PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network

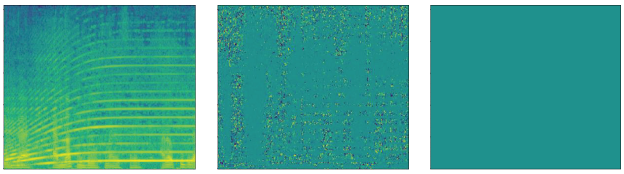
Dacheng Yin,[1]* **Chong Luo,**[2] **Zhiwei Xiong,**[1] **Wenjun Zeng**[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia
ydc@mail.ustc.edu.cn, {cluo, wezeng}@microsoft.com, zwxiong@ustc.edu.cn

Abstract

Time-frequency (T-F) domain masking is a mainstream approach for single-channel speech enhancement. Recently, focuses have been put to phase prediction in addition to amplitude prediction. In this paper, we propose a phase-and-harmonics-aware deep neural network (DNN), named PHASEN, for this task. Unlike previous methods which directly use a complex ideal ratio mask to supervise the DNN learning, we design a two-stream network, where amplitude stream and phase stream are dedicated to amplitude and phase prediction. We discover that the two streams should communicate with each other, and this is crucial to phase prediction. In addition, we propose frequency transformation blocks to catch long-range correlations along the frequency axis. Visualization shows that the learned transformation matrix implicitly captures the harmonic correlation, which has been proven to be helpful for T-F spectrogram reconstruction. With these two innovations, PHASEN acquires the ability to handle detailed phase patterns and to utilize harmonic patterns, getting 1.76dB SDR improvement on AVSpeech + AudioSet dataset. It also achieves significant gains over Google's network on this dataset. On Voice Bank + DEMAND dataset, PHASEN outperforms previous methods by a large margin on four metrics.

1 Introduction

Single-channel speech enhancement aims at separating the clean speech from a noise-corrupted speech signal. Existing methods can be classified into two categories according to the signal domain they work on. The time domain methods directly operate on the one-dimensional (1D) raw waveform of speech signals, while the time-frequency (T-F) domain methods manipulate the two-dimensional (2D) speech spectrogram. Mainstream methods in the second category formulate the speech enhancement problem as to predict a T-F mask over the input spectrogram. Early T-F masking methods only try to recover the amplitude of the target speech. When the importance of phase information was recognized, complex ideal ratio mask (cIRM) (Williamson, Wang, and Wang 2016) was proposed aiming at faithfully recovering the complex T-F spectrogram.



(a) Noisy signal (b) cIRM.im (c) Estimated cRM.im

Figure 1: Straightforward cIRM estimation does not achieve desired results. Although the imaginary part of the cIRM, as shown in (b), contains much information, that of a predicted cRM, as shown in (c), is almost zero.

Williamson et al. (Williamson, Wang, and Wang 2016) observed that, in Cartesian coordinates, structure exists in both real and imaginary components of the cIRM, so they designed deep neural network (DNN)-based methods to estimate the real and imaginary parts of cIRM. However, on large dataset AVSpeech, our evaluations of a modern DNN-based cIRM estimation method (Ephrat et al. 2018) shows that simply changing the training target to cIRM did not generate desired prediction results. Fig.1(a) shows the amplitude of the noisy signal where the stripe pattern is caused by noise. Fig.1(b) and (c) show the imaginary parts of the ideal mask and the estimated mask, respectively. Surprisingly, Fig.1(c) is almost zero, meaning that the estimated cIRM is downgraded to IRM. In another word, the phase information is not recovered at all.

This observation motivates us to design a novel architecture to improve the phase prediction. A straightforward idea is to separately predict amplitude mask and phase with a two-stream network. However, Willianson et al. (Williamson, Wang, and Wang 2016) also pointed out that, in polar coordinates, structure does not exist in the phase spectrogram. This suggests that independent phase estimation is very difficult, if not completely impossible. In view of this, we add two-way information exchange for the two-stream architecture, so that the predicted amplitude can guide the prediction of phase. Results show that such information exchange is critical to the successful phase predic-

*This work was carried out while Dacheng Yin was an intern at MSRA.

tion of the target speech.

In the design of the amplitude stream, we find that conventional CNN kernels which are widely used in image processing do not capture the harmonics in T-F spectrogram. The reason is that correlations in natural images are mostly local while those in speech T-F spectrogram along the frequency axis are mostly non-local. In particular, at a given point of time, the value at a base frequency f_0 is strongly correlated with the values at its overtones. Unfortunately, previous DNN models cannot efficiently exploit harmonics although backbones like U-net (Jansson et al. 2017) and dilated 2D convolution (Ephrat et al. 2018) can increase the receptive field. In this paper, we propose to insert frequency transformation blocks (FTBs) to capture global correlations along the frequency axis. Visualization of FTB weights shows that FTBs implicitly learn the correlations among harmonics.

In a nutshell, we design a phase-and-harmonics-aware speech enhancement network, dubbed PHASEN, for monaural speech enhancement. The contributions of this work are three-fold:

- We propose a novel two-stream DNN architecture with two-way information exchange for efficient speech enhancement in T-F domain. The proposed architecture is capable of recovering phase information of the target speech.
- We design frequency transformation blocks in the amplitude stream to efficiently exploit global frequency correlations, especially the harmonic correlation in spectrogram.
- We carry out comprehensive experiments to justify the design choices and to demonstrate the performance superiority of PHASEN over existing noise reduction methods.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents the proposed PHASEN architecture and its implementation details. Section 4 shows the experimental results. Section 5 concludes this paper with discussions on limitations and future work.

2 Related Work

This section reviews both time-frequency domain methods and time-domain methods for single-channel speech enhancement. Within T-F domain methods, we are only interested in T-F masking methods. Special emphases are put to phase estimation and the utilization of harmonics.

2.1 T-F Domain Masking Methods

T-F domain masking methods for speech enhancement usually operate in three steps. First, the input time-domain waveform is transformed into T-F domain and represented by a T-F spectrogram. Second, a multiplicative mask is predicted based on the input spectrogram and applied to it. Last, an inverse transform is applied to the modified spectrogram to obtain the real-valued time-domain signal. The most widely used T-F spectrogram is computed by the short-time Fourier transform (STFT) and it can be convert back to time-domain signal by the inverse STFT (iSTFT). The key problems to be solved in T-F domain masking methods are what type of mask to be used and how to predict it.

Early T-F masking methods only try to estimate the amplitudes of a spectrogram by using real-valued ideal binary mask (IBM) (Hu and Wang 2001), ideal ratio mask (IRM) (Srinivasan, Roman, and Wang 2006; Narayanan and Wang 2013), or spectral magnitude mask (SMM) (Wang, Narayanan, and Wang 2014). After the enhanced amplitudes are obtained, they are combined with the noisy phase to produce the enhanced speech. Later, research (Paliwal, Wójcicki, and Shannon 2011) reveals that phase plays an important role in speech quality and intelligibility. In order to recover phase, phase sensitive mask (PSM) (Erdogan et al. 2015; Weninger et al. 2015) and cIRM (Williamson, Wang, and Wang 2016) are proposed. PSM is still a real-valued mask, extending SMM by simply adding a phase measure. In contrast, cIRM is a complex-valued mask which has the potential to faithfully recover both amplitude and phase of the clean speech.

Williamson et al. (Williamson, Wang, and Wang 2016) propose a DNN-based approach to estimate the real and imaginary components of the cIRM, so that both amplitude and phase spectra can be simultaneously enhanced. However, their experimental results show that using cIRM does not achieve significantly better results than using PSM. We believe that the potential of a complex mask is not fully exploited. In (Ephrat et al. 2018), a much deeper neural network with dilated convolution and bi-LSTM is employed for speech separation with visual clues. It also achieves state-of-the-art speech enhancement performance when visual clues are absent. We carry out experiments on the network and surprisingly find that the imaginary components of the estimated cIRM is almost zero. This suggests that directly using cIRM to supervise a single-stream DNN cannot achieve satisfactory results.

There exist some other methods (Takahashi et al. 2018; Takamichi et al. 2018; Masuyama et al. 2019) which process phase reconstruction asynchronously with amplitude estimation. Their goal is to reconstruct phase based on a given amplitude spectrogram, which could be the amplitude spectrogram of a clean speech or the output from any speech denoising model. In particular, Takahashi et al. (Takahashi et al. 2018) observe the difficulty in phase regression, so they treat the phase estimation problem as a classification problem by discretizing phase values and assigning class indices to them. While all these methods demonstrate the benefits of phase reconstruction, their approach does not fully utilize the rich information in the input noisy phase spectrogram.

2.2 Time Domain Methods

Time domain methods belong to the other camp for speech enhancement. We briefly mention several pieces of work here because they are proposed to avoid the phase prediction problem in T-F domain methods. SEGAN (Pascual, Bonafonte, and Serra 2017) uses generative adversarial networks (GANs) to directly predict the 1D waveform of the clean speech. Rethage et al. (Rethage, Pons, and Serra 2018) modify Wavenet for the speech enhancement task. conv-TasNet (Luo and Mesgarani 2019) uses a learnable encoder-decoder in time domain as an alternative to the hand-crafted STFT-iSTFT for a speech separation task. However, when it is ap-

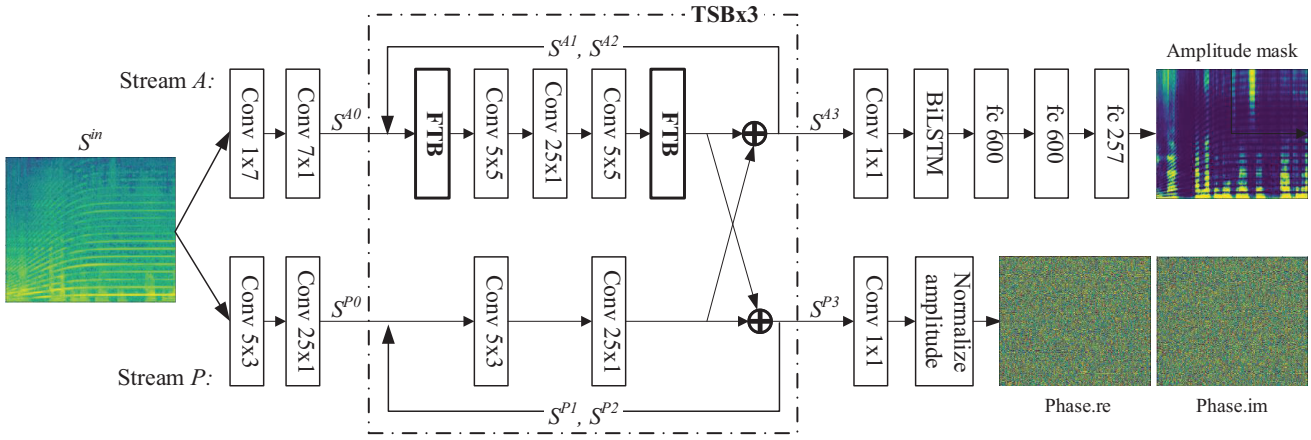


Figure 2: The proposed two-stream PHASEN architecture. The amplitude stream (Stream A) is in the upper portion and the phase stream (Stream P) is in the lower portion. The outputs of Stream A and Stream P are the amplitude mask and the estimated (complex) phase, respectively. Three two-stream blocks (TSBs) are stacked in the network.

plied to the speech enhancement task, the 2ms frame length appears to be too short. TCNN (Pandey and Wang 2019) adopts a similar approach as TasNet, but it uses non-linear encoder-decoder and longer frame length than TasNet. Although these methods divert around the difficult phase estimation problem, they also give up the benefits of speech enhancement in T-F domain, as it is widely recognized that most speech and noise patterns are separately distributed or easily distinguishable on T-F domain features. As a result, the performance of time domain methods is not among the first tier in the speech enhancement task.

2.3 Harmonics in Spectrogram

Plapous et al. (Plapous, Marro, and Scalart 2005) discover that common noise reduction algorithms suppress some harmonics existing in the original signal and then the enhanced signal sounds degraded. They propose to regenerate the distorted speech frequency bands by taking into account the harmonic characteristic of speech. Other research (Krawczyk and Gerkmann 2014; Mowlaee and Kulmer 2015) also show that phase correlation between harmonics can be used for speech phase reconstruction. A recent work (Wakabayashi et al. 2018) further propose a phase reconstruction method based on harmonic enhancement using the fundamental frequency and phase distortion feature. All these work demonstrate the importance of harmonics in speech enhancement. In this paper, we also try to exploit harmonic correlation, but this is achieved by designing an integral block in the end-to-end learning DNN.

3 PHASEN Architecture

3.1 Overview

The basic idea behind PHASEN is to separate the predictions of amplitude and phase, as the two prediction tasks may need different features. In our design, we use two parallel streams, denoted by stream A for amplitude mask prediction and stream P for phase prediction. The entire PHASEN architecture is shown in Fig. 2.

The input to the network is the STFT spectrogram, denoted by S^{in}. Here, $S^{in} \in \mathbb{R}^{T\times F\times 2}$ is a complex-valued spectrogram, where T represents the number of time steps and F represents the number of frequency bands. S^{in} is fed into both streams and two different groups of 2D convolutional layers are used to produce feature $S^{A_0} \in \mathbb{R}^{T\times F\times C_A}$ for stream A and feature $S^{P_0} \in \mathbb{R}^{T\times F\times C_P}$ for stream P. Here, C_A and C_P are the number of channels for stream A and stream P, respectively.

The key component in PHASEN is the stacked two-stream blocks (TSBs), in which stream A and stream P features are computed separately. Note that at the end of each TSB, stream A and stream P exchange information. This design is critical to the phase estimation, as phase itself does not have structure and is hard to estimate (Williamson, Wang, and Wang 2016). However, with the information from the amplitude stream, the features for phase estimation is significantly improved. In Section 4, we will visualize the difference between the estimated phase spectrograms when the information communication is present and absent. The output features of the three TSBs are denoted by S^{A_i} and S^{P_i}, for $i \in \{1, 2, 3\}$. They have the same dimensions as S^{A_0} and S^{P_0}. In stream A, frequency transformation blocks (FTBs) are used to capture non-local correlation along the frequency axis.

After the three TSBs, S^{A_3} and S^{P_3} are used to predict amplitude mask and phase. For S^{A_3}, channel is reduced to $C_r = 8$ by a 1×1 convolution, then reshaped into a 1D feature map, whose dimension is $T \times (F \cdot C_r)$, and finally fed into a Bi-LSTM and three fully connected (FC) layers to predict an amplitude mask $M \in \mathbb{R}^{T\times F\times 1}$. Sigmoid is used as activation function of the last FC layer. For the other FC layers, ReLU is used as activation function.

For S^{P_3}, a 1×1 convolution is used to reduce channel number to 2 to form a complex-valued feature map $S^{P_c} \in$

$\mathbb{R}^{T\times F\times 2}$, where the two channels correspond to the real and the imaginary parts. Then, amplitude of this complex feature map is normalized to 1 for each T-F bin. As such, the feature map only contains phase information. The phase prediction result is denoted by Ψ.

Finally, the predicted spectrogram can be computed by:

$$S^{out} = abs(S^{in}) \circ M \circ \Psi, \tag{1}$$

where $\circ$ denotes element-wise multiplication.

3.2 Two-Stream Blocks (TSBs)

Stream A In each TSB, three 2D convolutional layers are used for stream A to handle local time-frequency correlation of the input feature. To capture global correlation on frequency axis such as harmonic correlation, we propose frequency transformation blocks (FTBs) to be used before and after the three convolutional layers. The FTB design will be detailed in the next subsection. The combination of 2D convolutions and FTBs efficiently captures both global and local correlations, allowing the following blocks to extract high-level features for amplitude prediction. Stream A of each TSB performs the following computation:

$$S_0^{A_i} = FTB_{in}^i(S^{A_i}), \tag{2}$$

$$S_{j+1}^{A_i} = conv_j^{A_i}(S_j^{A_i}), \quad j \in \{0,1,2\}, \tag{3}$$

$$S_4^{A_i} = FTB_{out}^i(S_3^{A_i}). \tag{4}$$

Here, $conv_j^{A_i}$ represents the j-th convolutional layer in stream A of the i-th TSB. $S_{j+1}^{A_i}$ and $S_j^{A_i}$ represent its output and input, respectively. FTB_{in}^i and FTB_{out}^i represent the FTB before and after the three 2D convolutional layers. Each 2D convolutional layer is followed by batch normalization (BN) and activation function ReLU.

Stream P Stream P is designed to be light-weight. We only use two 2D convolutional layers in each TSB to process the input feature $S^{P_i}(i = 1, 2, 3)$. Mathematically,

$$S_0^{P_i} = S^{P_i}, \tag{5}$$

$$S_{j+1}^{P_i} = conv_j^{P_i}(S_j^{P_i}), \quad for \;\; j \in \{0,1\}. \tag{6}$$

Here, $conv_j^{P_i}$ represents the j-th convolutional layer in stream P of the i-th TSB. $S_{j+1}^{P_i}$ and $S_j^{P_i}$ denote its output and input, respectively. The second convolutional layer uses a kernel size of 25×1 to capture long-range time-domain correlation. Global Layer Normalization(gLN) is performed before each convolutional layer. In stream P, no activation function is used. We will later show in ablation studies that this choice increases performance.

Information Communication Information communication is critical to the success of the two-stream structure. Without the information from Stream A, Stream P by itself cannot successfully make phase prediction. Conversely, successfully predicted phases can also help Stream A to better predict amplitude. The communication takes place just before TSB generates output features. Let $S_4^{A_i}$ and $S_2^{P_i}$ be the amplitude features and phase features computed from eq. (4) and eq. (6), the output feature of TSB after information communication can be written as:

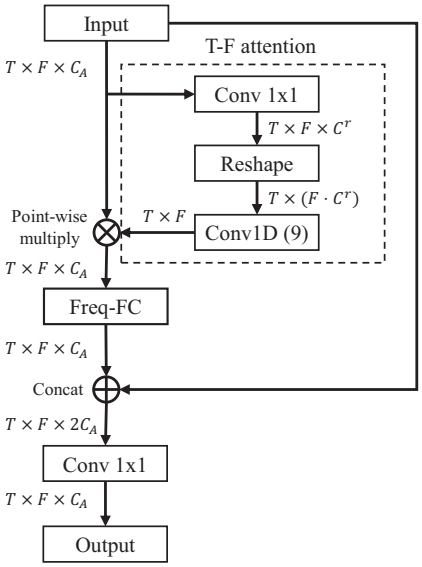


Figure 3: Flowchart of the proposed FTBs. Here, $C^r = 5$, and the kernel size of Conv 1D is 9.

$$S^{A_{i+1}} = f_{P2A}(S_4^{A_i}, S_2^{P_i}), \tag{7}$$

$$S^{P_{i+1}} = f_{A2P}(S_2^{P_i}, S_4^{A_i}), \tag{8}$$

where f_{P2A} and f_{A2P} are information communication functions of the two directions. In this work, we adopt the gating mechanism. For $i \in \{P2A, A2P\}$, we have:

$$f_i(x_1, x_2) = x_1 \circ Tanh(conv(x_2)). \tag{9}$$

Here, $\circ$ denotes element-wise multiplication and $conv$ represents a 1×1 convolution. The number of output channels is the same as the number of channels in x_1.

3.3 Frequency Transformation Blocks (FTBs)

Non-local correlations exist in a T-F spectrogram along the frequency axis. A typical example is the correlations among harmonics, which has been shown to be helpful for the reconstruction of corrupted T-F spectrograms. However, simply stacking several 2D convolution layers with small kernels cannot capture such global correlation. Therefore, we design FTBs to be inserted at the beginning and the end of each TSB, so that the output features of TSB have full-frequency receptive field. At the kernel of an FTB is the learning of a transformation matrix, which is applied on the frequency axis. Fig. 3 shows the flowchart of the proposed FTB. The three groups of operations in each FTB can be represented by:

$$S^a = f_{attn}(S^I), \tag{10}$$

$$S^{tr} = \text{FreqFC}(S^a), \tag{11}$$

$$S^O = conv(concat(S^{tr}, S^I)). \tag{12}$$

Eq. (10) describes the T-F attention module as highlighted in the dotted box in Fig. 3. With the input feature S^I, it uses

2D and 1D convolutional layers to predict an attention map, which is then point-wise multiplied to S^I to obtain S^a. The 2D 1×1 convolution reduces the channel number to $C^r = 5$ and the kernel size of the 1D convolution is 9.

Freq-FC is the key component in FTB. It contains a trainable frequency transformation matrix (FTM) which is applied to the feature map slice at each point in time. Let $X_{tr} \in \mathbb{R}^{F \times F}$ denote the trainable FTM and let $S^a(t_0) \in \mathbb{R}^{F \times C_A}$ ($t_0 \in \{0, 1, ..., T-1\}$) denote the feature slice at time step t_0. The transformation can be simply represented by the following equation:

$$S^{tr}(t_0) = X_{tr} \cdot S^a(t_0). \tag{13}$$

The transformed feature slice at time step t_0, denoted by $S^{tr}(t_0)$, has the same dimension as $S^a(t_0)$. Stacking them along the time axis and we can get the transformed feature map S^{tr}. After Freq-FC, each T-F bin in S^{tr} will contain the information from all the frequency bands of S^a. This allows the following blocks to exploit global frequency correlations for amplitude and phase estimation.

The output of an FTB, denoted by S^O, is calculated by concatenating S^{tr} with S^I and fusing them with a 1×1 convolution. In the proposed FTBs, batch normalization (BN) and ReLU are used after all convolutional layers as normalization method and activation function.

3.4 Implementation

PHASEN is implemented in Pytorch. The dimension of feature maps and the kernel size of convolutional layers are shown in Fig. 2 and Fig. 3. Both streams use convolution operation with zero padding, dilation=1 and stride=1, making sure the input and output feature map size are the same. All the conv layers' output channel in Stream A and P are 96 and 48, except the last 1x1 conv, respectively. The Bi-LSTM unit number is 600. All audios are resampled to 16kHz. STFT is calculated using Hann window, whose window length is 25ms. The hop length is 10ms and FFT size is 512.

The network is trained using MSE loss on the power-law compressed STFT spectrogram. The loss consists of two parts: amplitude loss L_a and phase-aware loss L_p.

$$L = 0.5 \times L_a + 0.5 \times L_p, \tag{14}$$

$$L_a = MSE(abs(S^{out}_{cprs}), abs(S^{gt}_{cprs})), \tag{15}$$

$$L_p = MSE(S^{out}_{cprs}, S^{gt}_{cprs}), \tag{16}$$

where S^{out}_{cprs} and S^{gt}_{cprs} are the power-law compressed spectrogram of output spectrogram S^{out} and ground truth spectrogram S^{gt}. The compression is performed on amplitude with $p = 0.3$ ($A^{0.3}$, where A is the amplitude of the spectrogram.)

Note that instead of only using pure phase, whole spectrogram (phase and amplitude) is taken into consideration for L_p. In this way, phase of T-F bins with higher amplitude is emphasized, helping the network to focus on the high amplitude T-F bins where most speech signals are located.

4 Experiments

4.1 Datasets

Two datasets are used in our experiments.

AVSpeech+AudioSet: This is a large dataset proposed by (Ephrat et al. 2018). Clean speech dataset AVSpeech is collected from YouTube, containing 4700 hours of video segments with approximately 150,000 distinct speakers, spanning a wide variety of people and languages. Noise dataset AudioSet (Gemmeke et al. 2017) contains a total of more than 1.7 million 10-second segments of 526 kinds of noise. 3-second segments $Speech_j$ and $Noise_k$ are firstly randomly sampled from clean speech and noise dataset, then the noisy speech Mix_i is calculated by $Mix_i = Speech_j + 0.3 \times Noise_k$. Mix_i and $Speech_j$ form a noisy-clean speech pair for training and testing. Because of the wide energy distribution in both datasets, the created noisy speech dataset has a wide range of SNR. In our experiments, 100k segments randomly sampled from AVSpeech dataset and the "Balanced Train" part of AudioSet are used to synthesize the training set, while the validation set is the same as the one used in (Ephrat et al. 2018), synthesized by the test part of AVSpeech dataset and the evaluation part of AudioSet.

Voice Bank+DEMAND: This is an open dataset[1] proposed by (Valentini-Botinhao et al. 2016). Speech of 30 speakers from the Voice Bank corpus (Ephrat et al. 2018) are selected as clean speech: 28 are included in the training set and 2 are in the validation set. The noisy speech is synthesized using a mixture of clean speech with noise from Diverse Environments Multichannel Acoustic Noise Database (DEMAND) (Thiemann, Ito, and Vincent 2013). A total of 40 different noise conditions are considered in training set and 20 different conditions are considered in test set. Finally, the training and test set contain 11572 and 824 noisy-clean speech pairs, respectively. Both speakers and noise conditions in the test set are totally unseen by the training set. Our system comparison is partly done on this dataset.

4.2 Evaluation Metrics

The following six metrics are used to evaluate PHASEN and state-of-the-art competitors. All metrics are better if higher.

- SDR (Vincent, Gribonval, and Févotte 2006): Signal-to-distortion ratio from the mir_eval library;
- PESQ: Perceptual evaluation of speech quality (from -0.5 to 4.5).
- STOI: Short-time objective intelligibility measure (from 0 to 1)
- CSIG (Hu and Loizou 2007): Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal (from 1 to 5).
- CBAK (Hu and Loizou 2007): MOS prediction of the intrusiveness of background noise (from 1 to 5).
- COVL (Hu and Loizou 2007): MOS prediction of the overall effect (from 1 to 5).
- SSNR: Segmental SNR .

4.3 Ablation Study

In the ablation study, networks of different settings are trained with the same random seed for 1 million steps. Adam

[1] https://datashare.is.ed.ac.uk/handle/10283/1942

Table 1: Ablation study on AVSpeech + AudioSet

Method	SDR(dB)	PESQ	STOI
PHASEN-baseline	15.08	2.87	0.844
PHASEN-1strm	15.99	2.98	0.856
PHASEN-w/o-FTBs	16.10	3.31	0.874
PHASEN-w/o-A2PP2A	16.13	3.33	0.876
PHASEN-w/o-P2A	16.62	3.38	0.880
PHASEN	**16.84**	**3.40**	**0.884**

optimizer with a fixed learning rate of 0.0002 is used and the batch size is set to 8. We use mean SDR, PESQ and STOI on test dataset as the evaluation metric.

The ablation results are shown in Table 1. Among these methods, PHASEN represents our full model. PHASEN-baseline represents a single-stream network which uses cIRM as training target. We use the network structure in stream A for PHASEN-baseline and replace the FTBs with 5×5 convolutions. The comparison between PHASEN and PHASEN-baseline shows that our two innovations, namely two-stream architecture and FTBs, provide a total improvement of 1.76dB on SDR, 0.53 on PESQ, and 0.04 on STOI.

Two-Stream Architecture PHASEN-1strm shows the performance of single-stream architecture with cIRM as training target. In this experiment, stream P and information communication are removed from PHASEN architecture, while FTBs are preserved. The output of stream A is the predicted cRM. Comparison between PHASEN-1strm and PHASEN shows that the two-stream architecture provides gain of 0.85dB on SDR, 0.42 on PESQ and 0.028 on STOI. The large gain on PESQ and STOI indicates the proposed two-stream architecture can largely improve the perceptual quality and intelligibility of the denoised speech.

FTBs The proposed method uses FTBs at both the beginning and the end of each TSB. In ablation study, PHASEN-w/o-FTBs try to replace all the FTBs in PHASEN architecture with 5×5 convolutions. By comparing PHASEN to PHASEN-w/o-FTBs we find that FTBs can provide 0.74 dB, 0.09 and 0.01 gain on SDR, PESQ and STOI, respectively. We have also tried to replace the FTBs on either location of each TSB with 5×5 convolutions. Both attempts result into 0.31dB-0.39dB drop on SDR, 0.03-0.05 drop on PESQ and 0.005 drop on STOI, showing that FTBs on both locations are equally important and the gain is accumulative.

In order for a better understanding of FTBs, we visualize the weights of X_{tr}, the matrix that reflects the learned global frequency correlation. From Fig. 4, we show that the energy map of X_{tr} resembles the harmonic correlation, especially when higher harmonics (larger H) are taken into consideration. This phenomenon confirms that FTBs really capture the harmonic correlation, and that harmonic correlation is really useful to a speech enhancement network, because the network can learn this correlation implicitly.

Information communication mechanism PHASEN-w/o-P2A, and PHASEN-w/o-A2PP2A are two settings that remove the information communication mechanism partly and fully. The former one removes the communication from stream P to stream A, and the latter one removes communication of both directions. In SDR and PESQ result, significant gain of 0.49dB and 0.05 is observed when comparing PHASEN-w/o-P2A to PHASEN-w/o-A2PP2A. This indicates that the information in the intermediate steps of amplitude prediction is very helpful to phase prediction. In comparison between our full model PHASEN and PHASEN-w/o-P2A, we also see that when integrating stream P information into stream A, the model gets 0.22dB gain on SDR and 0.02 gain on PESQ. This proves that phase feature can also help amplitude prediction.

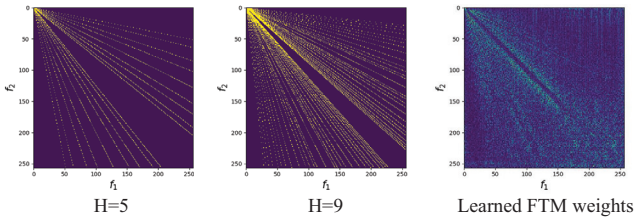


Figure 4: Comparison of different level of harmonic correlation: $f_2 = \frac{m}{n} f_1, m \neq n, m, n \in \{0, 1, ..., H\}$ and learned FTM weights. $f_1 = f_2 = 0$ is on the upper-left corner of each sub-figure.

Fig. 5 also confirms the above improvements through visualization. Here, because the predicted phase spectrogram has few visible patterns, we visualize $\Delta\Psi = \Psi/\Psi_{in}$, which represents the phase difference between predicted phase spectrogram and input noisy spectrogram. The division operation in this formula is on complex domain, and Ψ_{in} represents the phase spectrogram of input noisy speech. From the visualization, we can conclude that information communication mechanism not only significantly improves the phase prediction, but helps remove amplitude artifacts. To summarize, information communication of both directions are useful in PHASEN, while direction "A2P" plays a key role.

Other ablations Apart from the results shown in Table 1, we also perform ablations on activation function and normalization functions for stream P.

The proposed method uses no activation function on stream P. Though this design is counter-intuitive, it is actually inspired by previous work (Luo and Mesgarani 2019) and also supported by the ablation study. In fact, we try to add ReLU or Tanh as activation function after each, except the last, convolutional layer in stream P. However, this causes 0.02dB-0.16dB drop on SDR. Moreover, if ReLU is added after the last convolutional layer in stream P, a huge drop of 5.52dB and 0.2 is observed on SDR and PESQ.

The proposed method uses gLN in stream P and BN in stream A. We test other normalization method for each stream. A performance drop of 0.97dB and 0.12 on SDR and PESQ is observed if gLN is used in stream A, while a drop of 0.09dB and 0.02 on SDR and PESQ is observed if BN is used in stream P.

From these two experiments, we can observe significant

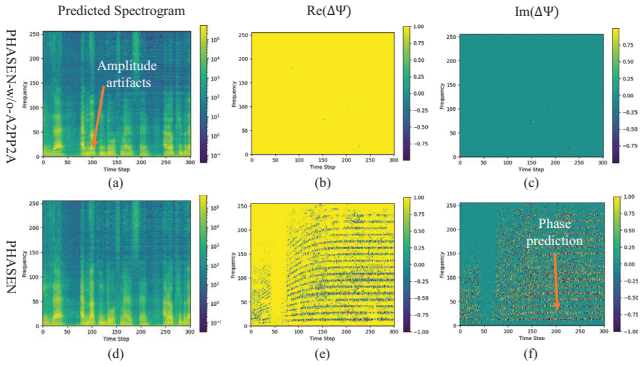


Figure 5: The effect of information communication mechanism. We use the same input noisy speech as in Fig.2. (a), (d): Amplitude of predicted spectrogram; (b), (e): real part of $\Delta\Psi$; (c), (f): imaginary part of $\Delta\Psi$. (a)-(c) are obtained without A2PP2A. Significant amplitude artifacts are observed in (a) on frequency bands where speech is overwhelmed by noise. In every T-F bins, (c) is almost zero, and (b) is almost one, indicating failure on phase prediction. In contrast, when A2PP2A is used, phase prediction is obviously visible in T-F bins where noise overwhelms speech, as (e) and (f) shows. Best viewed in color.

difference between phase prediction and amplitude mask prediction. This supports our design of using two streams to accomplish the two prediction tasks.

Additionally, we have tried to use complex convolution operations to replace the conv blocks in our architecture, but this did not make much difference on performance.

4.4 System Comparison

We carry out system comparison on both datasets mentioned in section 4.1.

AVSpeech + AudioSet On this large dataset we compare our method with two other recent methods, Conv-TasNet (Luo and Mesgarani 2019) and "Google" (Ephrat et al. 2018). Conv-TasNet is a time domain method. The result of Conv-TasNet is produced using the released code[2], trained for the same epochs and on the same data as our PHASEN. "Google" is a T-F domain masking method which uses cIRM as supervision. The method is intended for both speech enhancement and speech separation. We compare PHASEN with their audio-only, 1S+noise setting. The result in Table 2 shows that our method outperforms both Conv-TasNet and "Google". Note that this is achieved under the condition that we only use a small fraction of training step (1M/5M) and data (100k/2.4M) used by "Google". Such superior performance on large dataset demonstrates that our method can be generalized to various speakers and various kinds of noisy environments. It suggests that PHASEN is readily applicable to complicated real-world environment.

Voice Bank + DEMAND We also train our model on small but commonly-used dataset Voice Bank + DEMAND, so that we can fairly compare our PHASEN with many other methods. In this experiment, our network is trained on the training set for 40 epochs, with Adam optimizer using warm-up step number of 6000, learning rate of 0.0005, and batch size of 12.

[2]https://github.com/funcwj/conv-tasnet

Table 2: System comparison on AVSpeech + AudioSet

Method	SDR(dB)	PESQ	STOI
Conv-TasNet	14.19	2.93	0.833
Google(5M step, 2.4M data)	16.00	–	–
PHASEN(1M step, 100k data)	**16.84**	**3.40**	**0.884**

Table 3: System comparison on Voice Bank + DEMAND

Method	SSNR	PESQ	CSIG	CBAK	COVL
Noisy	1.68	1.97	3.35	2.44	2.63
SEGAN	7.73	2.16	3.48	2.94	2.80
Wavenet	–	–	3.62	3.23	2.98
DFL	–	–	3.86	3.33	3.22
MMSE-GAN	–	2.53	3.80	3.12	3.14
MDPhD	**10.22**	2.70	3.85	3.39	3.27
PHASEN	10.18	**2.99**	**4.21**	**3.55**	**3.62**

Table 3 shows the comparison result. Firstly, our method has very large gain over time-domain methods like SEGAN (Pascual, Bonafonte, and Serra 2017), Wavenet (Rethage, Pons, and Serra 2018), and DFL (Germain, Chen, and Koltun 2018) on all the five metrics, even though these time-domain methods are free of phase-prediction problem. This proves the advantage of our method over the time-domain methods on capturing phase-related information. Also, our method shows great improvement over time-frequency domain method like MMSE-GAN (Soni, Shah, and Patil 2018) on all metrics, indicating the superiority of our network design. Finally, we also compare our method with a recent hybrid model of time-domain and time-frequency domain called MDPhD (Kim et al. 2018). Our method significantly outperforms it on four metrics, and there is only a small difference of about 0.04dB on SSNR metric.

5 Conclusion

We have proposed a two-stream architecture with two-way information communication for efficient phase prediction in monaural speech enhancement. We have also designed a learnable frequency transformation matrix in the network. It implicitly learns a pattern that is consistent with harmonic correlation. Comprehensive ablation studies have been carried out, justifying almost every design choices we have made in PHASEN. Comparison with state-of-the-art systems on both AVSpeech+AudioSet and Voice Bank+DEMAND datasets demonstrates the superior performance of PHASEN. Note that the current design of PHASEN does not allow it to be used for low-latency applications, such as voice over IP. In the future, we plan to explore the potential of PHASEN in low-latency settings and mobile settings which require a smaller model size and shorter inference time. We also plan to expand this architecture to other related tasks such as speech separation.

References

Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W. T.; and Rubinstein, M. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Erdogan, H.; Hershey, J. R.; Watanabe, S.; and Le Roux, J. 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *ICASSP 2015*, 708–712. IEEE.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP 2017*, 776–780. IEEE.

Germain, F. G.; Chen, Q.; and Koltun, V. 2018. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*.

Hu, Y., and Loizou, P. C. 2007. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing* 16(1):229–238.

Hu, G., and Wang, D. 2001. Speech segregation based on pitch tracking and amplitude modulation. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, 79–82. IEEE.

Jansson, A.; Humphrey, E.; Montecchio, N.; Bittner, R.; Kumar, A.; and Weyde, T. 2017. Singing voice separation with deep u-net convolutional networks. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* 323–332.

Kim, J.-H.; Yoo, J.; Chun, S.; Kim, A.; and Ha, J.-W. 2018. Multi-domain processing via hybrid denoising networks for speech enhancement. *arXiv preprint arXiv:1812.08914*.

Krawczyk, M., and Gerkmann, T. 2014. Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(12):1931–1940.

Luo, Y., and Mesgarani, N. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(8):1256–1266.

Masuyama, Y.; Yatabe, K.; Koizumi, Y.; Oikawa, Y.; and Harada, N. 2019. Deep griffin–lim iteration. In *ICASSP 2019*, 61–65. IEEE.

Mowlaee, P., and Kulmer, J. 2015. Harmonic phase estimation in single-channel speech enhancement using phase decomposition and snr information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(9):1521–1532.

Narayanan, A., and Wang, D. 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7092–7096. IEEE.

Paliwal, K.; Wójcicki, K.; and Shannon, B. 2011. The importance of phase in speech enhancement. *speech communication* 53(4):465–494.

Pandey, A., and Wang, D. 2019. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019*, 6875–6879. IEEE.

Pascual, S.; Bonafonte, A.; and Serra, J. 2017. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

Plapous, C.; Marro, C.; and Scalart, P. 2005. Speech enhancement using harmonic regeneration. In *Proceedings.(ICASSP'05).*, volume 1, I–157. IEEE.

Rethage, D.; Pons, J.; and Serra, X. 2018. A wavenet for speech denoising. In *ICASSP 2018*, 5069–5073. IEEE.

Soni, M. H.; Shah, N.; and Patil, H. A. 2018. Time-frequency masking-based speech enhancement using generative adversarial network. In *ICASSP 2018*, 5039–5043. IEEE.

Srinivasan, S.; Roman, N.; and Wang, D. 2006. Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication* 48(11):1486–1501.

Takahashi, N.; Agrawal, P.; Goswami, N.; and Mitsufuji, Y. 2018. Phasenet: Discretized phase modeling with deep neural networks for audio source separation. In *Interspeech*, 2713–2717.

Takamichi, S.; Saito, Y.; Takamune, N.; Kitamura, D.; and Saruwatari, H. 2018. Phase reconstruction from amplitude spectrograms based on von-mises-distribution deep neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 286–290. IEEE.

Thiemann, J.; Ito, N.; and Vincent, E. 2013. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America* 133(5):3591–3591.

Valentini-Botinhao, C.; Wang, X.; Takaki, S.; and Yamagishi, J. 2016. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, 146–152.

Vincent, E.; Gribonval, R.; and Févotte, C. 2006. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing* 14(4):1462–1469.

Wakabayashi, Y.; Fukumori, T.; Nakayama, M.; Nishiura, T.; and Yamashita, Y. 2018. Single-channel speech enhancement with phase reconstruction based on phase distortion averaging. *TASLP* 26(9):1559–1569.

Wang, Y.; Narayanan, A.; and Wang, D. 2014. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 22(12):1849–1858.

Weninger, F.; Erdogan, H.; Watanabe, S.; Vincent, E.; Le Roux, J.; Hershey, J. R.; and Schuller, B. 2015. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International Conference on Latent Variable Analysis and Signal Separation*, 91–99. Springer.

Williamson, D. S.; Wang, Y.; and Wang, D. 2016. Complex ratio masking for monaural speech separation. *TASLP* 24(3):483–492.