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

Dacheng Yin1, Chong Luo2, Zhiwei Xiong1, and Wenjun Zeng2

1University of Science and Technology of China

2Microsoft Research Asia

ydc@mail.ustc.edu.cn, cluo@microsoft.com, zwxiong@ustc.edu.cn, wezeng@microsoft.com

Copyright 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Contents

[Abstract 1](#_Toc57384671)

[1.Introduction 2](#_Toc57384672)

[2. Related Work 4](#_Toc57384673)

[2.1 T-F Domain Masking Methods 4](#_Toc57384674)

[2.2 Time Domain Methods 5](#_Toc57384675)

[2.3 Harmonics in Spectrogram 5](#_Toc57384676)

[3 PHASEN Architecture 5](#_Toc57384677)

[3.1 Overview 5](#_Toc57384678)

[3.2 Two-Stream Blocks (TSBs) 7](#_Toc57384679)

[3.3 Frequency Transformation Blocks (FTBs) 9](#_Toc57384680)

[3.4 Implementation 9](#_Toc57384681)

[4 Experiments 10](#_Toc57384682)

[4.1 Datasets 10](#_Toc57384683)

[4.2 Evaluation Metrics 11](#_Toc57384684)

[4.3 Ablation Study 11](#_Toc57384685)

[4.4 System Comparison 14](#_Toc57384686)

[5 Conclusion 15](#_Toc57384687)

[References 16](#_Toc57384688)

# 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.

单通道语音增强主流方法是时频域掩蔽, 在振幅预测之外, 人们开始关注相位预测

In this paper, we **propose** a phase-and-harmonics-aware deep neural network (DNN), named PHASEN, for this task.

感知相位+谐波 的 DNN, PHA = Phase Harmonics Aware

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.

双流网络: 两路数据流, 分别用于相位预测和振幅预测

We discover that the two streams should communicate with each other, and this is crucial to phase prediction.

两路数据必须互相交流, 这对相位预测很重要

In addition, we **propose** frequency transformation blocks to catch long-range correlations along the frequency axis.

频域变换模块, 用于捕获频轴上的长程相关(长跨度上的相关性)

Visualization shows that the learned transformation matrix spontaneously captures the harmonic correlation, which has been proven to be helpful for T-F spectrogram reconstruction.

学习到的变换矩阵能捕获谐波相关性，这有助于T-F谱重建。

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.

能处理细致的相位pattern,能利用谐波pattern,

数据集: AVSpeech + AudioSet, Voice Bank + DEMAND

# 1.Introduction

Single-channel speech noise reduction 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.

单通道语音降噪的两类方法:

时域方法直接处理一维（1D）波形，时频（T-F）域方法处理二维（2D）语谱图。

Mainstream methods in the second category formulate the speech noise reduction problem as to predict a T-F mask over the input spectrogram.

第二类中的主流方法, 将语音降噪问题表述为预测输入语谱图上的T-F掩蔽

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.

早期的T-F掩蔽方法仅尝试恢复目标语音的幅度, 在意识到频域信息的重要性后, 出现了复数域的理想浮值掩蔽（cIRM）, 用于恢复复数域的T-F谱

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, in our evaluations of a modern DNN-based cIRM estimation method (Ephrat et al. 2018), we find that simply changing the training target to cIRM did not generate desired prediction results.

仅简单地将训练目标更改为cIRM并不会产生所需的预测结果

Fig.1(a) shows the amplitude of the noisy signal where the stripe pattern is caused by noise.

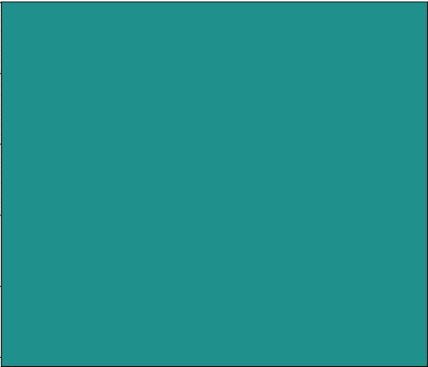
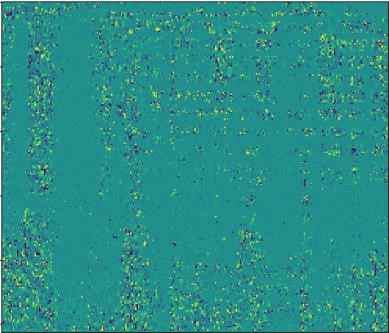
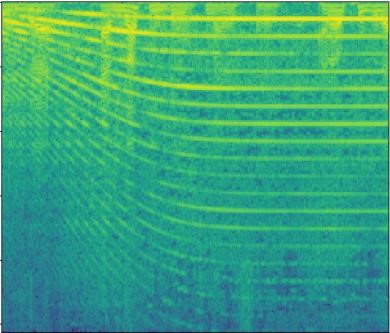
图1(a)显示的带噪信号中的条纹就是噪声

Fig.1(b) and (c) show the imaginary parts of the ideal mask and the estimated mask, respectively.

图1(b)和(c)分别显示了 理想掩蔽和估计掩蔽的虚部

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.

估计结果几乎为0, 意味着cIRM退化成了IRM, 也就是说相位信息完全无法恢复



(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.

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.

基本思路是使用双流网络分开预测幅度掩蔽和相位

However, Willianson et al. (Williamson, Wang, and Wang 2016) also pointed out that, in polar coordinates, structure does not exist in the phase spectrogram.

但是, 在极坐标中,相谱图不存在结构

This suggests that independent phase estimation is very difficult, if not completely impossible.

单独估计相位非常困难

In view of this, we add two-way information exchange for the twostream architecture, so that the predicted amplitude can guide the prediction of phase. Results show that such information exchange is critical to the successful phase prediction of the target speech.

对策:为双向流架构增加了双向信息交换, 以便幅度的预测可以指导相位的预测

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 f0 is strongly correlated with the values at its overtones.

常规CNN核无法捕获T-F谱图中的谐波,原因是T-F谱图在频轴上的非局部特性, 特别是基音和泛音的强相关性

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 spontaneously learn the correlations among harmonics.

本文提出 FTB(频率转换模块)来捕获频轴上的全局相关性

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

* We propose a novel two-stream DNN architecture with two-way information exchange for efficient speech noise reduction 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.

本文结构,

第二节:相关工作

第三节:PHASEN结构和实现细节

第四节:实验结果

第五节:总结和将来工作

# 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.

本节介绍了用于单通道语音增强的时频域方法和时域方法。 在时频域方法中，我们仅对时频掩蔽方法感兴趣。 特殊的重点放在相位估计和谐波利用上。

## 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 shorttime 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.

用于语音增强的T-F域屏蔽方法通常以三个步骤运行。 首先，将输入时域波形转换为T-F域，并由T-F频谱图表示。 其次，基于输入频谱图预测乘法掩码并将其应用。 最后，将逆变换应用于修改后的频谱图，以获得实值时域信号。 使用最广泛的T-F频谱图是通过短时傅立叶变换（STFT）计算的，并且可以通过逆STFT（iSTFT）将其转换回时域信号。 T-F域屏蔽方法要解决的关键问题是使用哪种类型的屏蔽以及如何对其进行预测。

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), 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, Wojcicki, 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) 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.

早期的TF屏蔽方法仅尝试通过使用实值理想二进制掩码（IBM）（Hu和Wang 2001），理想比率掩码（IRM）（Srinivasan，Roman和Wang 2006）或频谱幅度来估计频谱图的幅度。 遮罩（SMM）（Wang，Narayanan和Wang 2014）。 在获得增强的幅度之后，它们与噪声相位组合以产生增强的语音。 后来，研究（Paliwal，Wojcicki和Shannon 2011）发现，相位在语音质量和清晰度方面起着重要作用。 为了恢复相位，提出了相敏掩模（PSM）（Erdogan等人2015）和cIRM（Williamson，Wang和Wang 2016）。 PSM仍然是实值掩码，通过简单地添加相位测量来扩展SMM。 相反，cIRM是一个复数值掩码，它有可能忠实地恢复干净语音的幅度和相位。

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-ofthe-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.

威廉姆森等。 （Williamson，Wang，and Wang 2016）提出了一种基于DNN的方法来估计cIRM的实部和虚部，从而可以同时增强幅度和相位频谱。 但是，他们的实验结果表明，使用cIRM不会比使用PSM取得明显更好的结果。 我们认为，复杂面具的潜力尚未得到充分利用。 在（Ephrat et al.2018）中，采用了具有更深层的，具有扩张卷积和bi-LSTM的神经网络，用于通过视觉线索进行语音分离。 当没有视觉提示时，它还可以实现最先进的语音增强性能。 我们在网络上进行了实验，令人惊讶地发现，估计的cIRM的虚部几乎为零。 这表明直接使用cIRM来监督单流DNN不能取得令人满意的结果。

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.

还有其他一些方法（Takahashi等人2018; Takamichi等人2018; Masuyama等人2019），这些方法与幅度估计异步地处理相位重建。 他们的目标是基于给定的幅度谱图重建相位，该谱图可以是干净语音的幅度谱图，也可以是任何语音降噪模型的输出。 特别是，高桥等。 （Takahashi et al.2018）观察到相位回归的困难，因此他们通过离散化相位值并为其分配类别索引，将相位估计问题视为分类问题。 尽管所有这些方法都证明了相位重建的好处，但它们的方法并未充分利用输入噪声相位频谱图中的丰富信息。

## 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. convolutionTasNet (Luo and Mesgarani 2019) uses a learnable encoderdecoder in time domain as an alternative to the hand-crafted STFT-iSTFT for a speech separation task. However, when it is applied 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 nonlinear 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.

时域方法属于语音增强的另一阵营。我们在此简要提及几项工作，因为提出它们是为了避免T-F域方法中的相位预测问题。 SEGAN（Pascual，Bonafonte和Serra 2017）使用生成对抗网络（GANs）直接预测干净语音的一维波形。 Rethage等。 （Rethage，Pons和Serra 2018）修改Wavenet进行语音增强任务。 convolutionTasNet（Luo and Mesgarani 2019）在时域使用可学习的编码器/解码器，以替代手工制作的STFT-iSTFT进行语音分离任务。但是，将其应用于语音增强任务时，2ms的帧长度似乎太短。 TCNN（Pandey and Wang 2019）采用了与TasNet类似的方法，但是它使用了非线性编解码器，并且帧长比TasNet长。尽管这些方法绕开了困难的相位估计问题，但它们也放弃了T-F域中语音增强的好处，因为众所周知，大多数语音和噪声模式在T-F域特征上是独立分布或容易区分的。结果，时域的表现达不到语音增强任务的第一梯队

## 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.

Plapous等。 （Plapous，Marro和Scalart 2005）发现，常见的降噪算法会抑制原始信号中存在的某些谐波，然后增强的信号声音会降低。 他们提出通过考虑语音的谐波特性来再生失真的语音频带。 其他研究（Krawczyk和Gerkmann，2014； Mowlaee和Kulmer，2015）也表明，谐波之间的相位相关可用于语音相位重建。 最近的工作（Wakabayashi等人2018）进一步提出了一种基于谐波的相位重建方法，该谐波使用基频和相位失真特征进行增强。 所有这些工作证明了谐波在语音增强中的重要性。 在本文中，我们还尝试利用谐波相关，但这是通过在端到端学习DNN中设计一个积分模块来实现的。

# 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.

基本思想是将幅度和相位的预测分开，因为这两个预测任务需要不同的特征。使用两个并行流，用流A预测幅度掩码, 用流P预测相位。整个PHASEN架构如图2所示。

|  |
| --- |
| 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. |

The input to the network is the STFT spectrogram, denoted by *Sin*. Here, *Sin* ∈ R*T*×*F*×2 is a complex-valued spectrogram, where *T* represents the number of time steps and *F* represents the number of frequency bands. *Sin* is fed into both streams and two different groups of 2D convolutional layers are used to produce feature *SA*0 ∈ R*T*×*F*×*CA* for stream *A* and feature *SP*0 ∈ R*T*×*F*×*CP* for stream *P*. Here, *CA* and *CP* are the number of channels for stream *A* and stream *P*, respectively.

网络的输入是STFT频谱图，用Sin表示。在此，Sin∈RT×F×2是复值频谱图，其中T表示时间步数，F表示频带数。将Sin馈入两个流中，并使用两组不同的2D卷积层为流A生成特征SA0∈RT×F×CA，为流P生成特征SP0∈RT×F×CP。在这里，CA和CP是数字分别用于流A和流P的通道数。

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 *SAi* and *SPi*, for *i* ∈ {1*,*2*,*3}. They have the same dimensions as *SA*0 and *SP*0. In stream *A*, frequency transformation blocks (FTBs) are used to capture non-local correlation along the frequency axis.

PHASEN中的关键组件是堆叠的双流块（TSBs），其中流A和流P的特征分开计算。注意，在每个TSB的末尾，流A和流P交换信息。这种设计对于相位估计至关重要，因为相位本身没有结构且难以估计（Williamson，Wang和Wang 2016）。但是，利用幅度流中的信息，相位估计的功能得到了显着改善。在第4节中，我们将可视化在存在和不存在信息沟通时估计的相位谱图之间的差异。对于i∈{1,2,3}，三个TSB的输出特征由SAi和SPi表示。它们的尺寸与SA0和SP0相同。在流A中，频率变换块（FTB）用于捕获沿频率轴的非局部相关性。

After the three TSBs, *SA*3 and *SP*3 are used to predict amplitude mask and phase. For *SA*3, channel is reduced to *Cr* = 8 by a 1 × 1 convolution, then reshaped into a 1D feature map, whose dimension is *T* × (*F* · *Cr*), and finally fed into a Bi-LSTM and three fully connected (FC) layers to predict an amplitude mask *M* ∈ R*T*×*F*×1. Sigmoid is used as activation function of the last FC layer. For the other FC layers, ReLU is used as activation function.

在三个TSB之后，使用SA3和SP3来预测幅度掩码和相位。对于SA3，通过1×1卷积将通道减小为Cr = 8，然后重塑为尺寸为T×（F·Cr）的一维特征图，最后馈入Bi-LSTM和三个完全连接（FC ）层以预测振幅掩模M∈RT×F×1。 Sigmoid用作最后一个FC层的激活功能。对于其他FC层，将ReLU用作激活功能。

For *SP*3, a 1 × 1 convolution is used to reduce channel number to 2 to form a complex-valued feature map *SPc* ∈ R*T*×*F*×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.

对于SP3，使用1×1卷积将通道数减少为2，以形成复值特征图SPc∈RT×F×2，其中两个通道分别对应于实部和虚部。然后，对于每个T-F bin，将此复杂特征图的幅度归一化为1。这样，特征图仅包含相位信息。相位预测结果用Ψ表示。

Finally, the predicted spectrogram can be computed by:

最后，可以通过以下方式计算预测的频谱图：

*Sout* = *abs*(*Sin*) ◦ *M* ◦ Ψ*,* (1)

where ◦ denotes element-wise multiplication.

其中◦表示逐元素乘法。

## 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 highlevel features for amplitude prediction. Stream *A* of each TSB performs the following computation:

在每个TSB中，用三个2D卷积层处理输入特征的局部时频相关性。在三个卷积层之前和之后使用的频率变换块（FTB）来捕获频率轴上的全局相关性(如谐波相关性)。2D卷积和FTB的组合有效地捕获了全局和局部相关性，从而允许以下块提取用于幅度预测的高级特征。每个TSB的流A执行以下计算：

*S*0*Ai* = *FTBini* (*SAi*)*,* (2)

*SjAi*+1= *convjAi*(*SjAi*)*, j* ∈ {0*,*1*,*2}*,* (3)

*S*4*Ai* = *FTBouti* (*S*3*Ai*)*.* (4)

Here,  represents the *j*-th convolutional layer in stream *A* of the *i*-th TSB. and *SjAi* represent its output and input, respectively. *FTBini* and *FTBouti* 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 *SPi*(*i* = 1*,*2*,*3).

**流A** 表示第i个TSB的流A中的第j个卷积层。 和*SjAi*分别表示其输出和输入。 FTBini和FTBouti代表三个2D卷积层之前和之后的FTB。 每个2D卷积层之后是批处理归一化（BN）和激活函数ReLU。

流P的设计较为轻巧。 我们在每个TSB中仅使用两个2D卷积层来处理输入特征SPi（i = 1,2,3）。

Mathematically,

 (5)

*.* (6)

Here,  represents the *j*-th convolutional layer in stream *P* of the *i*-th TSB. and 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.

**流P** 表示第i个TSB的流P中的第j个卷积层。 和分别表示其输出和输入。 第二个卷积层使用25×1的内核大小来捕获远程时域相关性。 全局层归一化（gLN）在每个卷积层之前执行。 在流P中，不使用激活功能。 稍后我们将在消融研究中表明，这种选择可以提高性能。

**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  and 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:

**信息沟通** 信息沟通对于两流结构的成功至关重要。 没有来自流A的信息，流P本身无法成功进行相位预测。 相反，成功预测的相位也可以帮助流A更好地预测振幅。 通信恰在TSB生成输出功能之前进行。 令和为从等式（4）和等式 （6）计算出的幅度特征和相位特征。该信息沟通后TSB的输出特征可写为：

|  |  |
| --- | --- |
| *,* | (7) |
| *,* | (8) |

|  |
| --- |
| T  -  F attention  Concat  Conv  1  1  x  𝑇  ×  𝐹  ×  𝐶  𝐴  𝑇×𝐹×𝐶𝐴  𝑇  ×  𝐹  ×  2  𝐶  𝐴  𝑇×𝐹×𝐶𝐴  Reshape  Conv1D (9)  Conv 1x1  Freq  -  FC  Input  Output  𝑇  ×  𝐹  ×  𝐶  𝑟  𝑇  ×  (  𝐹  ⋅  𝐶  𝑟  )  𝑇×𝐹×𝐶A  Point  -  wise  multiply |
| Figure 3: Flowchart of the proposed FTBs. Here, *Cr* = 5, and the kernel size of Conv 1D is 9. |

where *fP*2*A* and *fA*2*P* are information communication functions of the two directions. In this work, we adopt the attention mechanism. For *i* ∈ {*P*2*A,A*2*P*}, we have:

其中fP2A和fA2P是两个方向的信息沟通功能。 在这项工作中，我们采用注意机制。 对于i∈{P2A，A2P}，我们有：

*fi*(*x*1*,x*2) = *x*1 ◦ *Tanh*(*conv*(*x*2))*.* (9)

Here, ◦ 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.

此处，◦表示逐元素乘法，conv表示1×1卷积。 输出通道数与x1中的通道数相同。

## 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 fullfrequency 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:

沿频率轴的T-F频谱图中存在非局部相关性。 一个典型的例子是谐波之间的相关性，这已被证明有助于重建损坏的T-F频谱图。 但是，仅将几个2D卷积层与小内核堆叠在一起就无法捕获这种全局相关性。 因此，我们将FTB设计为插入每个TSB的开头和结尾，以使TSB的输出特征具有全频接收场。 FTB的核心是学习变换矩阵，该变换矩阵应用于频率轴。 图3示出了所提出的FTB的流程图。 每个FTB中的三组操作可以表示为：

*Sa* = *fattn*(*SI*)*,* (10)

*Str* = FreqFC(*Sa*)*,* (11)

*SO* = *conv*(*concat*(*Str,SI*))*.* (12)

Eq. (10) describes the T-F attention module as highlighted in the dotted box in Fig. 3. With the input feature *SI*, it uses 2D and 1D convolutional layers to predict an attention map, which is then point-wise multiplied to *SI* to obtain *Sa*. The 2D 1×1 convolution reduces the channel number to *Cr* = 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 *Xtr* ∈ R*F*×*F* denote the trainable FTM and let *Sa*(*t*0) ∈ R*F*×*CA* (*t*0 ∈ {0*,*1*,...,T* − 1}) denote the feature slice at time step *t*0. The transformation can be simply represented by the following equation:

等式 （10）将TF注意模块描述为在图3中的虚线框中突出显示。利用输入特征SI，它使用2D和1D卷积层来预测注意图，然后将其逐点乘以SI以获取Sa 。 2D 1×1卷积将通道数减少为Cr = 5，一维卷积的内核大小为9。

Freq-FC是FTB中的关键组件。 它包含一个可训练的频率变换矩阵（FTM），该矩阵在每个时间点应用于特征图切片。 设Xtr∈RF×F表示可训练的FTM，设*Sa*(*t*0) ∈RF×CA（t0∈{0,1，...，T − 1}）表示时间步t0的特征切片。 转换可以简单地由以下等式表示：

*Str*(*t*0) = *Xtr* · *Sa*(*t*0)*.* (13)

The transformed feature slice at time step *t*0, denoted by *Str*(*t*0), has the same dimension as *Sa*(*t*0). Stacking them along the time axis and we can get the transformed feature map *Str*. After Freq-FC, each T-F bin in *Str* will contain the information from all the frequency bands of *Sa*. This allows the following blocks to exploit global frequency correlations for amplitude and phase estimation.

The output of an FTB, denoted by *SO*, is calculated by concatenating *Str* with *SI* 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.

在时间步t0处变换的特征切片（由*Str*(*t*0)表示）具有与*Sa*(*t*0)相同的维度。 沿时间轴堆叠它们，我们可以获得变换后的特征图Str。 在Freq-FC之后，Str中的每个T-F bin将包含来自Sa的所有频带的信息。 这允许以下块利用全局频率相关性进行幅度和相位估计。

FTB的输出（用SO表示）是通过将*Str*与SI串联在一起，并通过1×1卷积将它们融合来计算的。 在提出的FTB中，在所有卷积层之后使用批处理归一化（BN）和ReLU作为归一化方法和激活函数。

## 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 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 powerlaw compressed STFT spectrogram. The loss consists of two parts: amplitude loss *La* and phase-aware loss *Lp*.

PHASEN在Pytorch中实现。 特征图的尺寸和卷积层的内核大小如图2和图3所示。两个流均使用零填充，dilation = 1和stride = 1的卷积运算，确保输入和输出特征图的大小为 相同。 所有音频都重新采样到16kHz。 STFT使用Hann窗口计算，其窗口长度为25ms。 跳长为10ms，FFT大小为512。

使用powerlaw压缩STFT频谱图上的MSE损失来训练网络。 Loss由两部分组成：幅度Loss La和相位感知Loss Lp。

|  |  |
| --- | --- |
| *L* = 0*.*5 × *La* +0*.*5 × *Lp,* | (14) |
| *,* | (15) |
| *,* | (16) |

whereare the power-law compressed spectrogram of output spectrogram *Sout* and ground truth spectrogram *Sgt*. 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 *Lp*. 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.

用表示输出频谱图Sout和地面真实频谱图Sgt的幂律压缩频谱图。 用取值p = 0.3（A0.3，其中A是频谱图的振幅）对振幅执行压缩。

请注意，对于Lp，不仅要考虑使用纯相位，还要考虑整个频谱图（相位和幅度）。 以此方式，放大了具有较高幅度的T-F仓的相位，从而帮助网络集中于大多数语音信号所位于的高幅度T-F仓。

# 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). Audios from AVSpeech dataset are used as clean speech. It is collected from YouTube, containing 4700 hours of video segments with approximately 150,000 distinct speakers, spanning a wide variety of people and languages. The noisy speech is a mixture of the above clean speech segments with AudioSet (Gemmeke et al. 2017), which contains a total of more than 1.7 million 10-second segments of 526 kinds of noise. The noisy speech is synthesized by a weighted linear combination of speech segments and noise segments: *Mixi* = *Speechj* + 0*.*3 × *Noisek*, where *Speechj* and *Noisek* are 3-second segments randomly sampled from speech and noise dataset. *Mixi* and *speechj* form a noisy-clean speech pair. 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]](#footnote-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.

我们的实验中使用了两个数据集。

AVSpeech + AudioSet：这是（Ephrat等人2018）提出的大型数据集。来自AVSpeech数据集的音频用作纯净语音。它是从YouTube收集的，其中包含4700小时的视频片段，大约有15万名不同的演讲者，涉及各种各样的人和语言。嘈杂的语音是以上干净语音片段与AudioSet的混合（Gemmeke et al.2017），其中包含超过170万个10秒片段，共526种噪声。嘈杂的语音是由语音段和噪声段的加权线性组合合成的：Mixi = Speechj + 0.3×Noisek，其中Speechj和Noisek是从语音和噪声数据集中随机采样的3秒段。 Mixi和Speechj构成一个噪声干净的语音对。在我们的实验中，从AVSpeech数据集和AudioSet的``平衡训练''部分随机采样的10万个片段用于合成训练集，而验证集与（Ephrat等人2018）中使用的相同，由AVSpeech数据集的测试部分和AudioSet的评估部分。

Voice Bank + DEMAND：这是（Valentini-Botinhao等人2016）提出的开放数据集。语音库语料库（Ephrat等人2018）的30位演讲者的语音被选为干净的语音：训练集中包括28位，验证集中包括2位。嘈杂的语音是将干净的语音与来自Diverse Environments多通道声学噪声数据库（DEMAND）的噪声混合而成（Thiemann，Ito和Vincent，2013年）。训练集中考虑了40种不同的噪声条件，而测试集中考虑了20种不同的条件。最后，训练和测试集分别包含11572和824个干净的语音对。训练集完全看不到测试集中的说话者和噪声状况。我们的系统比较部分地在此数据集上完成。

## 4.2 Evaluation Metrics

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

* SDR (Vincent, Gribonval, and Fevotte 2006): Signal-to-´ distortion ratio from the mir eval library;
* PESQ: Perceptual evaluation of speech quality (from -0.5 to 4.5).
* 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 .

以下六个指标用于评估PHASEN和最先进的竞争对手。 所有这些指标越高越好。

•SDR（Vincent，Gribonval和Fevotte，2006年）：镜像库中的信噪比；

•PESQ：语音质量的感知评估（从-0.5到4.5）。

•CSIG（Hu和Loizou，2007年）：仅对语音信号造成的信号失真（从1到5）的平均观点分数（MOS）预测。

•CBAK（Hu和Loizou，2007年）：MOS对背景噪声干扰的预测（从1到5）。

•COVL（Hu and Loizou 2007）：MOS对总体影响的预测（从1到5）。

•SSNR：分段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 Table 1: Ablation study on AVSpeech + AudioSet optimizer with a fixed learning rate of 0.0002 is used and the batch size is set to 8. We use mean SDR and PESQ on test dataset as the evaluation metric.

在消融研究中，使用相同的随机种子对不同设置的网络进行了1百万步的训练。 亚当表1：使用固定学习率为0.0002的AVSpeech + AudioSet优化器进行的消融研究，并且批大小设置为8。我们使用测试数据集上的平均SDR和PESQ作为评估指标。

|  |  |  |
| --- | --- | --- |
| Method | SDR(dB) | PESQ |
| PHASEN-baseline | 15.08 | 2.87 |
| PHASEN-1strm | 15.99 | 2.98 |
| PHASEN-w/o-FTBs | 16.10 | 3.31 |
| PHASEN-w/o-A2PP2A | 16.13 | 3.33 |
| PHASEN-w/o-P2A | 16.62 | 3.38 |
| PHASEN | 16.84 | 3.40 |

The ablation results are shown in Table 1. Among these methods, PHASEN represents our full model. PHASENbaseline 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 of 1.76dB improvement on SDR and 0.53 improvement on PESQ.

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 0.85dB gain on SDR and 0.42 gain on PESQ. The large gain on PESQ indicates the proposed two-stream architecture can largely improve the perceptual quality of the denoised speech.

FTBs The proposed method uses FTBs at both the beginning and the end of each TSB. In ablation study, PHASENw/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 and 0.09 gain on SDR and PESQ, 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 and 0.03-0.05 drop on PESQ, 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 *Xtr*, the matrix that reflects the learned global frequency correlation. From Fig. 4, we show that the energy map of *Xtr* 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 spontaneously.

Information communication mechanism PHASENw/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.

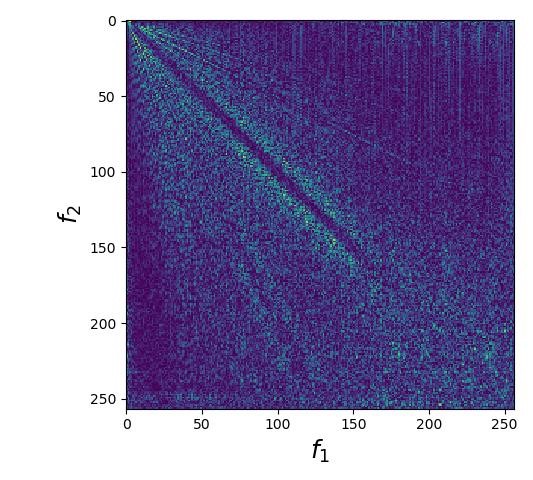
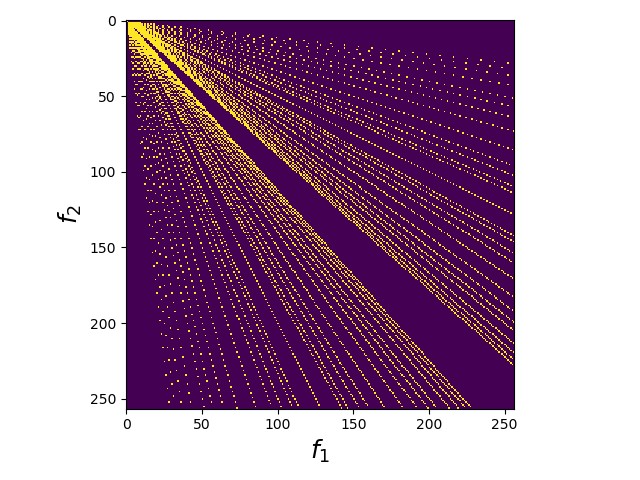
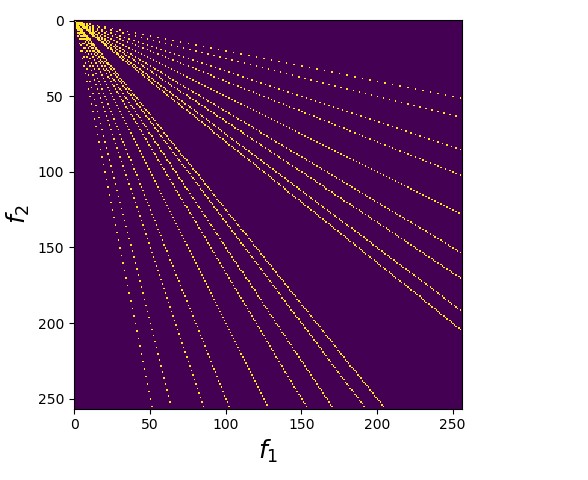
消融结果如表1所示。在这些方法中，PHASEN代表了我们的完整模型。 PHASENbaseline代表使用cIRM作为训练目标的单流网络。我们将流A中的网络结构用于PHASEN基线，并用5×5卷积替换FTB。 PHASEN与PHASEN基准线之间的比较表明，我们的两项创新（即双流架构和FTB）在SDR上总共提高了1.76dB，在PESQ上总共提高了0.53。

两流体系结构PHASEN-1strm显示了以cIRM为训练目标的单流体系结构的性能。在该实验中，从PHASEN体系结构中删除了流P和信息沟通，同时保留了FTB。流A的输出是预测的cRM。 PHASEN-1strm和PHASEN之间的比较表明，两流架构在SDR上提供0.85dB的增益，在PESQ上提供0.42的增益。 PESQ的巨大收益表明，所提出的两流体系结构可以大大提高去噪语音的感知质量。

FTB建议的方法在每个TSB的开头和结尾都使用了FTB。在消融研究中，PHASENw / o-FTB尝试用5×5卷积替换PHASEN体系结构中的所有FTB。通过将PHASEN与PHASEN-w / o-FTB进行比较，我们发现FTB可以分别在SDR和PESQ上提供0.74 dB和0.09的增益。我们还尝试用5×5卷积替换每个TSB任一位置上的FTB。两种尝试都会导致SDR下降0.31dB-0.39dB，而PESQ下降0.03-0.05，这表明两个位置的FTB同等重要，并且增益是累积的。

为了更好地了解FTB，我们将Xtr的权重可视化，该矩阵反映了所学的全局频率相关性。从图4中我们可以看出，Xtr的能量图类似于谐波相关性，尤其是在考虑了高次谐波（H较大）的情况下。这种现象证实了FTB确实捕获了谐波相关性，并且谐波相关性对于语音增强网络确实有用，因为该网络可以自发地学习此相关性。

信息沟通机制PHASENw / o-P2A和PHASEN-w / o-A2PP2A是两个设置，部分和完全删除了信息沟通机制。前者将流P到流A的通信删除，而后者将两个方向的通信都删除。在SDR和PESQ结果中，将PHASEN-w / o-P2A与PHASEN-w / o-A2PP2A进行比较时，可观察到0.49dB和0.05的显着增益。这表明幅度预测的中间步骤中的信息对于相位预测非常有帮助。比较我们的完整模型PHASEN和PHASEN-w / o-P2A，我们还发现，将流P信息集成到流A中时，该模型在SDR上获得0.22dB的增益，在PESQ上获得0.02的增益。这证明了相位特征也可以帮助幅度预测。



H=5 H=9 Learned FTM weights

Figure 4: Comparison of different level of harmonic correlation: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 ∆Ψ = Ψ*/*Ψ*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 difference between phase prediction and amplitude mask prediction. This supports our design of using two streams to accomplish the two prediction tasks.

图5还通过可视化确认了上述改进。在这里，由于预测的相位谱图几乎没有可见的图案，因此我们可视化∆Ψ =Ψ/Ψin，它表示预测的相位谱图和输入噪声频谱图之间的相位差。该公式中的除法运算是在复数域上进行的，in表示输入有声语音的相位谱图。从可视化中，我们可以得出结论，信息沟通机制不仅可以显着改善相位预测，而且可以消除振幅伪影。总而言之，双向信息交流在PHASEN中非常有用，而方向“ A2P”起着关键作用。

其他烧蚀除了表1中显示的结果外，我们还对流P的激活函数和归一化函数执行烧蚀。

所提出的方法在流P上不使用激活函数。尽管这种设计是违反直觉的，但它实际上是受到先前工作的启发（Luo和Mesgarani 2019），并且得到了消融研究的支持。实际上，我们尝试在每个流之后添加ReLU或Tanh作为激活函数，除了流P中的最后一个卷积层。但是，这会导致SDR下降0.02dB-0.16dB。此外，如果在流P的最后一个卷积层之后添加ReLU，则在SDR和PESQ上会观察到5.52dB和0.2的巨大下降。

所提出的方法在流P中使用gLN，在流A中使用BN。我们针对每个流测试其他归一化方法。如果在流A中使用gLN，则在SDR和PESQ上性能下降0.97dB和0.12，而在流P中使用BN时，在SDR和PESQ上性能下降0.09dB和0.02。

从这两个实验中，我们可以观察到相位预测和幅度模板预测之间的显着差异。这支持了我们使用两个流来完成两个预测任务的设计。

(

a)

(

b

)

)

d

(

(

c)

f

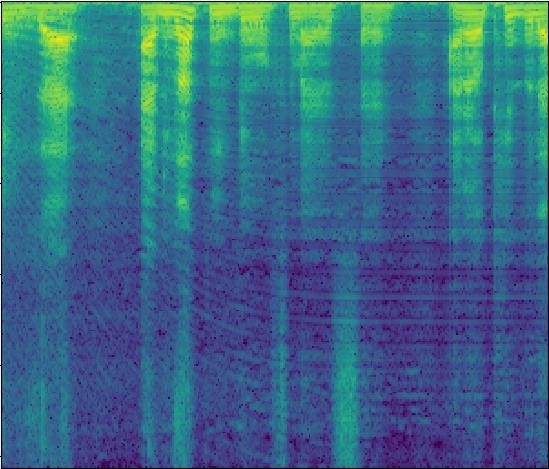
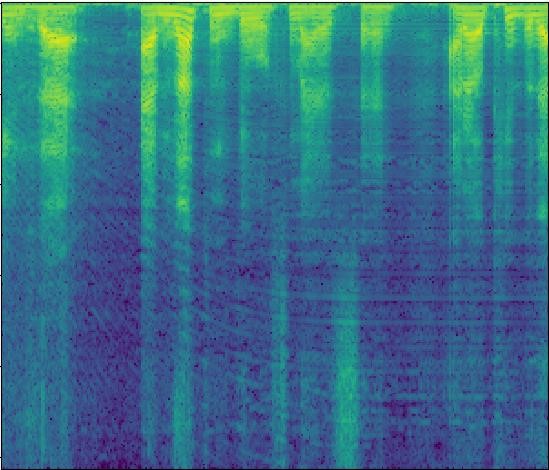
(

)

)

e

(



Amplitude

artifacts

PHASEN

-

w/o

-

A2PP2A

PHASEN

Predicted Spectrogram

Re(

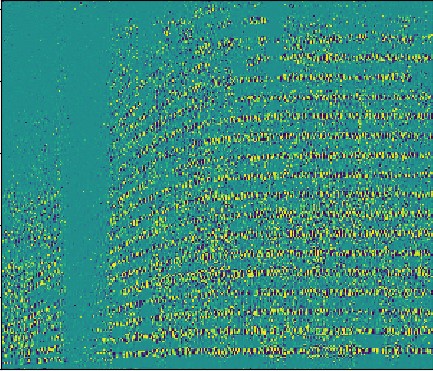
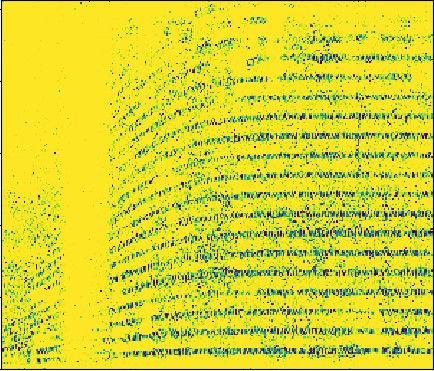
ΔΨ

)

Im(

ΔΨ

)



Phase

prediction

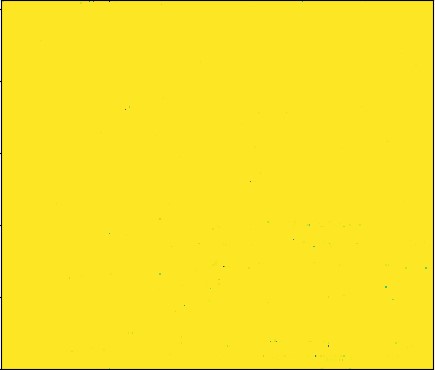
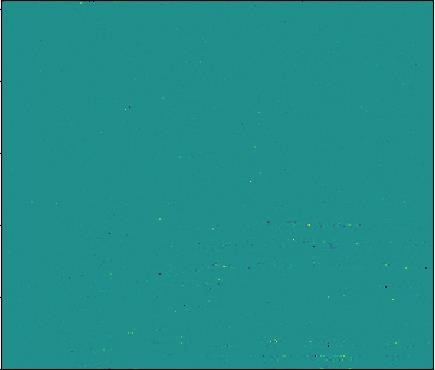


Figure 5: The effect of information communication mechanism. Best viewed in color. We use the same input noisy speech as the one in Fig.2 to produce the results. (a),(b),(c): Amplitude of predicted spectrogram, real part, and imaginary part of ∆Ψ in setting PHASEN-w/o-A2PP2A. Significant amplitude artifacts 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. (d),(e),(f): Amplitude of predicted spectrogram, real part, and imaginary part of ∆Ψ in setting PHASEN. In (e) and (f), phase prediction is obviously visible in T-F bins where noise overwhelms speech. (d) also shows fewer artifacts on amplitude spectrogram.

## 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, ConvTasNet (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]](#footnote-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 noise reduction and speech separation. We compare PHASEN with their audio-only, 1S+noise setting. The result in Table 2 shows that our method outperforms both ConvTasNet 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 Apart from using large dataset, we also train our model on small but commonly-used dataset Table 2: System comparison on AVSpeech + AudioSet Voice Bank + DEMAND, so that we can fairly compare our PHASEN with many other methods. In this experiment, our network is trained on 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.

我们对4.1节中提到的两个数据集进行系统比较。

AVSpeech + AudioSet在这个庞大的数据集上，我们将我们的方法与其他两种最新方法进行了比较，ConvTasNet（Luo和Mesgarani 2019）和“ Google”（Ephrat等人，2018）。 Conv-TasNet是时域方法。使用发布的代码生成Conv-TasNet的结果，并针对与我们的PHASEN相同的纪元和相同的数据进行训练。 “ Google”是一种使用cIRM作为监督的T-F域掩蔽方法。该方法既用于语音降噪又用于语音分离。我们将PHASEN与仅音频，1S +噪声设置进行了比较。表2中的结果表明，我们的方法优于ConvTasNet和“ Google”。请注意，这是在我们仅使用“ Google”使用的一小部分训练步骤（1M / 5M）和数据（100k / 2.4M）的情况下实现的。这种在大型数据集上的出色性能表明，我们的方法可以推广到各种说话者和各种嘈杂的环境。这表明PHASEN很容易适用于复杂的现实环境。

语音库+ DEMAND除了使用大型数据集以外，我们还在较小但常用的数据集上训练模型。表2：AVSpeech + AudioSet语音库+ DEMAND上的系统比较，以便我们可以将PHASEN与许多其他方法进行合理比较。在此实验中，我们的网络在40个纪元的训练集上进行了训练，使用Adam优化器，其预热步数为6000，学习率为0.0005，批次大小为12。

|  |  |  |
| --- | --- | --- |
| Method | SDR(dB) | PESQ |
| Conv-TasNet | 14.19 | 2.93 |
| Google(5M step, 2.4M speech) | 16.00 | – |
| PHASEN(1M step, 100k speech) | 16.84 | 3.40 |

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 timedomain 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.

表3示出了比较结果。 首先，我们的方法在所有方面都比SEGAN（Pascual，Bonafonte和Serra 2017），Wavenet（Rethage，Pons和Serra 2018）和DFL（Germain，Chen和Koltun 2018）等时域方法有很大的收益。 五个指标，即使这些时域方法没有相位预测问题。 这证明了我们的方法在捕获相位相关信息方面优于时域方法。 此外，我们的方法在所有指标上都比MMSE-GAN（Soni，Shah和Patil 2018）等时频域方法有了很大的改进，表明了我们网络设计的优越性。 最后，我们还将我们的方法与最近的时域和时频域混合模型MDPhD进行了比较（Kim等人2018）。 我们的方法在四个指标上的性能明显优于它，而在SSNR指标上仅存在约0.04dB的微小差异。

# 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 spontaneously 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 stateof-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.

我们提出了一种具有双向信息沟通的两流体系结构，用于单声道语音增强中的有效相位预测。 我们还设计了网络中可学习的频率转换矩阵。 它自发地学习与谐波相关一致的模式。 已经进行了全面的消融研究，证明了我们在PHASEN中所做的几乎每个设计选择都是合理的。 在AVSpeech + AudioSet和语音库+ DEMAND数据集上与最新系统的比较证明了PHASEN的卓越性能。 请注意，PHASEN的当前设计不允许将其用于低延迟应用程序，例如IP语音。 将来，我们计划在需要较小模型尺寸和更短推理时间的低延迟设置和移动设置中探索PHASEN的潜力。 我们还计划将该体系结构扩展到其他相关任务，例如语音分离。

# References

[Ephrat et al. 2018] 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 et al. 2015] Erdogan, H.; Hershey, J. R.; Watanabe, S.; and Le Roux, J. 2015. Phase-sensitive and recognitionboosted speech separation using deep recurrent neural networks. In *ICASSP 2015*, 708–712. IEEE.

[Gemmeke et al. 2017] 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 humanlabeled dataset for audio events. In *ICASSP 2017*, 776–780. IEEE.

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

[Hu and Loizou 2007] 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 and Wang 2001] 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 et al. 2017] 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 et al. 2018] 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 and Gerkmann 2014] 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 and Mesgarani 2019] 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 et al. 2019] Masuyama, Y.; Yatabe, K.; Koizumi, Y.; Oikawa, Y.; and Harada, N. 2019. Deep griffin–lim iteration. In *ICASSP 2019*, 61–65. IEEE.

[Mowlaee and Kulmer 2015] 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.

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

[Pandey and Wang 2019] 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, Bonafonte, and Serra 2017] Pascual, S.; Bonafonte, A.; and Serra, J. 2017. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.

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

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

[Soni, Shah, and Patil 2018] 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, Roman, and Wang 2006] Srinivasan, S.; Roman, N.; and Wang, D. 2006. Binary and ratio timefrequency masks for robust speech recognition. *Speech Communication* 48(11):1486–1501.

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

[Thiemann, Ito, and Vincent 2013] 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 et al. 2016] 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, Gribonval, and Fevotte 2006] Vincent, E.; Gribon-´ val, R.; and Fevotte, C. 2006. Performance measurement in´ blind audio source separation. *IEEE transactions on audio, speech, and language processing* 14(4):1462–1469.

[Wakabayashi et al. 2018] Wakabayashi, Y.; Fukumori, T.; Nakayama, M.; Nishiura, T.; and Yamashita, Y. 2018. Single-channel speech enhancement with phase reconstruction based on phase distortion averaging. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26(9):1559–1569.

[Wang, Narayanan, and Wang 2014] 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.

[Williamson, Wang, and Wang 2016] Williamson, D. S.; Wang, Y.; and Wang, D. 2016. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*

24(3):483–492.

1. https://datashare.is.ed.ac.uk/handle/10283/1942 [↑](#footnote-ref-1)
2. https://github.com/funcwj/conv-tasnet [↑](#footnote-ref-2)