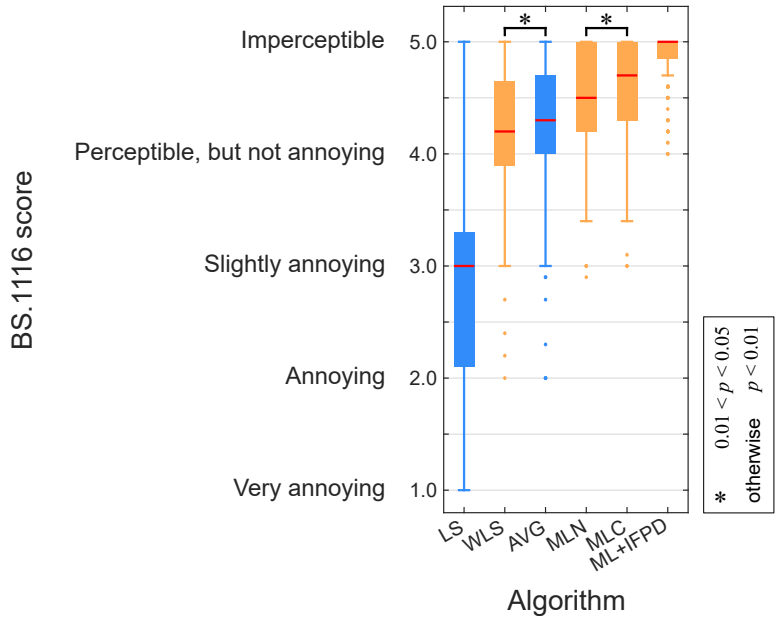


Figure 3.8: Subjective scores of phase reconstruction algorithms.

modulations in the frequency of the reconstructed time-domain signals. As a consequence, the signals estimated using `GL` and `LS` often contains artifacts, such as reverberation and buzz. The same-color regions at high-amplitude TF bins became larger for other methods. Especially, by using the IFPD, `ML+IFPD` clearly improved the phase relationships between TF bins, illustrated with the smooth phase difference spectrogram in Fig. 3.7(j). This finding in this example is reflected in, and consistent with, the objective results.

Fig. 3.8 illustrates the results of the BS.1116 test for the two-stage algorithms, i.e., the measures of perceptual impairments on the reconstructed signals compared with the original signal. The subjective scores in Fig. 3.8 expose a similar trend to the objective scores in Fig. 3.6, in which `ML+IFPD` surpassed other methods. The differences between the scores are also confirmed with the small p-values of the paired sample t-test.

The above observations confirmed that `ML+IFPD` outperforms the other methods. In addition, for the second stage of two-stage phase reconstruction algorithms, the ML-based methods are better than the LS-based and circular average-based methods. The experimental results also indicate the efficacy of using amplitude weights in improving the conventional LS-based method.

Although achieving high objective and subjective scores, a limitation of the two-stage ap-

proaches is that the waveform may differ from the original signal as we focus on the phase relationship between the TF bins but not the absolute value of the phase. This is a common problem for phase reconstruction when only the amplitude information is available [47]. In other applications, when the noisy/mixed phases are available, they can be used as an initialization for the proposed methods to reduce the problem. Another limitation of the proposed methods is that they require multiple models to estimate the phase features in the first stage, which may be a drawback in real-time applications. A possible solution is to use multitask learning, i.e., to combine all the models in the first stage into one model with multiple outputs.

3.6 Conclusion

In this chapter, this thesis presented two approaches for the second stage of two-stage phase reconstruction algorithms. The first method is to add the amplitude weights to a conventional LS-based method. The second method is based on the ML with the *von Mises* distribution, which is optimized using the regularized Newton's method and coordinate descent. In the theory discussion, this thesis analyzed the GD properties and introduced a GD-normalization formula by compensating for the phase shift introduced by the commonly used zero-starting window function. This thesis also investigated the roles of the phase relationships between TF bins in the ISTFT. On the basis of the analysis, this thesis applied the IFPD to the phase reconstruction. Both objective and subjective experiments showed that the performance of our ML-based method using the IFPD is superior to other methods that use only the IF and GD. The results also suggest that ML-based methods perform better than other methods in the second stage, and the use of amplitude weights significantly improves the results of the conventional LS-based method.

Chapter 4

DNN-based online phase reconstruction

4.1 Introduction

In addition to the offline processing, online phase reconstruction is desired in many application, such as low-latency audio source separation [75] and incremental text-to-speech [76]. Many phase reconstruction algorithms require iteration or future frame information to estimate the current-frame phase, which may only be feasible offline. For real-time settings, some modifications have been made. [77] proposed a real-time version of the Griffin–Lim algorithm, called the real-time spectrogram inversion algorithm (RTISI), which iteratively reconstructs the signal frame-by-frame with an effective initialization scheme. In a non-iterative manner, the single-pass spectrogram inversion (SPSI) [78] utilizes a phase-locking technique related to a phase vocoder. [79] proposed a real-time adaptation of the PGHI (i.e., RTPGHI) with one or even zero look-ahead frames. Although these methods have achieved promising results, they are still suboptimal due to some approximations used, e.g., the harmonic model assumption in the SPSI and the phase derivative approximation in the RTPGHI. Among phase reconstruction approaches, DNN-based methods have significant potential for real-time applications, as they can be easily adapted by using a causal model. Additionally, DNNs have a robust modeling capability to learn the underlying structure of the target signals.

However, most of the conventional DNN-based phase reconstruction methods do not consider the distinct properties of the phase at different TF bins. In the inverse STFT (ISTFT), the amplitude acts as a scaling factor for the TF bins, and the phase determines their relative

position, thus ensuring their proper combination. If the amplitude is low, regardless of the values of the phase, the contribution of the TF bin to the reconstructed signal will be small. Training a model with the phase at these low-amplitude bins may not bring much benefit and might even restrict the model from learning useful information in the high-amplitude bins. Another property of the phase is that, unlike the amplitude, its values depend linearly on the frequency. At high frequencies, the phase changes quickly along the time, leading to instability. These unstable phase elements usually yield high errors, which may impede the model from fitting the more stable phase elements at low frequencies.

Taking into account the varying properties of the phase, the aim of this chapter is to improve DNN-based methods for real-time phase reconstruction from the amplitude, including proposing new loss functions and data augmentation scheme. Starting with the *von Mises* distribution-based loss functions as in [41] and [80], this thesis imposes weights on the phase loss with respect to frequency to control the effect of unstable phase elements at high frequencies. This thesis also leverages amplitude weights to separate the importance of the phase at different TF bins. In addition, the IFPD is included in the loss function to improve the connection of phase elements along the frequency. The proposed loss functions are utilized to train a causal DNN architecture for real-time applications. To improve the generalization of the models, the training data is augmented by randomly shifting the signals in the time domain before calculating the STFT for each training epoch.

The remainder of this chapter is organized as follows. The conventional DNN-based phase reconstruction is reviewed in Section 4.2. The proposed methods are then described in Section 4.3. In Section 4.4, this thesis discusses the experiments and presents the results. Finally, this thesis concludes the chapter in Section 4.5.

4.2 Conventional loss functions for DNN-based phase reconstruction

To handle the periodicity of the phase, [41] modeled the phase using the *von Mises* distribution, which is a circular distribution. With the assumption of a constant κ, the phase loss function is defined as

$$\mathcal{L}_{\mathrm{p}}(\mathbf{\Phi}, \hat{\mathbf{\Phi}}) = -\sum_{k,\ell} \mathcal{C}^{\mathrm{p}}_{k,\ell} \triangleq -\sum_{k,\ell} \cos(\Phi_{k,\ell} - \hat{\Phi}_{k,\ell}). \tag{4.1}$$

The modeling and derivation of the loss function has been defined in Chapter 3.

With the observation that the GD has a similar structure to the amplitude spectrogram, [41] utilized the GD to improve the performance of the phase reconstruction algorithm. By modeling the GD with the *von Mises* distribution, [41] introduced a multitask-learning loss function for phase reconstruction, which can be expressed as

$$\mathcal{L}_{\mathrm{pgd}}(\mathbf{\Phi}, \hat{\mathbf{\Phi}}) = -\sum_{k,\ell} \left(\lambda_{\mathrm{p}} \mathcal{C}^{\mathrm{p}}_{k,\ell} + \lambda_{\mathrm{gd}} \mathcal{C}^{\mathrm{gd}}_{k,\ell} \right), \tag{4.2}$$

where

$$\mathcal{C}^{\mathrm{gd}}_{k,\ell} = \cos(U_{k,\ell} - \hat{U}_{k,\ell}), \tag{4.3}$$

and λ_{p} and λ_{gd} are the weights for the loss components.

4.3 Proposed phase reconstruction

This section introduces several improvements to the DNN-based phase reconstruction. Specifically, the *Von Mises* mixture model is used to mitigate the sign indetermination problem in Section 4.3.1. Loss functions that incorporate weights are presented in Section 4.3.2 and the IFPD in used in Section 4.3.3. In addition, the data augmentation scheme for training the DNN is described in Section 4.3.4.

4.3.1 *Von Mises* mixture model-based loss function

An idea for mitigating the sign indetermination problem is to reconstruct the phase of either $x(n)$ or $-x(n)$. The phases of $x(n)$ and $-x(n)$ have a difference of π and can be modeled by a *von Mises* mixture model with two mixture components, as

$$\mathcal{F}(\Phi_{k,\ell}|\mu, \kappa) = \frac{1}{2} f(\Phi_{k,\ell}|\mu, \kappa) + \frac{1}{2} f(\Phi_{k,\ell}|\mu + \pi, \kappa). \tag{4.4}$$

Assuming $\kappa = 1$, the *von Mises* mixture model-based loss function is defined using maximum likelihood, as

$$\begin{aligned} \mathcal{L}_{\mathrm{vmm}}(\mathbf{\Phi}, \hat{\mathbf{\Phi}}) &= -\sum_{k,\ell} \mathcal{C}^{\mathrm{vmm}}_{k,\ell} \\ &\triangleq -\sum_{k,\ell} \log \left(e^{\cos(\Phi_{k,\ell} - \hat{\Phi}_{k,\ell})} + e^{-\cos(\Phi_{k,\ell} - \hat{\Phi}_{k,\ell})} \right). \end{aligned} \tag{4.5}$$

A problem in training DNNs with $\mathcal{L}_{\text{vmm}}(\mathbf{\Phi}, \hat{\mathbf{\Phi}})$ is that the reconstructed phase may be inconsistent. Specifically, some phase elements may converge to the phase of $x(n)$ while others to the phase of $-x(n)$. To ensure a consistent reconstructed phase, this thesis uses the IF and GD losses to enhance the dependencies of phase elements along time and frequency. The loss function combining IF and GD losses is defined as

$$\mathcal{L}_{\text{vmmifgd}}(\mathbf{\Phi}, \hat{\mathbf{\Phi}}) = -\sum_{k,\ell} \left(\lambda_{\text{vmm}} \mathcal{C}_{k,\ell}^{\text{vmm}} + \lambda_{\text{if}} \mathcal{C}_{k,\ell}^{\text{if}} + \lambda_{\text{gd}} \mathcal{C}_{k,\ell}^{\text{gd}} \right), \tag{4.6}$$

where

$$\mathcal{C}_{k,\ell}^{\text{if}} = \cos(V_{k,\ell} - \hat{V}_{k,\ell}), \tag{4.7}$$

and λ_{vmm} and λ_{if} are the weights for the loss components.

4.3.2 Weighted loss functions

At high frequencies, we utilize the *von Mises* mixture model-based phase loss, $\mathcal{L}_{\text{vmm}}$, to mitigate the sign indetermination problem. For low frequencies, we use the *von Mises* distribution-based phase loss, $\mathcal{L}_{\text{p}}$, which we have found to be effective through empirical testing. Weights are incorporated into the loss function as follows.

Frequency-specific weights

Instead of using a fixed weight for all frequency bins, this thesis utilizes weights that vary in accordance with frequencies to control the impact of $\mathcal{L}_{\text{p}}$ and $\mathcal{L}_{\text{vmm}}$ on the loss function. The loss function with frequency-specific weights is defined as

$$\begin{aligned} \mathcal{L}_{\text{fw}}(\mathbf{\Phi}, \hat{\mathbf{\Phi}}) = -\sum_{k,\ell} \Big(& \alpha_{\text{p}}(k) \mathcal{C}_{k,\ell}^{\text{p}} + \alpha_{\text{vmm}}(k) \mathcal{C}_{k,\ell}^{\text{vmm}} \\ & + \lambda_{\text{if}} \mathcal{C}_{k,\ell}^{\text{if}} + \lambda_{\text{gd}} \mathcal{C}_{k,\ell}^{\text{gd}} \Big), \end{aligned} \tag{4.8}$$

where $\alpha_{\text{p}}(k)$ and $\alpha_{\text{vmm}}(k)$ are the weights of $\mathcal{C}_{k,\ell}^{\text{p}}$ and $\mathcal{C}_{k,\ell}^{\text{vmm}}$, respectively. The preliminary idea for the weights is illustrated in Fig. 4.1, in which $\mathcal{C}_{k,\ell}^{\text{p}}$ is used at low frequencies with high weights while $\mathcal{C}_{k,\ell}^{\text{vmm}}$ is assigned low weights to reduce the effect of unstable phase elements at high frequencies. k_1 and k_2 are the boundary frequencies for the phase losses. The weights for $\mathcal{C}_{k,\ell}^{\text{if}}$ and $\mathcal{C}_{k,\ell}^{\text{gd}}$ are constant because, unlike the phase, the values of the IF and GD are not linearly dependent on the frequency.

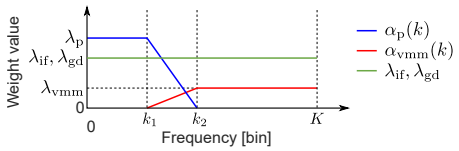


Figure 4.1: Illustration of weights of $\mathcal{L}_{\rm fw}$.

Amplitude weights

In addition to frequency-specific weights, amplitude weights are also used to emphasize the importance of high-amplitude TF bins. By incorporating amplitude weights, (4.8) becomes

$$\mathcal{L}_{\rm afw}(\boldsymbol{\Phi}, \hat{\boldsymbol{\Phi}}) = -\sum_{k,\ell} W_{k,\ell}\Big(\alpha_{\rm p}(k)\mathcal{C}^{\rm p}_{k,\ell} + \alpha_{\rm vmm}(k)\mathcal{C}^{\rm vmm}_{k,\ell} + \lambda_{\rm if}\mathcal{C}^{\rm if}_{k,\ell} + \lambda_{\rm gd}\mathcal{C}^{\rm gd}_{k,\ell}\Big), \tag{4.9}$$

where

$$W_{k,\ell} = \begin{cases} |X_{k,\ell}|, & \text{if } |X_{k,\ell}| < \gamma \\ \gamma, & \text{otherwise} \end{cases}, \tag{4.10}$$

and γ is the weight cutoff, which is used to reduce the gap between low and high amplitudes, thereby preventing the model from excessively fitting the phase at high-amplitude TF bins.

4.3.3 Integrating the IFPD to the loss function

Conventional loss functions utilize the GD loss to preserve the phase relationship across frequencies. However, as a phase difference between two consecutive bins as in (2.3), the GD may only capture the local relationship. Meanwhile, all frequency bins in a frame are interdependent because they are calculated from all the samples in the signal frame. To enhance the connections of the reconstructed phase elements along the frequency we integrate $U^{(i)}_{k,\ell}$ into the loss function as

$$\mathcal{L}_{\rm afw\_gd+}(\boldsymbol{\Phi}, \hat{\boldsymbol{\Phi}}) = -\sum_{k,\ell} W_{k,\ell}\Big(\alpha_{\rm p}(k)\mathcal{C}^{\rm p}_{k,\ell} + \alpha_{\rm vmm}(k)\mathcal{C}^{\rm vmm}_{k,\ell} + \lambda_{\rm if}\mathcal{C}^{\rm if}_{k,\ell} + \sum_{i\in\boldsymbol{S}}\lambda_{{\rm gd}(i)}\mathcal{C}^{{\rm gd}(i)}_{k,\ell}\Big), \tag{4.11}$$

where $\boldsymbol{S}$ is a set of frequency hops used to calculate $U^{(i)}_{k,\ell}$. $\mathcal{C}^{{\rm gd}(i)}_{k,\ell}$ is defined similarly to $\mathcal{C}^{\rm gd}_{k,\ell}$, and $\lambda_{{\rm gd}(i)}$ is its weight.

It is worth noting that the same technique can be applied to the IF to enhance phase relationships along the time. For the scope of this paper, we only consider the GD extension.

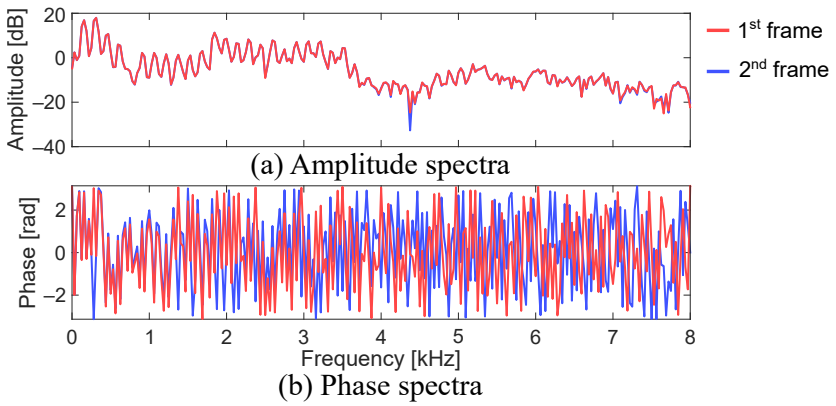


Figure 4.2: Example of two frames with shift of 1 sample.

4.3.4 Data augmentation

The phase is well-known to be sensitive to the waveform shift, as even a small shift of the signal in the time domain can lead to a significant change in the phase spectrogram. Fig. 4.2 illustrates the amplitude and phase spectra of two signal frames with the shift of 1 sample. We can see here that the phase spectra are very different while the amplitude spectra are almost the same. When training DNNs to estimate the spectral information, conventional methods usually calculate the STFT of the signal once and use it for every epoch. In other words, since the typical window shift is larger than 1 sample, if one frame in Fig. 4.2 is used for training, the other will be ignored, even though both frames contain useful information about variations of the phase.

To augment the training data, for each epoch, we randomly shift the signals by m samples before calculating the STFT. The shifted signal is defined as

$$x'(n) = x(n + m). \tag{4.12}$$

This is equivalent to removing the first m samples of the signal. The shift m is limited in $[0, R)$. If m is equal to the window shift R, frame ℓ of $x'(n)$ is identical to frame $(\ell + 1)$ of $x(n)$.

The augmentation can be extended in cases where high-resolution data are available. By shifting the signal before resampling it to the target sampling rate, we will be able to achieve a shift of less than 1 sample.

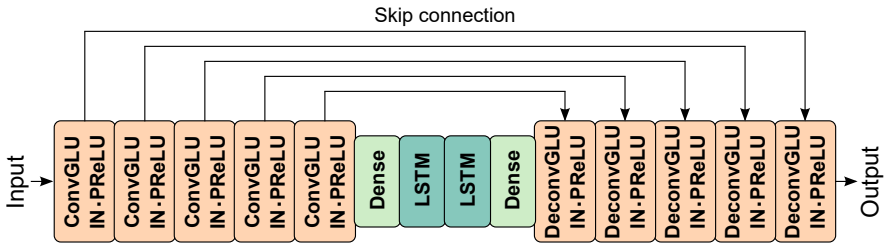


Figure 4.3: Diagram of convolutional recurrent network.

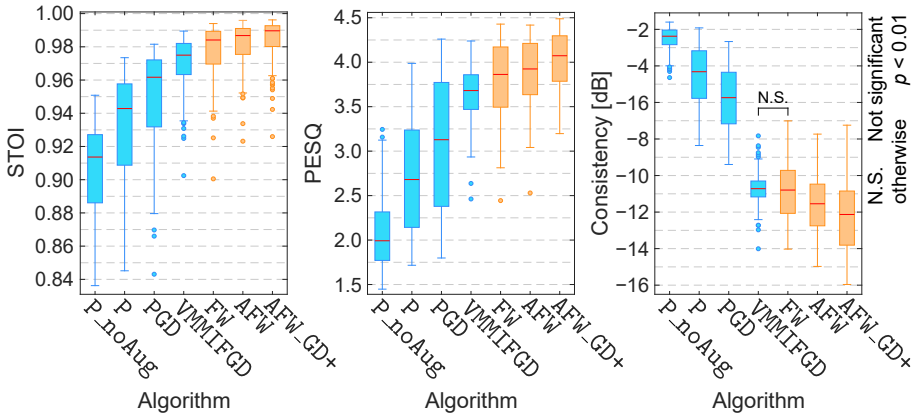
4.4 Experiments and results

4.4.1 Experimental setup

The experiment is divided into two parts. First, we compared the performances of the DNN-based phase reconstruction using the proposed loss functions $\mathcal{L}_{\text{fw}}$ (hereafter, `FW`), $\mathcal{L}_{\text{afw}}$ (hereafter, `AFW`), and $\mathcal{L}_{\text{afw\_gd+}}$ (hereafter, `AFW_GD+`) with the conventional loss functions $\mathcal{L}_{\text{p}}$ [41] (hereafter, `P`), $\mathcal{L}_{\text{pgd}}$ [41] (hereafter, `PGD`), and $\mathcal{L}_{\text{vmmifgd}}$ [80] (hereafter, `VMMIFGD`). For a fair comparison, the proposed data augmentation was applied to all methods. To evaluate the efficacy of the data augmentation scheme, we trained a model using $\mathcal{L}_{\text{p}}$ without augmenting the data (hereafter, `P_noAug`). In the second part of the experiment, we compared the proposed method, `AFW_GD+`, with other non-DNN real-time phase reconstruction methods including RTISI (hereafter, `RTISI`) [77], SPSI (hereafter, `SPSI`) [78], and RTPGHI (hereafter, `RTPGHI`) [79]. For these non-DNN methods, we set the number of look-ahead frames to zero so that they are all causal. We also included the offline version of `RTPGHI` (i.e., `PGHI` [81]) for comparison. PESQ, STOI and consistency measure are also used for evaluation metrics.

The training data were the training set of the TIMIT Acoustic-Phonetic Continuous Speech Corpus [67], which consists of recordings of 462 speakers of eight dialects of American English each reading ten sentences. The sampling rate is $16\,\text{kHz}$. The test was conducted on 300 samples (150 men and 150 women) randomly selected from the test set of TIMIT.

The weights were selected empirically for this preliminary work. We fixed λ_{if} and λ_{gd} to 1.0 and then varied the other weights for several values around 1.0. As a result, for all loss functions, λ_{p} was set to 1.0, and λ_{vmm} was set to 0.1. The boundary frequencies k_1 and k_2 were set to 25 and 100, respectively. After normalizing the speech signals to the active level [82] of $-30\,\text{dB}$, the cutoff γ was set to 0.07. For $\mathcal{L}_{\text{afw\_gd+}}$, we considered only one extension of the GD,



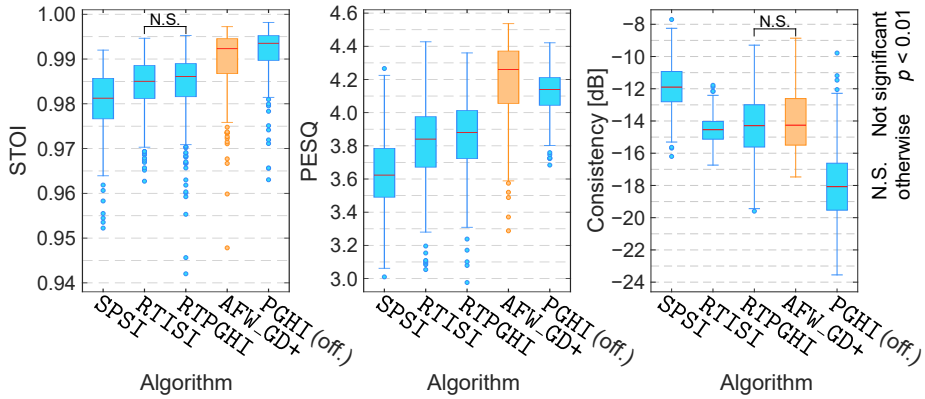
(a) STOI (b) PESQ (c) Consistency

Figure 4.4: Performances of different loss functions for DNN-based phase reconstruction, where blue and red respectively indicate conventional and proposed methods.

i.e., $\boldsymbol{S} = \{1, 2\}$, corresponding to the weights $\lambda_{\mathrm{gd}(1)} = 1.0$ and $\lambda_{\mathrm{gd}(2)} = 0.1$.

For real-time phase reconstruction, we utilized a causal DNN architecture, i.e., the convolutional recurrent network (CRN) [83], as shown in in Fig. 4.3. The encoder and decoder were designed symmetrically, each comprised five convolutional/deconvolutional layers with gated linear units [84] (ConvGLU/DeconvGLU). For each layer, we used a kernel size of 2×3 (time $\times$ frequency), stride of $(1, 2)$, and number of channels of 64. We also applied the instance normalization (IN) [85] and parametric rectified linear unit (PReLU) after each layer, except for the last layer. Temporal information was modeled by two layers of long short-term memory (LSTM) with 256 units per layer. The dense layers were utilized to convert the dimensions of the output/input of the encoder/decoder to the input/output of the LSTM layers. In total, the model consisted of nearly 2.2 million parameters.

The input of the models was the log amplitude spectrogram normalized to zero mean and unit variance. The output was the phase spectrogram. The STFT was calculated using a Hamming window with a 32-ms length, 8-ms shift, and 512-point DFT. The Adam optimizer was used with a batch size of 4 audio samples and the learning rate of 10^{-5}. For the first part of the experiments, each model was trained for 1000 epochs. The model in the second part was trained for $10\,000$ epochs.



(a) STOI (b) PESQ (c) Consistency

Figure 4.5: Performances of real-time phase reconstruction algorithms (except PGHI).

4.4.2 Experimental results and discussion

Fig. 4.4 compares the performances of loss functions for DNN-based phase reconstruction, which exhibit a similar pattern for all metrics. In the comparison of the first two methods, `P` performs notably better than `P_noAug`, even though they use the same loss function $\mathcal{L}_{\mathrm{p}}$. This highlights the efficacy of the proposed data augmentation. Although the shifting technique is not a novel approach, it may have been overlooked in training DNNs for estimating the amplitude because the amplitude changes slowly over time, and this type of data augmentation may not have much of an effect. However, the experimental results here demonstrate that the shifting technique can be highly beneficial for phase reconstruction. Fig. 4.4 also shows that, in comparison with the conventional loss functions, the proposed loss functions `FW`, `AFW`, and `AFW_GD+` gradually lead to a better performance, thereby demonstrating the efficacy of the frequency-specific weights, amplitude weights, and the extended GD in DNN-based phase reconstruction. We have found that these DNN-based models for phase reconstruction perform better when the fundamental frequency of the signal is low and become less stable when the fundamental frequency is high. This leads to the large ranges of their scores as well as overlaps between these score distributions. However, the paired sample t-test demonstrated that all the improvements are statistically significant, except between the consistency measures of `VMMIFGD` and `FW`. In addition, the final proposed method, `AFW_GD+`, clearly outperforms all conventional methods.

Fig. 4.5 presents a comparison of real-time phase reconstruction algorithms, with the offline algorithm `PGHI` as a reference. The results here reveal that the proposed method achieves

superior performances in PESQ and STOI while maintaining a comparable consistency measure to other real-time algorithms. In addition, the proposed method even outperforms the offline `PGHI` algorithm in PESQ and slightly underperforms in STOI. These results demonstrate the efficacy of the DNN-based method in real-time phase reconstruction.

Another advantage of the DNN-based methods is their flexibility for adaptation to various applications. For example, when a noisy/mixed phase is available, it can easily be incorporated as input to improve the performance of the model. A drawback of the conventional non-DNN methods is that they usually require a clean amplitude to estimate the phase. In contrast, DNNs can estimate the phase by using any features that contain the phase information, even if they are not clean. However, the DNN-based phase reconstruction still faces the challenge of rapid phase changes at high frequencies. Although this paper proposes using low weights for the phase loss to reduce the sensitivity, it does not fully address the problem of effectively reconstructing the high-frequency phase. Possible directions for future work include utilizing other advanced DNN architectures to better model the phase sensitivity and incorporating other phase features to enhance the phase structure.

4.5 Conclusions

This chapter presented improvements to DNN-based real-time phase reconstruction. Utilizing the varying properties of the phase as a basis, this thesis proposed loss functions that incorporate frequency-specific weights, amplitude weights, and an extension of the GD. In addition, a data augmentation scheme was introduced to improve the model generalization. Experimental results demonstrated the efficacy of the data augmentation and the superior performance of the proposed loss functions compared to conventional loss functions. The results also showed that the proposed method outperforms other non-DNN real-time phase reconstruction methods.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

This thesis investigated the phase-based features and proposed phase reconstruction algorithms for both offline and online conditions.

The wrapping issue obscures the underlying structure of the phase, making it difficult to extract useful information. To address this issue, an approach is to transform the phase into alternative representations. In Chapter 2, this thesis explored several phase-based features, including the IF, GD, IFD, RPS, and PD, which have been conventionally used. Additionally, two novel phase-based features, namely the DIF and IFPD, were introduced and their properties were investigated.

The aim of phase reconstruction is to estimate a phase spectrogram for a given amplitude spectrogram to reconstruct the time-domain signal. In Chapter 3, this thesis presented two approaches for the second stage of two-stage phase reconstruction algorithms, i.e., the weighted LS method and the ML-based method using the *von Mises* distribution. In the theory discussion, this thesis analyzed the GD properties and interpreted the roles of the phase relationships between TF bins in the ISTFT. Both objective and subjective experiments showed that the performance of the ML-based method using the IFPD is superior to other methods that use only the IF and GD. The results also suggest that ML-based methods perform better than other methods in the second stage, and the use of amplitude weights significantly improves the results of the conventional LS-based method.

For real-time phase reconstruction, in Chapter 4, this thesis proposed loss functions to train a causal DNN model. Utilizing various properties of the phase, frequency-specific weights, amplitude weights, and the IFPD were incorporated into the loss functions to enhance the training process. In addition, a data augmentation scheme was introduced to improve the model's generalization. Experimental results demonstrated the superior performance of the proposed method compared to other conventional DNN-based and non-DNN phase reconstruction algorithms.

5.2 Future Works

Future work includes investigating the properties and potential applications of phase-based features. One promising direction is to incorporate these phase features as auxiliary inputs, in addition to the amplitude, to improve the DNN models.

Regarding the two-stage phase reconstruction algorithms, future work will investigate the effects of the first stage on the final results. This investigation will include using other advanced DNN architectures and combining the models in the first stage into one model with multiple outputs.

In the context of online DNN-based phase reconstruction, a key research objective will be addressing the challenge of rapid phase changes at high frequencies. Potential approaches may involve leveraging other advanced DNN architectures to better model the phase sensitivity and incorporating other phase features to enhance the phase structure.

Furthermore, extending the application of phase reconstruction to other areas, such as STFT-based speech synthesis and speech enhancement, shows promise for future research directions.

Bibliography

[1] H. Von Helmholtz, On the Sensations of Tone as a Physiological Basis for the Theory of Music, Longmans, Green, 1912.

[2] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.30, no.4, pp.679–681, Aug. 1982.

[3] P. Vary and M. Eurasip, "Noise suppression by spectral magnitude estimation—mechanism and theoretical limits—," Signal Processing, vol.8, no.4, pp.387–400, Jul. 1985.

[4] F.A. Bilsen, "On the influence of the number and phase of harmonics on the perceptibility of the pitch of complex signals," Acta Acustica united with Acustica, vol.28, no.1, pp.60–65, Jan. 1973.

[5] R. Plomp and H.J.M. Steeneken, "Effect of phase on the timbre of complex tones," The Journal of the Acoustical Society of America, vol.46, no.2B, pp.409–421, 1969.

[6] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," Digital Signal Processing, vol.17, no.3, pp.578–616, May 2007.

[7] K.K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," Eighth European Conference on Speech Communication and Technology, pp.2117–2120, Sept. 2003.

[8] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.8, pp.1283–1294, Aug. 2015.

[9] K. Vijayan and K.S.R. Murty, “Analysis of phase spectrum of speech signals using allpass modeling,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.12, pp.2371–2383, Dec. 2015.

[10] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” IEEE Signal Processing Magazine, vol.32, no.2, pp.55–66, Mar. 2015.

[11] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, “Phase reconstruction based on recurrent phase unwrapping with deep neural networks,” 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.826–830, May 2020.

[12] L. Thieling, D. Wilhelm, and P. Jax, “Recurrent phase reconstruction using estimated phase derivatives from deep neural networks,” 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7088–7092, Jun. 2021.

[13] B.T. Nguyen, Y. Wakabayashi, K. Iwai, and T. Nishiura, “Two-stage phase reconstruction using DNN and von Mises distribution-based maximum likelihood,” 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.995–999, Dec. 2021.

[14] Y. Masuyama, K. Yatabe, K. Nagatomo, and Y. Oikawa, “Online Phase Reconstruction via DNN-Based Phase Differences Estimation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.31, pp.163–176, Nov. 2022.

[15] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, “Simple representation of signal phase for harmonic speech models,” Electronics letters, vol.45, no.7, pp.381–383, Mar. 2009.

[16] G. Degottex and D. Erro, “A uniform phase representation for the harmonic model in speech synthesis applications,” EURASIP Journal on Audio, Speech, and Music Processing, vol.2014, no.1, pp.1–16, Oct. 2014.

[17] B.T. Nguyen, Y. Wakabayashi, K. Iwai, and T. Nishiura, "Analysis of derivative of instantaneous frequency and its application to voice activity detection," Applied Acoustics, vol.181, pp.108–116, Oct. 2021.

[18] J. Muth, S. Uhlich, N. Perraudin, T. Kemp, F. Cardinaux, and Y. Mitsufuji, "Improving DNN-based music source separation using phase features," arXiv:1807.02710, 2018.

[19] 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, no.8, pp.1256–1266, May 2019.

[20] Z.Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.28, pp.1778–1787, May 2020.

[21] D. Ma, N. Hou, V.T. Pham, H. Xu, and E.S. Chng, "Multitask-based joint learning approach to robust ASR for radio communication speech," 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.497–502, Dec. 2021.

[22] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation," INTERSPEECH, pp.2713–2717, Sept. 2018.

[23] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J.R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," IEEE Journal of Selected Topics in Signal Processing, vol.13, no.2, pp.370–382, Mar. 2019.

[24] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.26, no.6, pp.1095–1105, Jun. 2018.

[25] 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, Sept. 2018.

[26] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.9, pp.1521–1532, Sept. 2015.

[27] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5560–5564, Mar. 2017.

[28] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.32, no.2, pp.236–243, Apr. 1984.

[29] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," INTERSPEECH, pp.1128–1132, Aug. 2017.

[30] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," INTERSPEECH, pp.3389–3393, Aug. 2017.

[31] Y. Saito, S. Takamichi, and H. Saruwatari, "Text-to-speech synthesis using STFT spectra based on low-/multi-resolution generative adversarial networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5299–5303, Apr. 2018.

[32] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," 2015 23rd European Signal Processing Conference (EUSIPCO), pp.1–5, Sept. 2015.

[33] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," International Conference on Digital Audio Effects (DAFx), pp.397–403, Sept. 2010.

[34] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Griffin–Lim like phase recovery via alternating direction method of multipliers," IEEE Signal Processing Letters, vol.26, no.1, pp.184–188, Nov. 2018.

[35] T. Peer, S. Welker, and T. Gerkmann, "Beyond Griffin-Lim: Improved Iterative Phase Retrieval for Speech," arXiv preprint arXiv:2205.05496, 2022.

[36] Z. Průša, P. Balazs, and P.L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.25, no.5, pp.1154–1164, Mar. 2017.

[37] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," Proceedings of the IEEE, vol.80, no.4, pp.520–538, Apr. 1992.

[38] R.M. Hegde, H.A. Murthy, and V.R.R. Gadde, "Significance of the modified group delay feature in speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.1, pp.190–202, Jan. 2007.

[39] 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, no.1, pp.63–76, Sept. 2018.

[40] A.A. Nugraha, K. Sekiguchi, and K. Yoshii, "A deep generative model of speech complex spectrograms," 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.905–909, May 2019.

[41] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," 2018 International Workshop on Acoustic Signal Enhancement (IWAENC), pp.286–290, Sept. 2018.

[42] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network," IEEE Journal of Selected Topics in Signal Processing, vol.15, no.1, pp.37–50, Oct. 2020.

[43] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," IEEE Transactions on Signal Processing, vol.62, no.16, pp.4199–4208, Jul. 2014.

[44] T. Gerkmann, "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4478–4482, Jul. 2014.

[45] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.28, pp.380–390, 2019.

[46] Z. Ouyang, H. Yu, W.P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5756–5760, May 2019.

[47] N. Sturmel, L. Daudet, *et al.*, "Signal reconstruction from STFT magnitude: A state of the art," International Conference on Digital Audio Effects (DAFx), pp.375–386, Sept. 2011.

[48] G. Degottex and N. Obin, "Phase distortion statistics as a representation of the glottal source: Application to the classification of voice qualities," Fifteenth Annual Conference of the International Speech Communication Association, pp.1633–1637, Sept. 2014.

[49] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4608–4611, May 2011.

[50] P.L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4844–4847, May 2011.

[51] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.22, no.12, pp.1931–1940, Dec. 2014.

[52] A.P. Stark and K.K. Paliwal, "Group-delay-deviation based spectral analysis of speech," INTERSPEECH, pp.1083–1086, Sept. 2009.

[53] A.P. Stark and K.K. Paliwal, "Speech analysis using instantaneous frequency deviation," Ninth Annual Conference of the International Speech Communication Association, pp.2602–2605, Sept. 2008.

[54] D. Friedman, "Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.10, pp.1121–1124, Apr. 1985.

[55] H.A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," Speech Communication, vol.10, no.3, pp.209–221, Aug. 1991.

[56] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," The Journal of the Acoustical Society of America, vol.105, no.3, pp.1912–1924, Mar. 1999.

[57] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," IEEE transactions on audio, speech, and language processing, vol.16, no.6, pp.1097–1111, Aug. 2008.

[58] R.M. Hegde, H.A. Murthy, and G.V.R. Rao, "Application of the modified group delay function to speaker identification and discrimination," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.1, pp.517–520, May 2004.

[59] L. Gu, Single-channel speech separation based on instantaneous frequency, Ph.D. thesis, Columbia University, 2010.

[60] V.K. Prasad, T. Nagarajan, and H.A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," Speech Communication, vol.42, no.3-4, pp.429–446, Apr. 2004.

[61] Y. Agiomyrgiannakis and Y. Stylianou, "Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech," IEEE Transactions on Audio, Speech, and Language Processing, vol.17, no.4, pp.775–786, 2009.

[62] D.C. Ghiglia and M.D. Pritt, Two-dimensional phase unwrapping: Theory, algorithms, and software, Wiley, 1998.

[63] B.N. Datta, Numerical linear algebra and applications, Siam, 2010.

[64] D.W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," Journal of the society for Industrial and Applied Mathematics, vol.11, no.2, pp.431–441, Jun. 1963.

[65] W. Kahan, "Accurate eigenvalues of a symmetric tri-diagonal matrix," tech. rep., Computer Science Dept., Stanford University, Jul. 1966.

[66] S.J. Wright, "Coordinate descent algorithms," Mathematical Programming, vol.151, no.1, pp.3–34, Mar. 2015.

[67] J.S. Garofolo, TIMIT acoustic phonetic continuous speech corpus, Linguistic Data Consortium, 1993.

[68] A.V.D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv:1609.03499, 2016.

[69] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on directional-statistics deep neural networks," Signal Processing, vol.169, p.107368, Apr. 2020.

[70] E. Anderson, Z. Bai, C. Bischof, L.S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, *et al.*, LAPACK users' guide, SIAM, 1999.

[71] 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," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.749–752, May 2001.

[72] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol.19, no.7, pp.2125–2136, Feb. 2011.

[73] Recommendation ITU-R BS.1116-3, Methods for the subjective assessment of small impairments in audio systems, Feb. 2015.

[74] M. Schoeffler, S. Bartoschek, F.R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," Journal of Open Research Software, vol.6, no.1, Feb. 2018.

[75] P. Magron and T. Virtanen, "Online spectrogram inversion for low-latency audio source separation," IEEE Signal Processing Letters, vol.27, pp.306–310, 2020.

[76] T. Yanagita, S. Sakti, and S. Nakamura, "Neural iTTS: Toward synthesizing speech in real-time with end-to-end neural text-to-speech framework," Proceedings of the 10th ISCA Speech Synthesis Workshop, pp.183–188, 2019.

[77] X. Zhu, G.T. Beauregard, and L.L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.5, pp.1645–1653, Jun. 2007.

[78] G.T. Beauregard, M. Harish, and L. Wyse, "Single pass spectrogram inversion," 2015 IEEE international conference on digital signal processing (DSP), pp.427–431, July. 2015.

[79] Z. Pruša and P.L. Søndergaard, "Real-time spectrogram inversion using phase gradient heap integration," International Conference on Digital Audio Effects (DAFx), pp.17–21, Sept. 2016.

[80] B.T. Nguyen, Y. Wakabayashi, G. Yuting, K. Iwai, and T. Nishiura, "Von mises mixture model-based DNN for sign indetermination problem in phase reconstruction," 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.957–961, Nov. 2022.

[81] Z. Průša, P. Balazs, and P.L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.25, no.5, pp.1154–1164, Mar. 2017.

[82] R. ITU, "P. 56: Objective measurement of active speech level," International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), 2011.

[83] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.28, pp.380–390, Nov. 2019.

[84] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, *et al.*, "Conditional image generation with PixelCNN decoders," Advances in neural information processing systems, vol.29, 2016.

[85] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv preprint arXiv:1607.08022, 2016.

List of publications

Journal Paper

1. **Nguyen Binh Thien**, Yukoh Wakabayasshi, Kenta Iwai, and Takanobu Nishiura, "Inter-Frequency Phase Difference for Phase Reconstruction Using Deep Neural Networks and Maximum Likelihood," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 31, pp. 1667-1680, DOI:10.1109/TASLP.2023.3268577, Apr. 2023.

2. **Nguyen Binh Thien**, Yukoh Wakabayasshi, Kenta Iwai, and Takanobu Nishiura, "Analysis of derivative of instantaneous frequency and its application to voice activity detection," Applied Acoustics, Volume 181, DOI: https://doi.org/10.1016/j.apacoust.2021.108116, Oct. 2021.

International Conference

1. **Nguyen Binh Thien**, Yukoh Wakabayasshi, Yuting Geng, Kenta Iwai, and Takanobu Nishiura, "Weighted Von Mises Distribution-based Loss Function for Real-time STFT Phase Reconstruction Using DNN," INTERSPEECH 2023, Dublin, Ireland, Aug. 2023. (Accepted)

2. **Nguyen Binh Thien**, Yukoh Wakabayashi, Yuting Geng, Kenta Iwai and Takanobu Nishiura, "Von Mises Mixture Model-Based DNN for Sign Indetermination Problem in Phase Reconstruction," APSIPA-ASC 2022, pp. 958-962, Thailand, Nov. 2022.

3. **Nguyen Binh Thien**, Yukoh Wakabayashi, Kenta Iwai and Takanobu Nishiura, "Two-stage phase reconstruction using DNN and von Mises distribution-based maximum likelihood,"

APSIPA-ASC 2021, pp. 995-999, Online, Dec. 2021.

4. **Nguyen Binh Thien**, Yukoh Wakabayashi, Takahiro Fukumori and Takanobu Nishiura, "Derivative of Instantaneous Frequency for Voice Activity Detection using Phase-based Approach, " APSIPA-ASC 2019, pp. 1168-1172, Lanzhou, China, Nov. 2019.

Domestic Conference

1. **Nguyen Binh Thien**, Yukoh Wakabayashi, Yuting Geng, Kenta Iwai, Takanobu Nishiura, "Two-stage phase reconstruction using inter-frequency phase difference," 日本音響学会2022年秋季研究発表会, pp. 299-300, Hokkaido, Sep. 2022.

2. **Nguyen Binh Thien**, Yukoh Wakabayashi, Kenta Iwai, Takanobu Nishiura, "Maximum likelihood estimation for phase reconstruction from its derivatives," 日本音響学会2021年秋季研究発表会, pp. 941-942, Online, Sep. 2021.

3. **Nguyen Binh Thien**, Yukoh Wakabayashi, Takahiro Fukumori, Takanobu Nishiura, “A Phase-based Voice Activity Detection using Statistical Likelihood Ratio of the Derivative of Instantaneous Frequency,” 日本音響学会2020年春季研究発表会, pp. 905-906, Saitama, Mar. 2020.

4. **Nguyen Binh Thien**, Yukoh Wakabayashi, Takahiro Fukumori, and Takanobu Nishiura, “Speech analysis using the second derivative of phase spectrum,” 日本音響学会2019年秋季研究発表会, pp.853-854, Shiga, Sep. 2019.