Phase-Aware Speech Enhancement Based on Deep Neural Networks

Naijun Zheng and Xiao-Lei Zhang

Abstract—**Short-time frequency transform (STFT) is fundamental in speech processing. Because of the difficulty of processing highly unstructured STFT phase, most speech-processing algorithms only operate with STFT magnitude, leaving the STFT phase far from explored. However, with the recent development of deep neural network (DNN) based speech processing, e.g., speech enhancement and recognition, phase processing is becoming more important than ever before as a new growing point of DNN-based methods. In this paper, we propose a phase-aware speech enhancement algorithm based on DNN. Specifically, in the training stage, when incorporating phase as a target, our core idea is to transform an unstructured phase spectrogram to its derivative along the time axis, i.e., instantaneous frequency deviation (IFD), which has a similar structure with its corresponding magnitude spectrogram. We further propose to optimize both IFD and magnitude jointly in a multiobjective learning framework. In the test stage, we propose a postprocessing method to recover the phase spectrogram from the estimated IFD. Experimental results demonstrate the effectiveness of the proposed method.**

Index Terms—**Deep neural network (DNN), phase estimation, speech enhancement, instantaneous frequency, harmonic model.**

I. Introduction

SPEECH enhancement has been studied extensively as a fundamental problem of signal processing. Speech enhancement techniques have been widely used in speech communication, speech analysis, speech recognition, etc. Short-time Fourier transform (STFT) is one of the bases of speech enhancement. It converts a speech signal in time domain to a spectro-temporal spectrogram, where the harmonic structure of the speech can be observed clearly. A STFT spectrogram can be decomposed to a magnitude spectrogram and a phase spectrogram. Most speech enhancement algorithms focused only on processing magnitude spectrograms during the last decades, due to the following reasons. First, some work indicated that the enhancement of magnitude is more important than that of phase, since a phase spectrogram can be enhanced iteratively by its associated magnitude spectrogram [1]. Moreover, a phase spectrogram seems randomly distributed and unstructured [2], which is difficult to be processed directly.

Phase processing is important to speech enhancement. Recently, phase has shown its strong relationship with speech quality [3], [4]. Phase processing has also received much attention than ever before. Examples include the consistent Wiener filtering [5] and phase reconstruction [6]. In these works, the derivatives of a phase spectrogram along the time and frequency axes, named *instantaneous frequency* (IF) [7] and *group delay* (GD) [8] respectively, show clear structures that are quite different from the randomly distributed and unstructured phase spectrogram.

To perform speech enhancement with a target of no matter whether magnitude-based or phase-based, one needs to construct a mapping function, either *model-based* or *data-driven*, from a noisy feature space to a clean target space. Recently, deep neural network (DNN) based speech enhancement, which is a data driven method that has shown its strong power in adverse environments since its first report [9], has received much attention [10]. DNN is a multilayer perceptron with more than one hidden layers. Each layer of DNN consists of a group of nonlinear hidden neurons in parallel. Due to the hierarchical structure and distributed representation at each layer, the data representation ability of DNN is exponentially more powerful than that of a shallow model when given the same number of nonlinear computational units. With the recent explosion of data and fast development of computing power, it is able to train very powerful DNN easily, which triggered the breakthrough of speech processing. The research on DNN-based speech enhancement methods focused on training targets [11]–[16], DNN models [9], [17]–[19], different types of noises [20]–[22], and different kinds of sensors [23]–[25], see [10] for an overview.

Most DNN-based speech enhancement methods use magnitude-aware training targets [11], either *mapping-based* or *masking-based*, leaving the noisy phase unprocessed. These methods do not fully utilize phase information for further improving the performance. Recently, manipulating on the full STFT expression of data is a new growing point of DNN-based speech enhancement. Examples include the phase-sensitive filter (PSF) [12] and complex-IRM (cIRM) [16]. However, the

Manuscript received July 18, 2017; revised October 21, 2017, January 2, 2018, March 7, 2018, and July 25, 2018; accepted September 9, 2018. Date of publication September 17, 2018; date of current version October 15, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671381, in part by the Shenzhen Science and Technology Plan under Grant JCYJ20170815161820095, and in part by the Shaanxi Natural Science Basic Research Program under Grant 2018JM6035. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tara N. Sainath. *(Corresponding author: Xiao-Lei Zhang.)*

N. Zheng is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: njzheng@stu.xidian.edu.cn).

X.-L. Zhang is with the Center for Intelligent Acoustics and Immersive Communications, School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xiaolei.zhang@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2018.2870742

above training targets [12], [16] are formulated in the *complex* rectangular coordinate system, where the phase and magnitude information exists in both the real and imaginary parts. These methods do not directly deal with the difficulty of processing a phase spectrogram which seems randomly distributed and highly unstructured. To our knowledge, no methods deal with phase spectrograms directly.

In this paper, we propose a phase-aware DNN-based speech enhancement method to deal with phase spectrograms directly. Specifically, to overcome the difficulty of processing a highly unstructured phase spectrogram, we employ the derivative of the phase spectrogram along the time axis, named *instantaneous frequency deviation* (IFD) [26], as the training target. Geometrically, IFD has a clear structure similar to its corresponding magnitude spectrogram. Theoretically, IFD is able to alleviate the wrapping problem of phase thanks to its derivative calculation along the time axis. We further propose to optimize IFD and magnitude jointly in a multi-objective learning framework. However, the estimated IFD in the test stage is only an estimate of the derivative of the phase spectrogram which cannot be used alone for recovering the phase spectrogram. To overcome this difficulty, we further propose a post-processing method which reconstructs the phase spectrogram by jointly processing the estimated IFD, estimated magnitude mask, and noisy phase. Our experimental results demonstrate that the proposed method outperforms the DNN method that estimates the magnitude spectrogram only in both matching and mismatching test environments.

The rest of the paper is organized as follows. In Section II, we present the motivation for conducting phase estimation. In Section III, we first propose a phase-aware training target of DNN, then present a multi-objective learning framework for jointly estimating magnitude and phase with DNN, and at last propose a post-processing method for phase recovery in the test stage. In Section IV, we report the empirical evaluation results of the proposed method. In Section V, we conclude our contributions.

II. Motivation

The STFT spectrogram of a signal $x\left(n\right)$ can be written as [27]:

$$\begin{aligned} X\left(k,l\right) &= \mathrm{STFT}\left\{x\left(n\right)\right\} \\ &= \sum_{n=0}^{N-1} x\left(lL+n\right) w\left(n\right) e^{-j\frac{2\pi}{N}kn} \end{aligned} \tag{1}$$

where k and l indicate the frequency band and frame index of the STFT spectrogram respectively, n indicates the time index, L is the time shift between two adjacent frames, $w(l)$ is an analysis window with a length of L_w for dividing the signal into frame segments, the overlapping ratio between two adjacent frames is $1 - L/L_w$, and N is the length the discrete Fourier transform (DFT). For speech enhancement, we denote the clean speech, additive noise, and corrupted noisy speech in time domain by $x(n)$, $z(n)$, and $y(n)$ respectively, with the T-F units of their corresponding spectrograms denoted by $X(k,l)$, $Z(k,l)$, and $Y(k,l)$ respectively. The value of a T-F unit is a complex number, which has two expressions: One is in the rectangular coordinate system, which decomposes the complex number into a real part and an imaginary part:

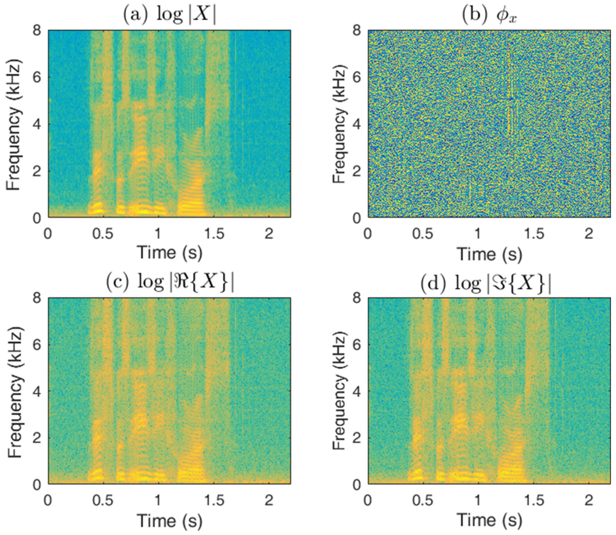

Fig. 1. Four parameters of a spectrogram.

$$X(k,l) = \Re\left\{X(k,l)\right\} + \Im\left\{X(k,l)\right\} \tag{2}$$

The other is in the polar coordinate system, which decomposes the complex number into a magnitude part and a phase part:

$$X(k,l) = |X(k,l)|e^{j\phi_x(k,l)} \tag{3}$$

For simplicity, we denote the phase (or magnitude) of the clean speech, noisy speech, and noise speech as *clean phase (or magnitude)*, *noisy phase (or magnitude)*, and *noise phase (or magnitude)* respectively. An example of the four parameters is illustrated in Fig. 1.

In this paper, we focus on developing speech enhancement methods working with the polar coordinate expression where the magnitude spectrogram has a clear structure, while the phase spectrogram does not have such a clear structure. Because it is efficient to suppress the background noise when the spectrogram of a noisy speech signal has a clear structure, magnitude-aware speech enhancement is much more popular than phase-aware speech enhancement.

Phase-aware speech enhancement is important. Here we give two examples that emphasize its importance. In [28], the authors synthesized an utterance by combining the spectral amplitude of a real-world utterance with an artificially generated spectral phase that is dramatically different from the spectral phase of the real-world utterance. The synthesized utterance, which sounds like a rock music, is quite different from the original speech. Similarly, in Fig. 2, we draw several speech signals that are synthesized from the same magnitude of an original noisy speech signal with the clean, noisy, and noise phase of the original speech respectively. Comparing the deviation between the synthesized speech signals and their clean counterpart, we observe that the noisy speech signal synthesized with the clean phase is very close to the clean speech signal, while the deviation of

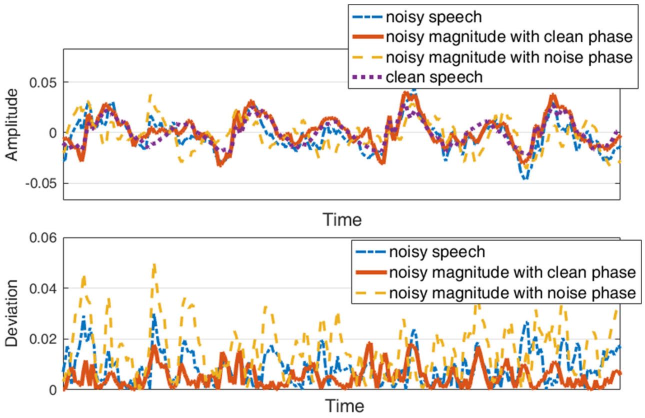

Fig. 2. On the importance of phase estimation. (a) Comparison of the noisy speech signals in time domain that are synthesized with different phase, where the noisy scenario is the factory noise at an SNR of -3 dB. (b) Comparison of the deviations of the noisy speech signals to their corresponding clean speech signal.

the noisy speech signal synthesized with the noise phase is even worse than that of the noisy speech signal. The average deviation of the synthesized signals with the clean, noisy, and noise phase from the corresponding clean speech is 0.0047, 0.0064, and 0.0089 respectively.

However, phase-aware speech enhancement is difficult, particularly for a DNN-based approach. The main difficulty is that the values of the T-F units in a phase spectrogram, as shown in Fig. 1 b, fluctuate rapidly along the time and frequency axes, and are uniformly distributed in the range of $[-\pi, \pi)$ [29] due to phase wrapping. As we know, it is difficult to train a statistical model, like DNN, with highly unstructured input or output training patterns. Another difficulty is that phase is a function of time. We take a sinusoidal signal in time domain as an example. Its spectral phase at the dominant frequency is a linear function of the frame indices (see Appendix A for the proof). Due to the dependency of phase to its frame indices, the deep models that do not take the time dependency into consideration is difficult to be applied, such as the standard feedforward DNN with stochastic gradient descent training. To overcome the above difficulties, a highly-structured new target derived from the phase expression as well as a post-processing method that recovers the original phase expression from the new target are strongly needed for DNN-based speech enhancement.

III. Method

In this section, we first present a phase-aware training target for DNN, then present the multi-objective training of DNN with the new target, and at last describe a post-processing method of the new target to reconstruct the enhanced phase.

A. A Phase-Aware Training Target for DNN

To derive a highly-structured phase-aware target, we need to overcome the difficulties of the phase wrapping and time-related effect as we have analyzed in Section II. Here we employ a variant of the STFT phase that is able to extract structured patterns from phase spectrograms: *instantaneous frequency* (IF) [7] which calculates the negative derivative of the phase spectrograms along the time axis. For discrete-time signals, IF is calculated by:

$$\begin{aligned}\mathrm{IF}_x(k,l) &= \mathrm{principle}(\phi_x(k,l) - \phi_x(k,l+1)) \\ &= \arg\left(X(k,l)\,X^*(k,l+1)\right)\end{aligned} \tag{4}$$

where the function $\mathrm{principle}(\cdot)$ denotes the selection of the principal values which projects the phase difference onto $[-\pi, +\pi)$, $\arg(\cdot)$ calculates the phase angle of a complex number, X^* denotes the complex conjugate of the complex number X, and the subscript x means that IF_x is calculated from the clean phase. The IF of the speech signal in Fig. 1 is demonstrated in Fig. 3 b.

However, as shown in Fig. 3 b, IF contains some parallel striation along the time axis. To eliminate the striation, we employ a deviation of IF, named *instantaneous frequency deviation* (IFD) [26]:

$$\mathrm{IFD}_x(k,l) = \mathrm{IF}_x(k,l) - \frac{2\pi}{N}kL \tag{5}$$

which measures how far an IF value strays from its center frequency $\frac{2\pi}{N}kL$. As shown in Fig. 3 c, IFD can give clear pitch and a harmonic structure similar to that in the magnitude spectrogram; compared with IF, IFD eliminates the striation caused by different center frequencies, which makes the structure of the speech more apparent.

IFD is able to alleviate the *phase wrapping* problem. Specifically, phase wrapping is a difficult problem that $-\pi$ and π are physically a same value, but we usually regard them as different states in the model training. IFD is an effective target that can concentrate the output value close to 0 and therefore reduce the probability of generating large phase values close to π or $-\pi$. We show the probability density functions of IFD with different frame overlaps in Fig. 4. From the figure, we can see that the probability density function of IFD at the position of $-\pi$ and π is small when the frame overlap is large, which greatly alleviates the phase wrapping problem. In this paper, the frame overlap is set to 75% of the frame length.

IFD can also be a common target of DNN. To support our claim, we focus on presenting the similarity between the magnitude and IFD, since many DNN-based speech enhancement techniques use clean magnitude spectrograms or their variants as targets. In [30], the author found that, when a Gaussian window is selected as the analysis window of STFT, a transform between the derivative of the STFT magnitude and the derivative of the STFT phase exists, which indicates that the temporal-spectral patterns of the derivatives contain similar information. Here, we compared IF, IFD, and the derivatives of the logarithm of the magnitude spectrogram along frequency axis in Fig. 3, where Hamming windows were used as the analysis windows of STFT. We find that IFD and the derivative of the magnitude spectrogram have a similar pattern, which reaches a similar conclusion with that in [30]. Our finding indicates that IFD can be used as a common target of DNN-based speech enhancement wherever magnitude spectrograms or their variants are used as targets.

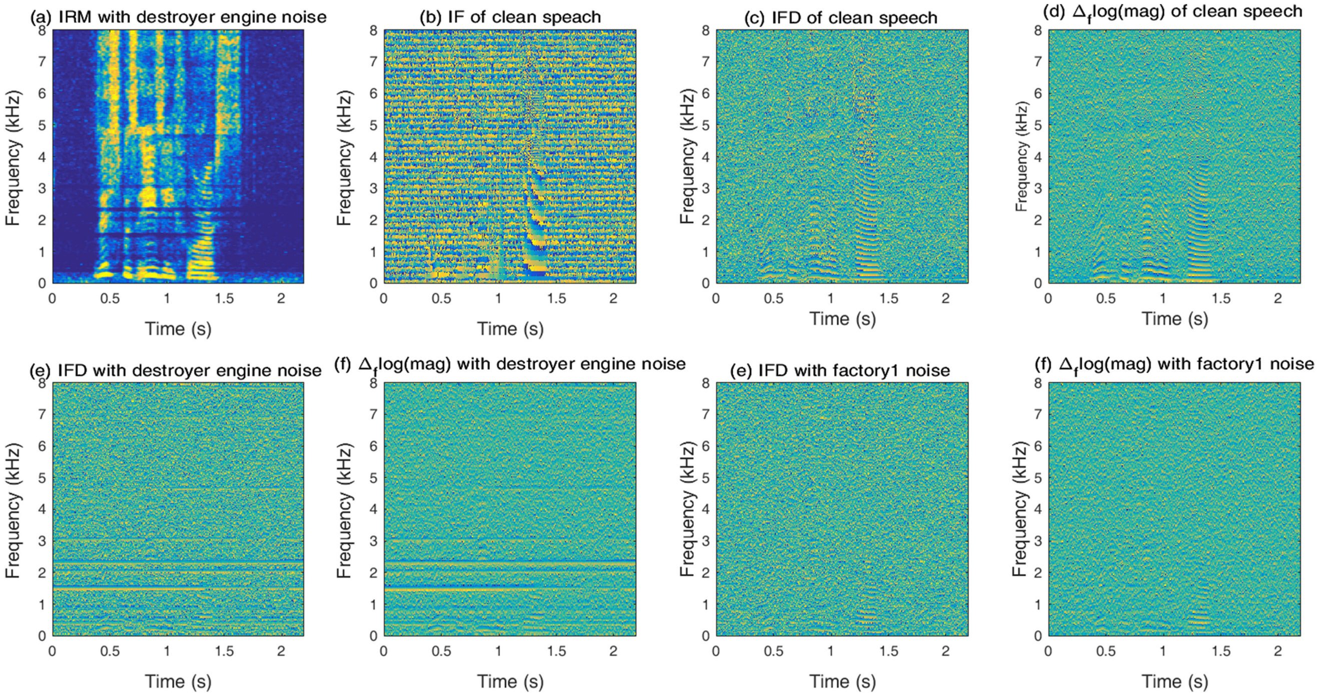

Fig. 3. Visualizations of IRM, IF, IFD, and the derivative of the logarithm of the magnitude spectrogram along the frequency axis (denoted by $\Delta_f \log(\mathrm{mag})$) in clean and noisy (destroyer engine noise or factory1 noise at -3 dB) environments.

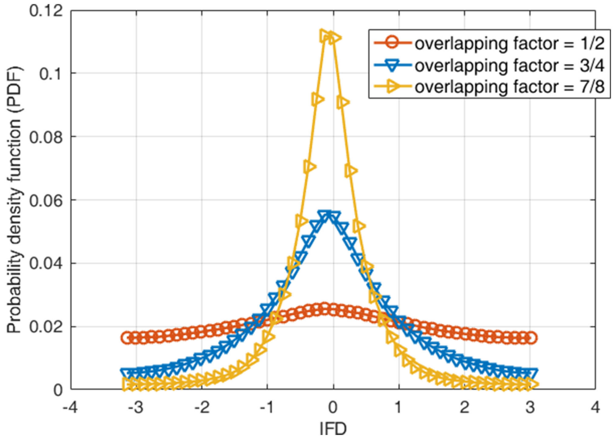

Fig. 4. Probabilistic density functions of IFD of clean speech with different frame overlapping factors, where the frame overlapping factor is defined as the percentage of the frame overlap over the frame length.

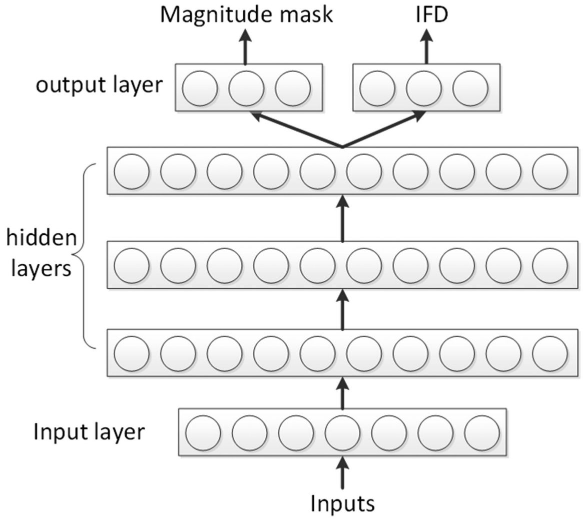

Fig. 5. Architecture of the phase-aware multi-objective DNN.

TABLE I
MAGNITUDE MASKS

Ideal ratio mask [31]	$\mathrm{IRM}_x(k,l) = \left(\frac{X^2(k,l)}{X^2(k,l)+Z^2(k,l)}\right)^\beta$
Ideal amplitude mask [12]	$\mathrm{IAM}_x(k,l) = \lvert X(k,l)\rvert / \lvert Y(k,l)\rvert$
Phase-sensitive filter [12]	$\mathrm{PSF}_x(k,l) = \Re\{X(k,l)/Y(k,l)\}$

B. Phase-Aware Multi-Objective DNN

1) A Phase-Aware Multi-Objective Learning Framework: We propose to estimate the magnitude and phase of STFT simultaneously by multi-objective learning. The proposed architecture is shown in Fig. 5, where a magnitude mask is defined as some energy or amplitude ratio between the clean speech and the background noise.The optimization target of the magnitude-aware objective is flexible. This paper investigates three kinds of magnitude masks listed in Table I, where $X^2(k,l)$ and $Z^2(k,l)$ denote the energy of the clean speech and noise at the (k,l)-th T-F unit respectively.

In Table I, parameter β of the ideal ratio mask (IRM) [31] is a scale parameter. We set β to 0.5 in this paper. PSF [12] calculates the real part of a complex-valued ratio. Note that the range of IRM is $[0,1]$, while the ranges of IAM and PSF are not $[0,1]$. To fit the IAM and PSF suitable to our DNN model whose output units are sigmoid functions, we truncate the ranges of IAM and PSF into $[0,1]$ in dimension.

The main difference of the multi-objective architecture from a standard DNN is that the output layer of the multi-objective architecture has two outputs: one for the magnitude mask estimation, and the other for the IFD estimation. The motivation for jointly learning the magnitude mask and IFD with the same input feature is that, as illustrated in Section III-A, the magnitude masks (e.g. IRM) and IFD have a similar structure, hence it is able to adopt the acoustic features that have been applied successfully to IRM-based methods for IFD as well.

The input acoustic feature of the multi-objective DNN is a concatenation of a set of complementary features and their deltas [32], where the complementary features include the amplitude modulation spectrogram, relative spectral transformed perceptual linear prediction coefficients, mel-frequency cepstral coefficients, and 64-channel Gammatone filterbank power spectra. All these features are noise robust ones.[1] They have been adopted by many magnitude-aware speech enhancement methods, e.g. [11]. We have also observed in the task of voice activity detection that the performance of DNN is improved gradually by combining more hand-engineering complementary features [33]. To explore the contextual information, we also expand the input feature from a single frame to a new vector that is centered at the frame and also incorporates the neighboring $2W$ frames, in other words, the window length of the input is set to $2W + 1$, where W is the half-window length.

We have been aware that DNN has the potential of extracting highly abstract features from original signals directly without hand-engineering features, given large enough training data. Here we still use the complementary features, leaving the completely end-to-end training of DNN as our future work.

2) DNN Training: The training procedure of the proposed multi-objective DNN is summarized in Fig. 6. To balance IFD and the magnitude mask on the training errors, we normalize IFD into the range of $[0, 1)$ and denote the normalized IFD as Ω_x:

$$\Omega_x(k,l) = \frac{1}{2\pi}\mathrm{IFD}_x(k,l) + \frac{1}{2} \tag{6}$$

Similarly, we use the symbol M_x to denote the ideal magnitude mask. We adopt the following two loss functions for the multi-objective DNN training: one is named mask approximation (MA):

$$\mathrm{MSE}_{\mathrm{MA}}(l) = \frac{1}{2N}\sum_{k=0}^{N-1}\left[\left(\mathrm{M}_x(k,l) - \hat{\mathrm{M}}_x(k,l)\right)^2 + \left(\Omega_x(k,l) - \hat{\Omega}_x(k,l)\right)^2\right] \tag{7}$$

and the other one is named masked signal approximation (mSA):

$$\mathrm{MSE}_{\mathrm{mSA}}(l) = \frac{1}{2N}\sum_{k=0}^{N-1}\left[|Y(k,l)|^2\left(\mathrm{M}_x(k,l) - \hat{\mathrm{M}}_x(k,l)\right)^2 + \left(\Omega_x(k,l) - \hat{\Omega}_x(k,l)\right)^2\right] \tag{8}$$

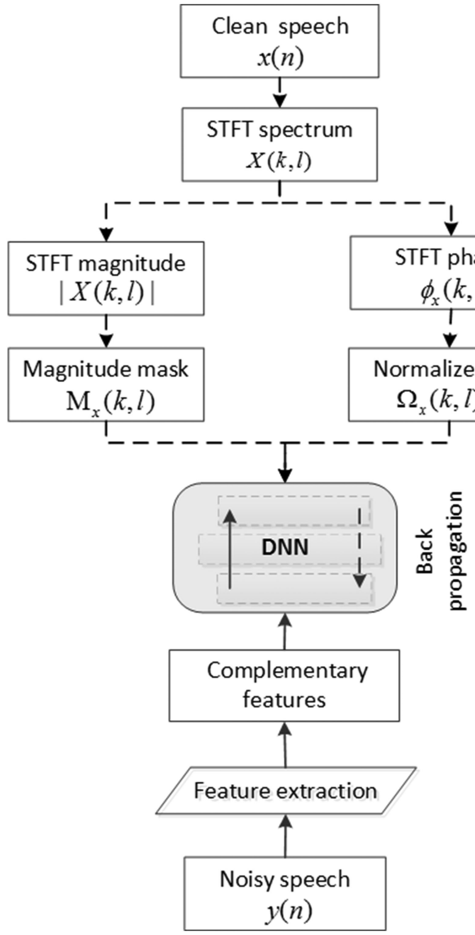

Fig. 6. Diagram of the training procedure of the phase-aware multi-objective DNN.

where the magnitude estimation part of mSA is motivated from [34] [35], $\hat{\mathrm{M}}_x$ and $\hat{\Omega}_x$ denote the estimations of magnitude mask and normalized IFD respectively, MSE is short for mean squared error.

We train each DNN model sufficiently by stochastic back-propagation. The number of training epochs should not only guarantee the convergence of the objective value but also prevent overfitting of the DNN to its training data. Empirically, when the MA loss function is adopted, the maximum epoch number is set as 80, and the model is selected according to the performance on the validation set. When the mSA loss function is adopted, we train the DNN model with the MA loss function for the first 40 epochs, and then continue to train the DNN model with the mSA loss function for at most 40 epochs [17]. The latter training scheme is denoted as MA+mSA.

To better capture the nonlinear variations of data, we set DNN to a depth of three hidden layers, each layer with 1024 hidden neurons. This network structure is commonly used, e.g. [11], [16]. The sigmoid activation function is used as the neurons in the output layer. The rectified linear unit (ReLU) [36] is used as the hidden neurons. The dropout regularization [37] is adopted with the dropout rate set to 0.2. The adaptive gradient descent [38] is used to train the multi-objective DNN with the moment set to 0.5 at the first 5 epochs and 0.9 for the rest.

[1] Our method is not limited to the features mentioned here.

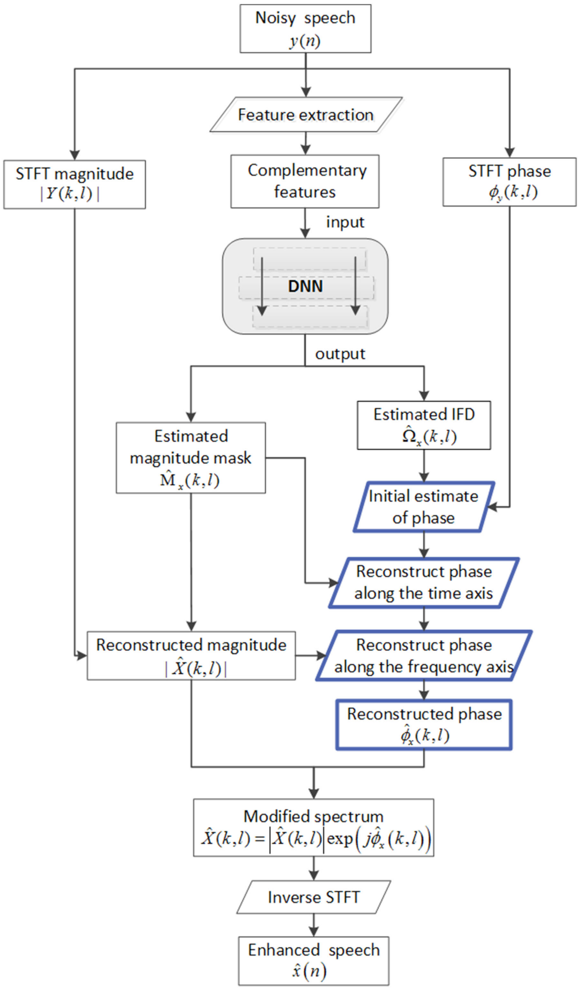

Fig. 7. Diagram of the test procedure of the phase-aware multi-objective DNN.

C. *Post-Processing for Phase Reconstruction in the Test Stage*

In the test stage, we propose a post-processing method to recover the phase spectrogram from an estimated IFD. In this subsection, all kinds of estimates are for the clean speech $x(n)$. Hence, for simplicity, we omit the subscript x from all kinds of estimates of the clean speech $x(n)$, for example, we denote the estimate $\hat{\phi}_x$ by $\hat{\phi}$, the estimate $\hat{\text{IF}}_x$ by $\hat{\text{IF}}$, and so forth.

The main difficulty of the phase recovery from the estimated IFD is that IFD is only the derivative of the phase spectrogram along the time axis. To recover the phase spectrogram, we also need a proper initial estimate of the phase in some T-F units. Then, based on the initial estimate, we can reconstruct the phase spectrogram along the time and frequency axes with the estimated IF by Eq. (4). Fig. 7 shows the entire test procedure, where the proposed post-processing method is highlighted in the boxes with blue bold borders. We present the post-processing method in detail as follows.

First of all, we recover the estimated IF from the estimated IFD by:

$$\hat{\text{IF}}(k,l) = 2\pi\left(\hat{\Omega}(k,l) - \frac{1}{2}\right) + \frac{2\pi}{N}Lk \tag{9}$$

Then, we conduct the following steps:

1) Initial Phase Estimation: Because the STFT spectrogram of the noisy speech $Y(k,l)$ is the summation of the spectrograms of the clean speech $X(k,l)$ and noise $Z(k,l)$, the noisy phase can be formulated as

$$\begin{aligned}\phi_y &= \arg\left(|X|e^{j\phi_x} + |Z|e^{j\phi_z}\right)\\ &= \phi_x + \arg\left(1 + \frac{|Z|}{|X|}e^{j(\phi_z - \phi_x)}\right).\end{aligned} \tag{10}$$

When the amplitude of the clean speech $|X|$ is much larger than the amplitude of the noise $|Z|$, the second term in Eq. (10) is close to 0. Then, the noisy phase can be approximated by the clean phase, i.e., $\phi_y \approx \phi_x$, which means the phase in the high local-SNR regions is nearly unchanged after the noise degradation. Fig. 8 a shows the distance between the noisy phase and the clean phase. From the figure, we observe clearly that the phase in the harmonic regions is nearly uncorrupted.

Based on the above finding, the first way of conducting the phase recovery is to select the T-F units of the noisy phase spectrogram that have high local SNRs as the initial estimate of the corresponding T-F units of the clean phase spectrogram, and then use the selected T-F units to recover the remaining T-F units, where the local SNRs are approximated by $\hat{M}$. Another way is to use the entire noisy phase spectrogram as the initial estimate of the clean phase spectrogram, and then use the local SNR of a T-F unit as the *reliability index* of its estimate. In this paper, we take the second approach:

$$\hat{\phi}_{\text{init}}(k,l) = \phi_y(k,l), \quad \forall k,\ \forall l. \tag{11}$$

2) Phase Reconstruction Along the Time Axis: We first generate $(2N_s + 1)$ frame-conditioned estimates for the (k,l)-th T-F unit by:

$$\hat{\phi}^i(k,l) = \begin{cases}\hat{\phi}_{\text{init}}(k,l+i) + \sum_{n=0}^{i-1}\hat{\text{IF}}(k,l+n), & \text{if } i \neq 0\\ \hat{\phi}_{\text{init}}(k,l+i), & \text{if } i = 0\end{cases},$$
$$\forall -N_s \leq i \leq N_s \tag{12}$$

where i is the frame distance between an initialized T-F unit and the target T-F unit.

The integration of the frame-conditioned estimates can be viewed as an interpolation problem. Several interpolation methods [39] can be applied, such as the nearest point interpolation or linear interpolation. Here we adopt the weighted sum linear interpolation: the final estimate $\hat{\phi}(k,l)$ is a weighted sum of the frame-conditioned estimates:

$$\hat{\phi}(k,l) = \frac{\sum_{i=-N_s}^{N_s}\left(s(i)\hat{\text{M}}(k,l+i)\right)\bar{\phi}^i(k,l)}{\sum_{i=-N_s}^{N_s}s(i)\hat{\text{M}}(k,l+i)} \tag{13}$$

where $s(i)$ denotes the proximity weight for $\bar{\phi}^i(k,l)$ and

$$\bar{\phi}^i(k,l) = unwrap(\hat{\phi}^i(k,l)\,|\hat{\phi}^i(k,l-1)) \tag{14}$$

with $unwrap(\cdot)$ as an unwrapping function to make the phase spectrogram smooth along the time axis.[2] The larger the distance

[2] The unwrapping function is the "unwrap" function in MATLAB.

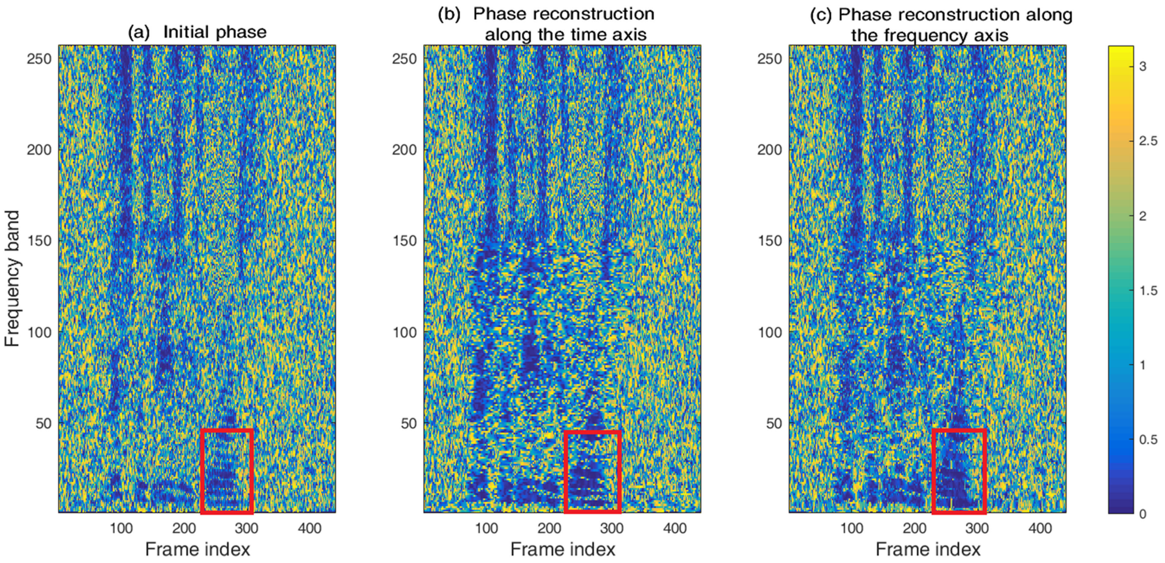

Fig. 8. Distance between the reconstructed phase and the clean phase, i.e., $|\mathrm{principle}(\hat{\phi}(k,l) - \phi_x(k,l))|$, in different phase reconstruction stages, where the dark regions indicate small deviations. The reconstruction process was conducted in the factory noise at an SNR of -3 dB. (a) Initial phase estimation. (b) Phase reconstruction along the time axis. (c) Phase reconstruction along the frequency axis.

$|i|$ is, the smaller the proximity weight $s(i)$ is. The influence of $\bar{\phi}^i(k,l)$ on $\hat{\phi}(k,l)$ is calculated by the product of $s(i)$ and $\hat{\mathrm{M}}(k,l+i)$, where $\hat{\mathrm{M}}(k,l+i)$ is the reliability index we have mentioned in Section III-C1. In this paper, $s(i)$ is set to the Hamming window:

$$s(i) = 0.54 + 0.46\cos\left(\frac{\pi i}{N_s}\right), \quad \forall - N_s \le i \le N_s \qquad (15)$$

where $(2N_s + 1)$ is the length of the Hamming window.

Eq. (13) indicates that the frame-conditioned estimate $\hat{\phi}^i(k,l)$ who has a small $\hat{\mathrm{M}}(k,l+i)$ contributes little to the final estimate $\hat{\phi}(k,l)$, which verifies the correctness of our strategy on the initial phase estimation, i.e., Eq. (11).

3) Phase Reconstruction Along the Frequency Axis: Because the initialized T-F units are only reliable in the high local SNR regions, there still exist large grooves between two neighboring harmonic bands that are not well estimated by the phase reconstruction along the time axis due to the low local SNRs. To recover the phase in these grooves, we employ the linear phase assumption [6] to reconstruct the phase along the frequency axis.

The main idea is that a speech signal can be represented by a sinusoidal model [40]. We first define the frequency index set of the harmonic bands at the l-th frame as $\mathcal{K}_l$, which is an index set of the local peaks of the estimated magnitude spectrogram $|\hat{X}(k,l)|$ along the frequency axis:

$$\mathcal{K}_l := \left\{ k \middle| \forall k : |\hat{X}(k,l)| > |\hat{X}(k \pm 1, l)| \right\}. \qquad (16)$$

where the symbol ":=" denotes the term "defined as" in mathematics. Then, after applying the STFT with a symmetric analysis STFT window $w(n)$ to the speech signal, the phases between two *harmonic bands*, supposed to be k_1 and k_2 ($\{k_1, k_2\} \subseteq \mathcal{K}_l$), are recalculated by the following equation:

$$\hat{\phi}(k,l) \approx \arg\left(\hat{X}(k_1,l)\frac{W(k-k_1)}{W(0)} + \hat{X}(k_2,l)\frac{W(N+k-k_2)}{W(0)}\right) \qquad (17)$$

where $k_1 < k < k_2$, $W(k)$ is the DFT of the analysis window $w(n)$ at the k-th frequency band, $\hat{X}(k_1,l)$ and $\hat{X}(k_2,l)$ are the estimated STFT spectrograms of the harmonic bands at the stage of phase reconstruction along the time axis:

$$\hat{X}(k_u,l) = |\hat{X}(k_u,l)|e^{j\hat{\phi}(k_u,l)}, \quad \forall k_u \in \mathcal{K}_l \qquad (18)$$

where $\hat{\phi}(k_u,l)$ is calculated by Eq. (13), and $|\hat{X}(k_u,l)|$ is the estimated magnitude which is calculated by masking the noisy magnitude $Y(k_u,l)$ with the estimated magnitude mask $\hat{M}_x(k_u,l)$ produced from DNN. The derivation of Eq. (17) is presented in Appendix B.

4) Overview of the Post-Processing Method: The post-processing method is summarized in Algorithm 1. To illustrate the effectiveness of Algorithm 1, we visualize the distance between the reconstructed phase and the original clean phase by $|\mathrm{principle}(\hat{\phi}(k,l) - \phi_x(k,l))|$ in Fig. 8, where the dark regions indicate small deviations. In the initialization stage, the dark regions are the high SNR regions identified by the estimated RM (Fig. 8a). After the phase reconstruction along the time axis, the dark regions are expanded horizontally (Fig. 8b). After the phase reconstruction along the frequency axis, the bright spots inside the dark regions are reduced, especially at the low frequency bands (Fig. 8c).

IV. Experiments

In this section, we first present the experimental settings in Section IV-A and then present the main results with the stan-

Algorithm 1: Post-Processing for Phase Reconstruction.

```
Input: Estimated IFD $\hat{\Omega}$, estimated magnitude mask $\hat{M}$,
       noisy magnitude $|Y(k,l)|$, and noisy phase $\phi_y(k,l)$
Output: Estimated phase $\hat{\phi}(k,l)$
1: Compute the estimated IF $\hat{\mathrm{IF}}(k,l)$ from Eq. (9)
2: Initialize phase estimates $\hat{\phi}_{\mathrm{init}}(k,l)$ from noisy phases
   by Eq. (11)
3: Get the frame-conditioned estimates of phase along the
   time axis, i.e., $\hat{\phi}^i(k,l)$, by Eq. (12)
4: Get the phase estimate $\hat{\phi}(k,l)$ by integrating the
   frame-conditioned estimates by Eq. (13)
5: Get the enhanced spectral magnitude $|\hat{X}|$ by masking
   the noisy spectral magnitude $|Y|$ with the output mask
   of DNN $\hat{M}$, where the DNN takes an ideal mask
   defined in Table I as the training target
6: Get the index set of the harmonic bands $\mathcal{K}$ by Eq. (16)
7: Recalculate the phase estimate $\hat{\phi}(k,l)$ around the
   harmonic bands along the frequency axis by Eq. (17)
```

dard feedforward DNN in Section IV-B. To study the effectiveness of the IFD-based phase estimation in detail, we analyze the effectiveness of the IFD-based phase estimation beyond the multi-objective DNN framework in Section IV-C, then study the effects of the hyperparameters on performance in Section IV-D, and at last report the results of the proposed method with long short-term memory networks in Section IV-E.

A. Experimental Settings

We first evaluated our algorithm on the TIMIT corpus [41] where each speaker utters 10 clean utterances. For each gender, we selected 136 speakers for training and 56 speakers for testing, which produced 1360 clean training utterances and 560 clean test utterances, respectively. We selected the babble, factory1, factory2, and buccaneer1 noises from the NOISEX-92 [42] database, and separated each noise signal to two parts, one for constructing training mixtures and the other for test. We mixed each clean utterance with 20 short noise segments at the SNR levels of -5, -3 and 0 dB respectively, where the 20 noise segments came from the 4 types of noises, each with 5 random noise segments. We mixed each clean test utterance with 1 short noise segment at the same SNR level as the training mixtures. Eventually, for each SNR level, we have 27200 training mixtures and 2240 test mixtures.

The training and test noise environments of the TIMIT corpus in the above experimental setting is matching. To study the generalization ability of the proposed methods, we further trained DNN models with all four kinds of noises in the training set and both genders at an SNR of -3 dB, and then evaluated the models in an unseen noise environment—destroyer operation room in NOISEX-92 at the same SNR.

We also conducted an experiment on the simulated training data of the 4th CHiME speech separation and recognition challenge (CHiME-4) [43]. CHiME-4 has 4 noisy environments which are the bus, cafe, pedestrian area, and street junction noises respectively, with 17 noise segments in total. We selected 50 and 4 speakers from the WSJ0 SI-84 training set of CHiME-4 for training and test respectively. We selected 1 noise segment from each noisy environment respectively as the additive noise of the test data, and used the remaining 13 noise segments as the additive noise of the training data. The genders in both the training and test speakers are also balanced. Finally, we have 4313 training mixtures and 1012 test mixtures.

We resampled all corpora to 16 kHz, and extracted 512-dimensional STFT features with the frame length set to 20 ms and the frame shift set to 5 ms, where the Hamming window is used as the analysis window of STFT.

As will be shown later, the difference between any two comparison methods is that their DNNs have different training targets. **Hence, we use the name of the training target of a comparison method to represent the method itself in this section.** For example, we use IRM+ IFD to represent the proposed phase-aware DNN that takes IRM and IFD as the training target. We denote the phase-aware DNN method with different magnitude mask estimation methods in Table I as **IRM+ IFD**, **IAM+ IFD**, and **PSF+ IFD** respectively. We set the length of the Hamming window, i.e., $(2N_s + 1)$ in Eq. (13), to 5, and set the window length of the input frames of the multi-objective DNN, i.e., $(2W + 1)$, to 5.

We compared with the following methods:

- **IRM:** DNN that takes the ideal ratio mask as the training target.
- **IAM:** DNN that takes the ideal amplitude mask as the training target.
- **PSF:** phase-sensitive filter [12].
- **cIRM:** complex ideal ratio mask [16].

We also reported oracle performance "*oracle*" which is produced from a method that first uses the above IRM-based DNN method to enhance the magnitude spectrograms and then reconstructs the signals with the clean phase spectrograms. We reported a lowerbound "*noisy*" as well, which evaluates the original noisy data.

We evaluated the performance from the perspective of speech quality and intelligibility. To show the effect of incorporating phase processing into speech enhancement, we used phased-aware metrics, including the perceptual evaluation of speech quality (PESQ) [44], [45], short-time objective intelligibility (STOI) [46], extended STOI (ESTOI) [47] and singal to distortion ratio (SDR) [48]. PESQ is a phase-aware metric for speech quality. Its score ranges from -0.5 to 4.5. The higher the PESQ score is, the better the predicted speech quality is. STOI evaluates the objective intelligibility of a degraded speech signal by computing the correlation of the temporal envelopes of the degraded speech signal and its clean reference. It has been shown empirically that STOI scores are strongly correlated with human speech intelligibility scores. ESTOI evaluates the objective intelligibility of a degraded speech signal by computing the spectral correlation coefficients of the degraded speech signal and its clean reference in short time segments. Unlike STOI, ESTOI does not assume that frequency bands are mutually independent. Both STOI and ESTOI scores range from 0 to 1. The higher the scores are, the better the predicted intelligibility is.

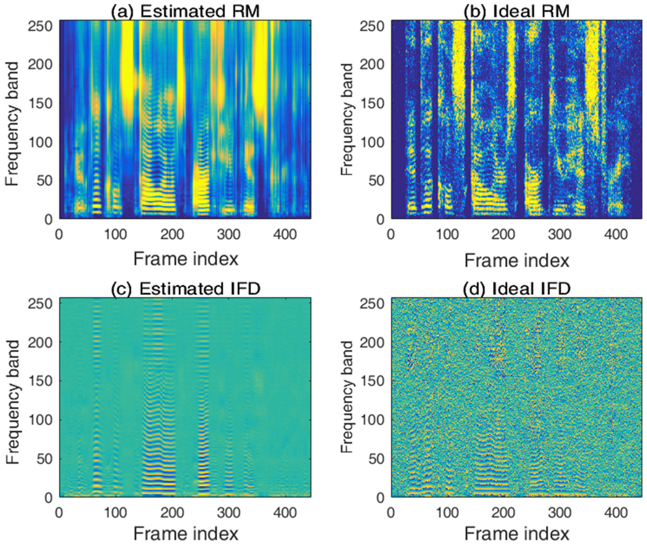

Fig. 9. Comparison between the outputs of the proposed method and the ground-truth labels in the factory1 noise at the SNR of −3 dB.

The SDR scores are computed by blind source separation evaluation measurements [48]. It has been widely used for evaluating speech quality.

B. Main Results

We show an output pattern of the proposed method in Fig. 9. From the figure, we observe that the estimated IFD values in the silent regions are mostly zero, which indicates that the silent regions have no structure.

We reported the comparison results[3] on males and females of TIMIT in Tables II and III respectively, where MA was used as the training loss function of all DNN models. From the tables, we observe the following experimental phenomena. (i) The proposed phase-aware methods improve the speech quality and intelligibility over their magnitude-based counterparts. That is to say, IRM+IFD outperforms IRM, IAM+IFD outperforms IAM, and PSF+IFD outperforms PSF. (ii) The overall performance of the proposed methods is higher than that of the comparison methods. For example, the SDR and ESTOI scores of the proposed methods are at least 0.4 and 0.02 higher than those of the comparison methods respectively on the female speakers at the SNR of 0 dB, and around 0.3 and 0.01 higher respectively on the male speakers at the SNR of 0 dB. (iii) The performance improvement on the female speakers is larger than that on the male speakers. (iv) The performance improvement in the three nonvocal noises, i.e., factory1, factory2, and buccaneer1 noises, is higher than that in the babble noise, since that the phase structure of the speech signals is different from the phase structures of the nonvocal noises, but similar with the phase structure of the babble noise.

The generalization performance of the comparison methods is shown in Table IV, where all DNN models were trained with the data from the both genders and in all four types of the noises at the SNR of −3 dB. The result on the "destroyer operation room" test noise scenario shows that the proposed methods improve the speech intelligibility.

We also reported the comparison results on CHiME-4 in Tables V and VI, where the loss functions were set to MA and MA+mSA respectively. The results indicate that the experimental conclusions are consistent across different loss functions.

C. Analysis of IFD-Based Phase Estimation

In this subsection, we investigate the effectiveness of the IFD-based phase estimation beyond the multi-objective DNN framework on both TIMIT and CHiME-4, so as to identify how much the phase enhancement component contributes to the performance improvement.

1) IFD-Based Phase Estimation With Noisy Magnitude: We evaluated a method that uses the enhanced phase spectrograms and noisy magnitude spectrograms to resynthesize signals in the time domain. The experiment was conducted on the female speakers of TIMIT in the factory1 noise environment at the SNR of 0 dB. The training target of DNN was IAM+IFD. Note that, because the magnitude mask estimation is a requirement for the proposed phase estimation method, here we still used it for the phase estimation, but not for the magnitude enhancement.

Figure 10 shows the experimental result, where the effects of the phase reconstruction along the time axis, phase reconstruction along the frequency axis, and their combinations are shown separately. As a comparison, we also show the performance of the *phase reconstruction along frequency* method in [6] which have the best performance among the three proposed methods in [6]. The comparison method [6] assumes that the phase shift between two adjacent bands that are close to their harmonic bands is a constant, which does not use magnitude information for its phase reconstruction. The comparison result shows that our phase reconstruction with the IFD estimates improves the speech quality and intelligibility.

As a complementary experiment, we applied the enhanced magnitude spectrograms to both our method and the method in [6] for the resynthesis of the time-domain signals. The comparison result in Table VII shows that our IAM+IFD performs better than the method in [6].

2) IFD-Based Phase Estimation With Oracle Magnitude Masks: We evaluated a method that uses the estimated IFD and oracle magnitude masks to resynthesize signals in the time domain, where we used the oracle magnitude masks for the initialization of the phase reconstruction stage of the IFD-based phase estimation. The experiment was conducted on the CHiME-4 corpus. Experimental result in Table VIII demonstrates the effectiveness of the proposed IFD-based phase estimation given oracle magnitude masks.

3) IFD-Based Phase Estimation With Single-Objective Networks: We compared the multi-objective network with a method that estimates the magnitude masks and IFD *separately* by two independent DNNs on CHiME-4. The comparison result in Table IX shows that training two separate networks yields similar performance with the multi-objective network in terms

[3] Some demos are uploaded to https://github.com/njzheng/speech-enhancement-DNN/blob/master/DemoPackage.zip.

TABLE II
PERFORMANCE COMPARISON ON THE FEMALE SPEAKERS OF THE TIMIT CORPUS. THE NUMBERS IN BOLD DENOTE THE BEST PERFORMANCE AMONG THE COMPARISON METHODS

SNR (dB)	Methods	Babble				Factory1				Factory2				Buccaneer1			
		PESQ	ESTOI	STOI	SDR	PESQ	ESTOI	STOI	SDR	PESQ	ESTOI	STOI	SDR	PESQ	ESTOI	STOI	SDR
-5	*noisy*	1.17	0.268	0.534	-4.81	1.09	0.251	0.518	-4.81	1.36	0.369	0.635	-4.82	1.05	0.224	0.506	-4.81
	IRM	1.51	0.435	0.648	1.47	1.64	0.450	0.671	3.55	2.03	0.597	0.782	6.52	2.12	0.549	0.763	5.92
	IRM+IFD (proposed)	**1.55**	0.448	0.653	1.70	1.71	0.469	0.680	4.04	2.11	0.618	0.793	7.09	2.21	0.573	**0.773**	6.53
	IAM	1.53	0.446	0.659	1.83	1.66	0.464	0.685	3.87	2.05	0.611	0.793	6.91	2.11	0.558	0.770	6.03
	IAM+IFD (proposed)	**1.57**	**0.460**	**0.665**	2.05	**1.72**	**0.481**	**0.692**	4.29	2.15	**0.633**	**0.804**	7.42	2.20	**0.583**	**0.779**	6.59
	PSF	1.49	0.426	0.634	2.60	1.67	0.447	0.656	5.12	2.11	0.592	0.776	8.01	2.13	0.535	0.748	7.42
	PSF+IFD (proposed)	1.53	0.442	0.641	2.85	**1.74**	0.467	0.669	5.60	**2.20**	0.614	0.787	**8.53**	**2.24**	0.559	0.759	**7.96**
	cIRM	1.53	0.410	0.629	**3.68**	1.66	0.429	0.650	**5.99**	2.13	0.571	0.768	**8.57**	2.14	0.529	0.743	**7.95**
	oracle	1.94	0.587	0.731	6.41	2.03	0.596	0.750	8.68	2.48	0.736	0.852	12.03	2.59	0.688	0.827	11.85
-3	*noisy*	1.31	0.321	0.581	-2.85	1.22	0.303	0.564	-2.86	1.51	0.422	0.679	-2.87	1.16	0.272	0.549	-2.86
	IRM	1.71	0.501	0.704	3.45	1.81	0.516	0.725	5.26	2.20	0.646	0.816	8.08	2.25	0.599	0.792	7.12
	IRM+IFD (proposed)	**1.76**	0.518	0.711	3.78	1.89	0.536	0.734	5.75	2.31	0.671	0.828	8.71	2.35	0.624	**0.803**	7.75
	IAM	1.73	0.513	0.716	3.83	1.83	0.530	0.737	5.57	2.23	0.660	0.827	8.51	2.24	0.610	0.801	7.28
	IAM+IFD (proposed)	**1.78**	**0.530**	**0.722**	4.11	**1.92**	**0.550**	**0.746**	6.06	2.33	**0.684**	**0.837**	9.01	2.35	**0.635**	**0.810**	7.87
	PSF	1.72	0.495	0.694	4.63	1.87	0.513	0.714	6.73	2.31	0.645	0.812	9.50	2.26	0.589	0.781	8.53
	PSF+IFD (proposed)	**1.77**	0.515	0.703	4.92	**1.93**	0.536	0.725	7.27	**2.40**	0.667	0.824	**10.02**	**2.39**	0.614	0.793	**9.09**
	cIRM	1.75	0.480	0.689	**5.53**	1.87	0.504	0.712	**7.55**	2.32	0.630	0.805	**10.02**	2.28	0.594	0.784	**9.06**
	oracle	1.71	0.645	0.781	8.38	2.20	0.654	0.797	10.43	2.65	0.774	0.878	13.62	2.73	0.726	0.852	12.99
0	*noisy*	1.53	0.402	0.652	0.10	1.44	0.391	0.640	0.10	1.75	0.503	0.741	0.09	1.35	0.350	0.618	0.10
	IRM	1.99	0.597	0.778	6.23	2.07	0.608	0.791	7.69	2.45	0.717	0.858	10.40	2.44	0.668	0.831	8.97
	IRM+IFD (proposed)	2.07	0.617	0.787	6.66	2.16	0.631	0.802	8.25	2.58	0.740	0.869	10.98	2.57	0.694	**0.843**	9.60
	IAM	2.01	0.610	0.789	6.63	2.09	0.622	0.802	8.01	2.48	0.729	0.867	10.79	2.44	0.679	0.840	9.17
	IAM+IFD (proposed)	**2.10**	**0.632**	**0.798**	7.07	2.18	**0.645**	**0.812**	8.58	2.61	**0.754**	**0.877**	11.35	2.57	**0.704**	**0.850**	9.75
	PSF	2.03	0.595	0.773	7.38	2.15	0.610	0.788	9.11	2.57	0.718	0.858	11.64	2.48	0.662	0.827	10.26
	PSF+IFD (proposed)	**2.11**	0.617	0.784	**7.88**	**2.24**	0.634	0.800	**9.64**	**2.69**	0.740	0.868	**12.17**	**2.62**	0.688	0.838	**10.83**
	cIRM	2.07	0.583	0.769	**8.08**	2.17	0.616	0.795	**9.82**	2.58	0.707	0.852	11.96	2.52	0.680	0.837	**10.80**
	oracle	2.43	0.725	0.844	11.25	2.46	0.729	0.852	12.91	2.92	0.825	0.909	15.96	2.94	0.779	0.883	14.76

TABLE III
PERFORMANCE COMPARISON ON THE MALE SPEAKERS OF THE TIMIT CORPUS

SNR (dB)	Methods	Babble				Factory1				Factory2				Buccaneer1			
		PESQ	ESTOI	STOI	SDR	PESQ	ESTOI	STOI	SDR	PESQ	ESTOI	STOI	SDR	PESQ	ESTOI	STOI	SDR
-5	*noisy*	1.52	0.278	0.546	-4.80	1.37	0.262	0.535	-4.81	1.65	0.377	0.649	-4.83	1.24	0.227	0.517	-4.80
	IRM	1.68	0.420	0.645	0.11	1.81	0.450	0.688	2.52	2.24	0.594	0.797	5.54	2.30	0.565	0.791	4.94
	IRM+IFD (proposed)	**1.70**	0.426	0.648	0.40	**1.85**	0.457	0.693	3.01	**2.28**	0.604	0.803	5.96	2.33	0.574	0.795	5.51
	IAM	**1.70**	**0.433**	**0.656**	0.35	1.81	0.450	0.688	2.52	2.25	0.603	**0.805**	5.77	2.29	0.573	**0.799**	5.04
	IAM+IFD (proposed)	**1.72**	**0.438**	**0.660**	0.64	**1.85**	**0.465**	**0.703**	3.16	**2.30**	**0.611**	**0.810**	6.16	**2.32**	**0.581**	**0.802**	5.55
	PSF	1.63	0.418	0.634	1.13	1.80	0.451	0.673	3.95	2.25	0.596	0.791	6.87	2.28	0.566	0.782	6.47
	PSF+IFD (proposed)	1.64	0.424	0.637	1.36	1.82	0.456	0.678	4.44	**2.31**	0.602	0.796	**7.37**	**2.35**	**0.578**	0.789	**7.05**
	cIRM	1.63	0.412	0.632	**1.72**	1.73	0.439	0.670	**4.60**	2.24	0.583	0.787	**7.28**	2.21	0.541	0.760	6.77
	oracle	1.98	0.550	0.717	5.31	2.08	0.580	0.759	7.96	2.56	0.726	0.861	11.36	2.64	0.689	0.848	11.41
-3	*noisy*	1.64	0.328	0.594	-2.85	1.49	0.313	0.582	-2.86	1.78	0.428	0.693	-2.88	1.34	0.275	0.561	-2.85
	IRM	1.88	0.485	0.704	2.14	2.00	0.512	0.738	4.23	2.40	0.644	0.830	7.11	2.44	0.611	0.816	6.16
	IRM+IFD (proposed)	**1.91**	0.492	0.708	2.40	**2.04**	0.522	0.745	4.72	2.45	0.655	0.836	7.57	**2.48**	0.622	0.822	6.74
	IAM	**1.91**	**0.499**	**0.717**	2.44	2.00	0.523	0.750	4.50	2.42	0.655	**0.839**	7.42	2.43	0.619	**0.826**	6.32
	IAM+IFD (proposed)	**1.93**	**0.506**	**0.720**	2.71	**2.03**	**0.533**	**0.756**	4.96	**2.47**	**0.664**	**0.844**	7.81	**2.47**	**0.630**	**0.829**	6.82
	PSF	1.85	0.485	0.696	3.15	2.00	0.517	0.730	5.63	2.44	0.650	0.828	8.41	2.43	0.614	0.811	7.58
	PSF+IFD (proposed)	1.87	0.491	0.700	3.43	**2.02**	0.523	0.735	**6.16**	**2.48**	0.656	0.833	**8.85**	**2.49**	**0.627**	0.818	**8.14**
	cIRM	1.86	0.480	0.696	**3.86**	1.95	0.506	0.727	**6.21**	2.42	0.638	0.824	**8.74**	2.36	0.597	0.797	7.88
	oracle	2.19	0.613	0.772	7.26	2.27	0.637	0.804	9.69	2.72	0.767	0.888	12.95	2.78	0.727	0.870	12.50
0	*noisy*	1.83	0.407	0.666	0.10	1.69	0.395	0.655	0.10	1.99	0.505	0.754	0.08	1.52	0.354	0.633	0.10
	IRM	2.17	0.584	0.782	5.11	2.26	0.602	0.803	6.72	2.65	0.712	0.870	9.45	2.65	0.675	0.852	8.05
	IRM+IFD (proposed)	2.21	0.593	0.788	5.48	**2.31**	0.613	0.810	7.19	2.70	0.722	0.876	9.82	**2.70**	0.688	0.858	8.59
	IAM	2.20	0.597	0.794	5.49	2.27	0.615	0.815	7.03	2.67	0.721	0.878	9.80	2.64	0.684	**0.860**	8.25
	IAM+IFD (proposed)	**2.24**	**0.609**	**0.802**	5.83	**2.31**	**0.625**	**0.821**	7.46	2.72	**0.733**	**0.884**	10.10	2.68	**0.696**	**0.865**	8.70
	PSF	2.17	0.588	0.781	6.21	2.29	0.610	0.801	8.01	2.70	0.721	0.872	10.62	2.66	0.681	0.850	9.31
	PSF+IFD (proposed)	2.20	0.598	0.787	6.56	**2.32**	**0.621**	0.810	**8.53**	**2.75**	**0.729**	0.877	**10.97**	**2.71**	**0.694**	0.857	**9.83**
	cIRM	2.18	0.583	0.781	**6.71**	2.26	0.603	0.799	**8.47**	2.68	0.717	0.869	**11.04**	2.58	0.678	0.849	9.62
	oracle	2.48	0.702	0.842	10.33	2.54	0.716	0.859	12.21	2.97	0.819	0.919	15.26	2.99	0.780	0.899	14.19

of PESQ, ESTOI, and STOI, and performs better than the latter in terms of SDR. However, the proposed method is more efficient than the method of training two separate deep networks.

D. Effects of the Hyperparameter N_s in Eq. (13) on Performance

Hyperparameter N_s is the window length in Eq. (13). In our above experiments, we have set N_s to 2. In this subsection, we analyze the effect of N_s on performance. Specifically, we plot the relative performance improvement of IAM+IFD over IAM with respect to different N_s in Fig. 11. From the figure, we observe that setting N_s to a relatively large window length leads to improved performance in terms of PESQ, ESTOI and STOI, and a performance drop in terms of SDR especially at high SNR levels. This phenomenon, which was caused by the time shifts introduced by the phase reconstruction, can be explained from Eq. (12) where large windows may violate the

TABLE IV
AVERAGE PERFORMANCE OF THE COMPARISON METHODS OVER BOTH GENDERS IN THE DESTROYER OPERATION ROOM NOISE TEST ENVIRONMENT ON THE TIMIT CORPUS AT −3 dB

Methods	PESQ	ESTOI	STOI	SDR
noisy	1.48	0.356	0.625	-2.85
IRM	1.61	0.491	0.725	3.60
IRM+IFD (proposed)	**1.69**	**0.498**	0.729	3.69
IAM	1.62	0.493	0.730	3.43
IAM+IFD (proposed)	1.65	**0.498**	**0.734**	3.39
PSF	1.59	0.490	0.719	4.64
PSF+IFD (proposed)	1.64	0.496	0.721	4.87
cIRM	**1.68**	0.466	0.705	**5.50**
oracle	1.96	0.618	0.789	8.97

TABLE V
PERFORMANCE OF THE COMPARISON METHODS WITH THE MA LOSS FUNCTION ON THE CHiME-4 CORPUS

Methods	PESQ	ESTOI	STOI	SDR
noisy	2.29	0.596	0.827	4.42
IRM	2.78	0.748	0.889	9.82
IRM+IFD (proposed)	**2.82**	0.756	0.893	10.15
IAM	2.81	0.756	0.893	10.25
IAM+IFD (proposed)	**2.85**	**0.763**	**0.897**	10.49
PSM	2.76	0.755	0.891	10.74
PSM+IFD (proposed)	2.80	**0.763**	**0.895**	**11.07**
cIRM	2.74	0.747	0.888	10.88
oracle	3.15	0.822	0.917	15.12

TABLE VI
PERFORMANCE OF THE COMPARISON METHODS WITH THE MA+mSA LOSS FUNCTION ON THE CHiME-4 CORPUS

Methods	PESQ	ESTOI	STOI	SDR
noisy	2.29	0.596	0.827	4.42
IRM	2.82	0.767	0.898	10.64
IRM+IFD (proposed)	**2.87**	**0.774**	**0.902**	10.93
IAM	2.82	0.767	**0.899**	10.68
IAM+IFD (proposed)	**2.86**	**0.773**	**0.902**	10.97
PSM	2.70	0.766	0.896	11.38
PSM+IFD (proposed)	2.75	**0.772**	**0.899**	**11.61**
oracle	3.15	0.836	0.924	16.29

short-time stability of the phase, and the corresponding errors caused by this violation will be accumulated during the phase reconstruction in Eq. (13). To illustrate this error accumulation problem more apparently, we showed a reconstructed utterance in Fig. 12 where N_s was set to 7. From the figure, we observe a clear time shift between the reconstructed utterance and the original clean utterance.

E. Results With LSTM and BLSTM Networks

All of the above experiments were conducted with the standard feedforward deep neural networks. To our knowledge, the state-of-the-art deep models for speech enhancement are long-short term memory (LSTM) and bidirectional long-short term memory (BLSTM) [12]. In this subsection, we investigated the performance of the proposed method with the LSTM or BLSTM network on CHiME-4. The parameter settings of the two networks are as follows. The adaptive gradient descent algorithm

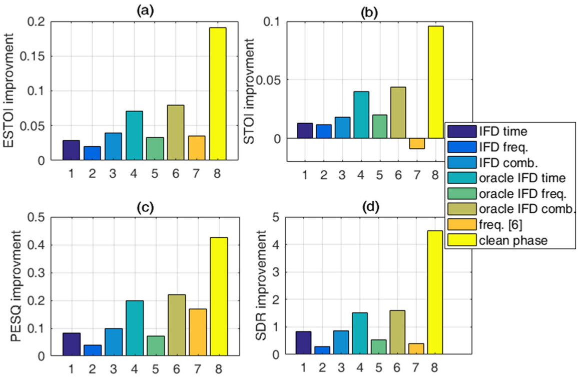

Fig. 10. Performance of the speech signals resynthesized from the IFD-based phase estimates and the noisy magnitude spectrograms, where the experiment was conducted on the females of TIMIT in the factory1 noise at the SNR of 0 dB. The legends "IFD time", "IFD freq.", and "IFD comb." denote the phase reconstruction along the time axis, phase reconstruction along the frequency axis, and their combinations of the proposed IFD-based phase estimation, respectively. As a comparison, the "oracle IFD" legends mean that the phase is calculated from the oracle IFD and ideal magnitude mask (instead of the output of the multi-objective DNN); the legend "freq. [6]" denotes the phase reconstruction along frequency method in [6].

TABLE VII
PERFORMANCE COMPARISON OF THE PROPOSED IAM+IFD AND THE PHASE RECONSTRUCTION METHOD IN [6] GIVEN THE ENHANCED MAGNITUDE SPECTROGRAMS (i.e., IAM+FREQ.)

Methods	PESQ	ESTOI	STOI	SDR
noisy	1.44	0.39	0.640	0.100
IAM+freq. [6]	2.12	0.62	0.790	7.730
IAM+IFD (proposed)	**2.19**	**0.65**	**0.810**	**8.580**

TABLE VIII
PERFORMANCE OF THE IFD-BASED PHASE ESTIMATION WITH ORACLE MAGNITUDE MASKS ON THE CHiME-4 CORPUS, WHERE THE MA LOSS FUNCTION ARE ADOPTED BY ALL DEEP MODELS

Methods	PESQ	ESTOI	STOI	SDR
noisy	2.29	0.596	0.827	4.42
oracle IRM with noisy phase	3.66	0.904	0.957	12.54
oracle IRM+IFD	3.84	0.917	0.963	13.14
oracle IAM with noisy phase	3.71	0.925	0.969	13.12
oracle IAM+IFD	**3.89**	**0.938**	**0.975**	13.53
oracle PSM with noisy phase	3.75	0.908	0.960	14.96
oracle PSM+IFD	**3.86**	0.919	0.965	**15.35**

TABLE IX
PERFORMANCE COMPARISON BETWEEN THE MULTI-OBJECTIVE DNN AND THE METHOD THAT ESTIMATES THE IFD AND MAGNITUDE MASKS SEPARATELY BY TWO INDEPENDENT DNNs ON CHiME-4, WHERE THE MA LOSS FUNCTION ARE ADOPTED BY ALL DEEP MODELS

Methods	PESQ	ESTOI	STOI	SDR
noisy	2.29	0.596	0.827	4.42
IRM	2.78	0.748	0.889	9.82
IRM+IFD (proposed)	2.82	0.756	0.893	10.15
IRM+IFD (separate)	2.83	0.758	0.894	10.28
IAM	2.81	0.756	0.893	10.25
IAM+IFD (proposed)	**2.85**	0.763	**0.897**	10.49
IAM+IFD (separate)	**2.86**	**0.765**	**0.897**	10.66
PSM	2.76	0.755	0.891	10.74
PSM+IFD (proposed)	2.80	0.763	0.895	11.07
PSM+IFD (separate)	2.79	**0.764**	0.895	**11.14**
oracle	3.26	0.837	0.926	15.73

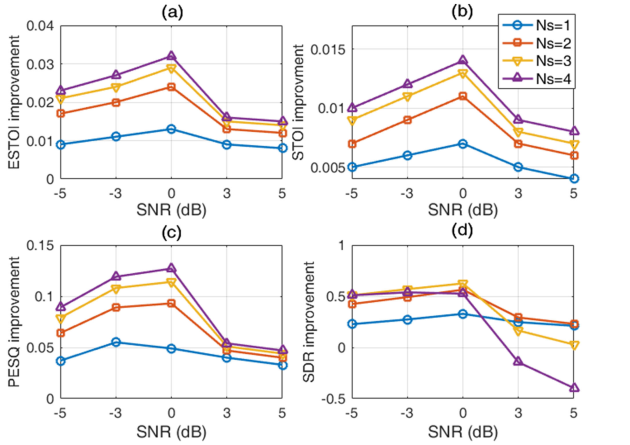

Fig. 11. Relative performance improvement of IAM+IFD over IAM at different Hamming half-window lengths N_s on the females at the factory1 noise.

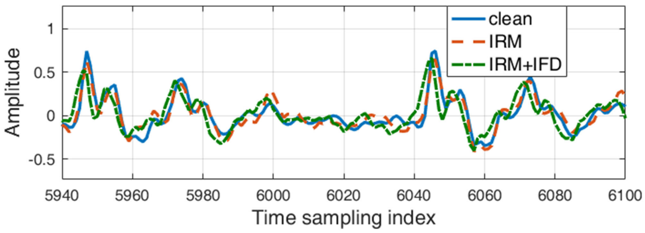

Fig. 12. The phenomenon of time shift generated by the proposed method.

TABLE X
PERFORMANCE OF THE COMPARISON METHODS WITH THE LSTM NETWORK THAT ADOPTS THE MA LOSS FUNCTION ON THE CHIME-4 CORPUS

Methods	PESQ	ESTOI	STOI	SDR
noisy	2.29	0.596	0.827	4.42
IRM	2.77	0.737	0.885	10.08
IRM+IFD (proposed)	2.85	0.753	0.893	10.22
IAM	2.81	0.748	0.891	10.41
IAM+IFD (proposed)	2.83	**0.759**	**0.897**	10.55
PSM	2.85	0.749	0.889	11.07
PSM+IFD (proposed)	**2.89**	**0.760**	**0.894**	**11.24**
cIRM	**2.89**	0.757	0.892	11.03
oracle	3.13	0.819	0.917	15.41

was used to train the networks. The initial learn rate was set to 0.001. The maximum number of epochs was set to 80. The mini-batch size was set to 50. Each network has two hidden layers with 256 hidden units per layer. For the BLSTM network, half of the hidden units per layer were used for the forward direction, and the other half were used for the backward direction.

The comparison results on CHiME-4 are listed in Tables X and XI. From the two tables and Table V, we observe that the effectiveness of the proposed method is consistent across different types of deep models. Moreover, LSTM reaches higher SDR than the standard DNN, and BLSTM performs the best among the three models.

We have also analyzed the effectiveness of the IFD-based phase estimation that adopts BLSTM as the base deep model in a similar experimental setting with Section IV-C. The experimental conclusions with DNN and BLSTM are consistent.

TABLE XI
PERFORMANCE OF THE COMPARISON METHODS WITH THE BLSTM NETWORK THAT ADOPTS THE MA LOSS FUNCTION ON THE CHIME-4 CORPUS

Methods	PESQ	ESTOI	STOI	SDR
noisy	2.29	0.596	0.827	4.42
IRM	2.85	0.755	0.893	10.18
IRM+IFD (proposed)	2.87	0.762	0.897	10.27
IAM	2.86	0.765	0.899	10.45
IAM+IFD (proposed)	2.91	**0.775**	**0.903**	10.57
PSM	2.92	0.767	0.895	11.30
PSM+IFD (proposed)	**2.95**	**0.772**	**0.899**	**11.37**
cIRM	**2.96**	0.768	0.897	11.21
oracle	3.26	0.837	0.926	15.73

V. CONCLUSIONS

In this paper, we have proposed a phase-aware speech enhancement method based on DNN. The method introduces a new training target for DNN-based speech enhancement methods, named IFD, which is the derivative of the clean phase spectrogram along the time axis. In the training stage, it optimizes IFD and a magnitude mask simultaneously in a multi-objective learning framework. In the test stage, it first uses noisy phase as an initial phase estimate, and then conducts phase reconstruction along the time and frequency axes with the estimated IFD and the estimated magnitude mask. To our knowledge, this method is the first DNN-based approach that directly processes phase spectrograms which appear to be randomly distributed and highly unstructured. Moreover, the proposed new target is a general one. It can be adopted by any algorithms who use magnitude spectrograms or their variants as targets. The proposed post-processing method also provides a robust way for reconstructing the phase spectrograms under the difficult situation where the initial point of the derivatives of the phase spectrograms is unknown. We have evaluated the proposed method in several adverse environments at low SNR levels. The experimental results show that the proposed method outperforms the counterparts that do not conduct phase processing, in terms of both speech quality and intelligibility.

APPENDIX A

A sinusoid signal $s(n)$ in time domain is written by

$$s\left(n\right) = 2A\left(n\right)\cos\left(\Omega \cdot n + \varphi\right) \tag{19}$$

where A is the magnitude, Ω is the normalized angular frequency and φ is the initial phase. The STFT of $s(n)$ in the dominant frequency band $k = \frac{\Omega}{2\pi}N$ can be derived as

$$\begin{aligned} S\left(k,l\right) &= \sum_{n=0}^{N-1} s\left(lL+n\right) w\left(n\right) e^{-j\frac{2\pi}{N}kn} \\ &= Ae^{j\left(\Omega lL+\varphi\right)} \sum_{n=0}^{N-1} w\left(n\right). \end{aligned} \tag{20}$$

From Eq. (20), we can see that the phase at the dominant frequency is a linear function of the frame index l with slope ΩL.

Appendix B

We first represent a short-time speech segment signal $x(n)$ as a sum of sinusoids [40]:

$$x(n) = \sum_{h=0}^{H-1} 2A_h(n)\cos(\Omega_h \cdot n + \varphi_h) \tag{21}$$

where H is the number of harmonics, h denotes the harmonic index, $2A_h$ is the corresponding real magnitude, and Ω_h is the normalized harmonic frequency.

We then calculate the STFT of $x(n)$ with a causal symmetric STFT analysis window $w(n)$, where we assume that the length of DFT N is long enough to resolve the harmonics for typical sounds and $w(n)$. Suppose there are two adjacent harmonics with their normalized frequencies as $\Omega_1 = \frac{2\pi}{N}k_1$ and $\Omega_2 = \frac{2\pi}{N}k_2$ respectively.

The T-F units of the STFT spectrogram between the two harmonic bands k_1 and k_2, denoted by $X(k,l)$, can be represented by:

$$\begin{aligned} X(k,l) &= \sum_{n=0}^{N-1} x(lL+n)\,w(n)\,e^{-j\frac{2\pi}{N}kn} \\ &\approx \sum_{n=0}^{N-1}\left(A_1(lL+n)\,e^{j(\Omega_1\cdot(lL+n)+\varphi_1)}\right)w(n)\,e^{-j\frac{2\pi}{N}kn} \\ &\quad + \sum_{n=0}^{N-1}\left(A_2(lL+n)\,e^{j(\Omega_2\cdot(lL+n)+\varphi_2)}\right)w(n)e^{-j\frac{2\pi}{N}kn} \\ &\approx A_{k_1,l}e^{j(\Omega_1 lL+\varphi_1)}\sum_{n=0}^{N-1} w(n)\,e^{j\left(\Omega_1-\frac{2\pi}{N}k\right)n} \\ &\quad + A_{k_2,l}e^{j(\Omega_2 lL+\varphi_2)}\sum_{n=0}^{N-1} w(n)e^{j\left(\Omega_2-\frac{2\pi}{N}k\right)n}, \\ &\qquad \forall k_1 < k < k_2 \end{aligned} \tag{22}$$

where the approximation is made under the following two assumptions. The first assumption is that there is no interference from other neighbor harmonics. The second assumption is that the amplitude of the h-th harmonic in time domain over the period of a single frame l is a constant, i.e., $A_u(lL+n) \approx A_{k_u,l}, \forall u = 1,2$. Then, with $\Omega_1 = \frac{2\pi}{N}k_1$ and $\Omega_2 = \frac{2\pi}{N}k_2$, Eq. (22) can be simplified to:

$$\begin{aligned} X(k,l) &\approx A_{k_1,l}e^{j(\Omega_1 lL+\varphi_1)}W(k-k_1) \\ &\quad + A_{k_2,l}e^{j(\Omega_2 lL+\varphi_2)}W(N+k-k_2) \end{aligned} \tag{23}$$

where $W(k)$ is the DFT of $w(n)$ which is a complex number.

Given $X(k_u,l) = A_{k_u,l}e^{j(\Omega_u lL+\varphi_u)}W(0)$, the phase of $X(k,l)$ can be approximated from the harmonic bands by:

$$\begin{aligned} \phi_x(k,l) &\approx \arg\Big(A_{k_1,l}e^{j(\Omega_1 lL+\varphi_1)}W(k-k_1) \\ &\quad + A_{k_2,l}e^{j(\Omega_2 lL+\varphi_2)}W(N+k-k_2)\Big) \\ &= \arg\left(A_{k_1,l}e^{j(\Omega_{h1} lL+\varphi_1)}W(0)\,\frac{W(k-k_1)}{W(0)}\right. \\ &\quad \left. + A_{k_2,l}e^{j(\Omega_2 lL+\varphi_2)}W(0)\,\frac{W(N+k-k_2)}{W(0)}\right) \\ &= \arg\left(X(k_1,l)\frac{W(k-k_1)}{W(0)}\right. \\ &\quad \left. + X(k_2,l)\frac{W(N+k-k_2)}{W(0)}\right). \end{aligned} \tag{24}$$

Finally, in the test stage, the phase of the (k,l)-th T-F unit can be estimated by replacing $X(k_i,l)$ in Eq. (24) by its estimate $\hat{X}(k_i,l)$:

$$\begin{aligned} \hat{\phi}_x(k,l) &\approx \arg\left(\hat{X}(k_1,l)\,\frac{W(k-k_1)}{W(0)}\right. \\ &\quad \left. + \hat{X}(k_2,l)\,\frac{W(N+k-k_2)}{W(0)}\right). \end{aligned} \tag{25}$$

Acknowledgment

We thank the associate editor and anonymous reviewers for their helpful comments.

References

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[2] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[3] K. Paliwal, K. Wjcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[4] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Commun.*, vol. 81, pp. 1–29, 2016.

[5] J. L. Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[6] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[7] D. Friedman, "Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1985, pp. 1121–1124.

[8] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2281–2289, Sep. 1992.

[9] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[10] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[11] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

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

[13] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[16] D. S. Williamson, Y. Wang, and D. L. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[17] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2014, pp. 577–581.

[18] F. Weninger *et al.*, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Proc. Int. Conf. Latent Variable Anal., Signal Separation*, 2015, pp. 91–99.

[19] X.-L. Zhang and D. L. Wang, “A deep ensemble learning method for monaural speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[20] K. Han, Y. Wang, and D. L. Wang, “Learning spectral mapping for speech dereverberation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4628–4632.

[21] M. Kolbæk, Z.-H. Tan, and J. Jensen, “Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[22] B. Wu, K. Li, M. Yang, and C.-H. Lee, “A reverberation-time-aware approach to speech dereverberation based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[23] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, “Factored spatial and spectral multichannel raw waveform CLDNNs,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5075–5079.

[24] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, “Improved MVDR beamforming using single-channel mask prediction networks,” in *Proc. INTERSPEECH*, 2016, pp. 1981–1985.

[25] X. Xiao *et al.*, “Deep beamforming networks for multi-channel speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5745–5749.

[26] K. K. Paliwal and A. P. Stark, “Speech analysis using instantaneous frequency deviation,” *Proc. INTERSPEECH*, 2008, pp. 2602–2605.

[27] J. Benedetto, *Applied and Numerical Harmonic Analysis*. Cambridge, MA, USA: Birkhauser, 2013.

[28] J. Le Roux, “Phase-controlled sound transfer based on maximally-inconsistent spectrograms,” in *Proc. Acoustical Soc. Japan Spring Meeting*, Mar. 2011, no. 1-Q-51.

[29] E. Loweimi, J. P. Barker, and T. Hain, “Statistical normalisation of phase-based feature representation for robust speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5310–5314.

[30] S. Shimauchi, S. Kudo, Y. Koizumi, and K. Furuya, “On relationships between amplitude and phase of short-time Fourier transform,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 676–680.

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

[32] Y. Wang, K. Han, and D. Wang, “Exploring monaural features for classification-based speech segregation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.

[33] X.-L. Zhang and J. Wu, “Deep belief networks based voice activity detection,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.

[34] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, “Deep clustering and conventional networks for music separation: Stronger together,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 61–65.

[35] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.

[36] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 315–323.

[37] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” 2012. [Online]. Available: http://arxiv.org/abs/1207.0580.

[38] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 2121–2159, 2011.

[39] P. O’Shea, A. Z. Sadik, and Z. M. Hussain, “Multirate digital signal processing,” in *Digital Signal Processing*. Berlin, Heidelberg, Germany: Springer, 2011, pp. 209–227.

[40] M. E. Deisher and A. S. Spanias, “Speech enhancement using state-based estimation and sinusoidal modeling,” *Acoustical Soc. America J.*, vol. 102, pp. 1141–1148, Aug. 1997.

[41] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” [CD-ROM]. National Bureau of Standards, vol. 93, 1993.

[42] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[43] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[44] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[45] A. Gaich and P. Mowlaee, “On speech quality estimation of phase-aware single-channel speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 216–220.

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

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

[48] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

Naijun Zheng received the B.S. and M.S. degrees in communication engineering from Xidian University, Xi’an, China, in 2015 and 2018, respectively. He is currently working toward the Ph.D. degree in systems engineering and engineering management from The Chinese University of Hong Kong, Shatin, Hong Kong. His research interests include speech enhancement and speaker recognition.

Xiao-Lei Zhang (S’08–M’12) received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Full Professor with the Center for Intelligent Acoustics and Immersive Communications, and the School of Marine Science and Technology, Northwestern Polytechnical University, Xi’an, China. He was a Postdoctoral Researcher with the Perception and Neurodynamics Laboratory, The Ohio State University.

His research interests include audio and speech signal processing, machine learning, statistical signal processing, and artificial intelligence. He has published over 30 journal articles and conference papers in IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE/ACM Transaction on Audio, Speech, and Language Processing, *Neural Networks*, IEEE Transactions on Cybernetics, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, *ICASSP, Interspeech*, etc. and co-edited a text book in statistics. He received the first-class Beijing Science and Technology Award. He serves/served as an associate editor of five international journals including *Neural Networks* and *EURASIP Journal on Audio, Speech, and Music Processing*.