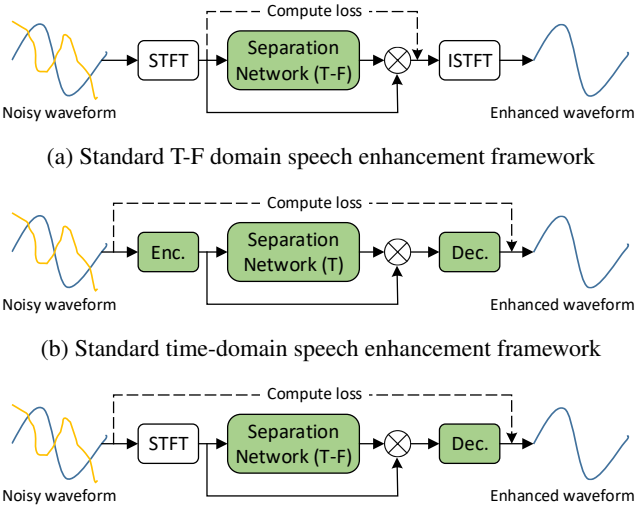
Joint Time-Frequency and Time Domain Learning for Speech Enhancement

Chuanxin Tang, **Chong Luo**, **Zhiyuan Zhao**, **Wenxuan Xie** and **Wenjun Zeng**
Microsoft Research Asia
{chutan, cluo, zhiyzh, wenxie, wezeng}@microsoft.com

Abstract

For single-channel speech enhancement, both time-domain and time-frequency-domain methods have their respective pros and cons. In this paper, we present a cross-domain framework named TFT-Net, which takes time-frequency spectrogram as input and produces time-domain waveform as output. Such a framework takes advantage of the knowledge we have about spectrogram and avoids some of the drawbacks that T-F-domain methods have been suffering from. In TFT-Net, we design an innovative dual-path attention block (DAB) to fully exploit correlations along the time and frequency axes. We further discover that a sample-independent DAB (SDAB) achieves a good trade-off between enhanced speech quality and complexity. Ablation studies show that both the cross-domain design and the SDAB block bring large performance gain. When logarithmic MSE is used as the training criteria, TFT-Net achieves the highest SDR and SSNR among state-of-the-art methods on two major speech enhancement benchmarks.



(a) Standard T-F domain speech enhancement framework

(b) Standard time-domain speech enhancement framework

(c) Proposed cross-domain speech enhancement framework

Figure 1: Illustration of different speech enhancement frameworks. The color-filled modules are learnable neural networks. The separation network, either in T-F domain or in time domain, separates the features of clean speech from that of noise.

1 Introduction

Single-channel speech enhancement addresses the problem of recovering clean speech from a noise-corrupted speech. According to the signal domain they work in, existing methods can be classified into two categories: time-frequency (T-F) domain methods and time-domain methods.

The T-F domain methods operate on the two-dimensional time-frequency spectrogram. As shown in Fig.1a, they often use short-time Fourier transform (STFT) to convert a raw audio signal into a T-F spectrogram. Then a T-F mask is predicted by the separation network. Finally, the output spectrogram is converted back to a time domain signal by the inverse STFT (ISTFT). T-F masking methods have been a great success for computational auditory scene analysis (CASA). They are also the mainstream methods for speech enhancement in the deep learning era. The theoretical ground for T-F analysis is that auditory patterns, such as proximity in frequency and time, harmonicity, and common amplitude and frequency modulation, are revealed on a T-F spectrogram [Wang and Chen, 2018]. However, the T-F domain methods are facing some challenges. First, STFT transforms the input signal into a complex domain with magnitude and phase. The phase of a signal is difficult to modify, but ignoring it will create a performance upper bound. Second, in supervised deep learning, a loss is computed at the output of the learnable network. It is a natural choice for T-F methods to compute the mean square error (MSE) loss for the T-F spectrogram. However, minimizing the MSE of spectrogram does not necessarily results in maximized signal-to-distortion ratio (SDR) of the output speech. This is known as the metric mismatch problem.

Time-domain methods try to avoid or solve these two problems by directly processing the raw waveform. Fig.1b shows the standard time-domain speech enhancement framework. Time-domain methods directly model the mixture waveform using an encoder-decoder framework and perform the separation on the output of the encoder. As phase no longer exists, the phase prediction problem is avoided. Besides, the loss is computed at the output of the learnable decoder. One is free

to pick whatever metric that is important to the speech quality. As such, the metric mismatch problem is solved. However, the separation network in a time-domain method cannot leverage the known auditory patterns on a T-F spectrogram. This will potentially bring a performance loss.

A natural question to ask is whether it is possible to take advantages of methods in the two domains and avoid their respective drawbacks. Luckily, we find that the advantages of the T-F domain methods are mainly reflected in the front of the network and those of the time-domain methods are mainly in the latter part of the network. But an obstacle to connect these two parts is that the encoder and the decoder have to be paired in the conventional wisdom. Thanks to the powerful learning capability of deep neural network, this is no longer an issue with a learnable decoder. Fig.1c shows the flow chart of the proposed cross-domain framework, dubbed TFT-Net.

In TFT-Net, the input waveform is converted to T-F spectrogram through the predefined STFT. The subsequent T-F domain separation network can make full use of the auditory patterns revealed on the spectrogram. After the masked spectrogram is obtained, a learnable decoder can fully recover any spectrum modification after STFT. Note that supervision signal is given only after the decoder, and we do not intend to recover the ground-truth T-F spectrogram before the decoder. In fact, the output spectrogram is in a latent domain defined by the learned decoder. In the design of the separation network, we find that long-range correlations on the T-F spectrogram are important to the denoising performance. To tackle the high complexity, we propose dual-path attention blocks (DAB) to exploit the correlations along time and frequency dimensions in parallel. We further discover that the attention maps for different samples resemble each other, indicating sample-independent correlation is sufficient. Therefore, we employ sample-independent DABs (SDABs) in TFT-Net to balance the performance and computational cost.

In a nutshell, the contributions of our work are three-fold:

- We propose a cross domain learning framework TFT-Net for speech enhancement. It takes advantages of both time domain and T-F domain approaches.
- We propose a novel sample-independent dual-path attention block (SDAB) to capture long-range temporal and frequency correlations with low computational cost.
- We perform comprehensive ablation studies and evaluate the proposed system TFT-Net on two major speech enhancement benchmarks. Results show that TFT-Net outperforms the state-of-the-art methods.

2 Related Work

This section reviews single-channel speech enhancement methods in different domains. Focuses are put on masking-based methods in which the separation network generates a mask that describes the relationship between clean speech features and noisy features.

2.1 T-F Domain Methods

In a typical T-F domain masking method, as shown in Fig.1a, a pair of predefined encoder and decoder are used to convert the signal between time-domain waveform and T-F domain spectrogram. STFT and ISTFT are the most widely used transforms for the encoder and the decoder, respectively. A complete T-F spectrogram is composed of magnitude and phase. Early T-F masking methods only try to estimate the magnitude of the spectrogram due to the difficulty in modifying the phase. Later, research [Paliwal *et al.*, 2011] reveals that phase also plays an important role in speech enhancement. There is a performance upper bound if only the magnitude is enhanced. PSM [Erdogan *et al.*, 2015] and cIRM [Williamson *et al.*, 2016] are then proposed to recover the phase in the spectrogram. While PSM is still a real-valued mask, cIRM is a complex-valued mask which can potentially recover both amplitude and phase of the clean speech.

In order to predict cIRM, the authors [Williamson *et al.*, 2016] propose a DNN-based approach to estimate the real and imaginary components of the cIRM. However, the 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 this work. In [Ephrat *et al.*, 2018], a much deeper neural network with dilated convolution and bi-LSTM is designed. Dilated convolution enlarges the receptive field and bi-LSTM learns long-range correlations along the time axis. More recently, Yin et al. [Yin *et al.*, 2019] propose a two-stream network for both magnitude and phase estimation. A frequency transformation block (FTB) is used in the front of the network to capture harmonic correlations along the frequency axis. A bi-LSTM is used at the latter part of the network to capture temporal dependencies. In [Kim *et al.*, 2019], the spectrogram is treated as a time sequence and a Transformer is employed to capture long-range temporal correlations.

All these T-F domain methods benefit from the rich auditory patterns in the T-F spectrogram. However, we notice that when the long-range correlations along the time and the frequency axes are both considered, they are always learned in separate steps in prior works. We believe that learning long-range correlations along both axes is important, as harmonics exist in speeches [Plapous *et al.*, 2005], [Krawczyk and Gerkmann, 2014], [Wakabayashi *et al.*, 2018] and distinguishing noise needs long-term statistics. We also believe that learning the two types of correlations should be performed in parallel and the learned information needs to be fully fused. This becomes the motivation of our DAB design.

2.2 Time Domain Methods

Time-domain methods have emerged recently to tackle the difficulty in phase estimation and the metric mismatch problem in T-F domain methods. The most influential work in this category is Conv-TasNet [Luo and Mesgarani, 2019] which uses a pair of learnable encoder-decoder in time domain as an alternative to the predefined STFT-ISTFT. Before Conv-TasNet, SEGAN [Pascual *et al.*, 2017] uses generative adversarial networks (GANs) to directly predict the 1D waveform of the clean speech. In [Rethage *et al.*, 2018], Wavenet is modified for speech enhancement. TCNN [Pandey and Wang, 2019] adopts a similar approach as Conv-TasNet, but it uses nonlinear encoder-decoder and longer frame length than Conv-TasNet.

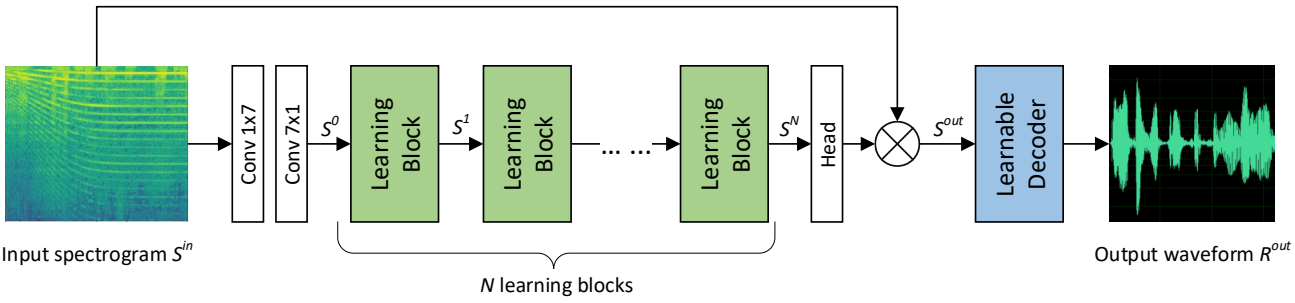


Figure 2: Illustration of the proposed TFT-Net framework for speech enhancement. The input signal is the T-F domain spectrogram and the output signal is the time-domain signal. TFT-Net stacks several learning blocks to learn a multiplicative mask for the spectrogram. The predicted spectrogram in the latent domain is decoded into time-domain signal through a learnable decoder.

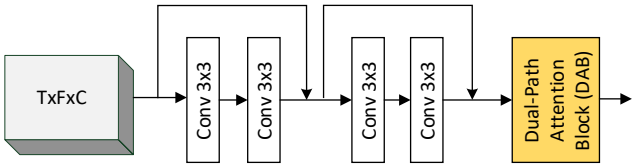


Figure 3: Each learning block is composed of four 3×3 convolutional layers with skip connections followed by a DAB.

These methods achieve their design goal in the sense that they avoid the problems of T-F domain methods. However, the most impressive performance is achieved by Conv-TasNet in the speech separation task. When applied to speech enhancement, its performance is inferior to the top-performing T-F domain method [Yin *et al.*, 2019]. A possible reason for this phenomenon is that speech and noise patterns are easily distinguishable on T-F domain representations, and time-domain methods cannot make use of such prior knowledge.

3 The Proposed Scheme

We design a speech enhancement system with two objectives in mind. First, it should make full use of the prior knowledge about the T-F spectrogram. To realize this, the pipeline starts from the spectrogram and we design dual-path attention blocks to exploit the long-range correlations along both the time and frequency axes. Second, it should avoid the drawbacks of T-F domain methods. To this end, we propose an innovative cross-domain framework which directly uses time-domain metric to supervise the network learning.

3.1 The TFT-Net Framework

The most prominent feature of the proposed TFT-Net framework is that it starts with T-F spectrogram and directly outputs time-domain waveform. Fig.2 gives detailed illustration of the TFT-Net framework. Compared to Fig.1c, we omit the predefined STFT module so that we can concentrate on the end-to-end fully learnable parts.

Formally, the input noisy waveform $R^{in} \in \mathbb{R}$ is first transformed into STFT spectrogram, denoted by S^{in} in Fig.2. 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 two two-dimensional convolution layers to produce feature $S^0 \in \mathbb{R}^{T \times F \times C}$, where C is the number of channels. S^0 is then fed into N stacked learning blocks.

Each learning block consists of two parts, as shown in Fig.3. The first part is a group of 3×3 convolution layers that focus on capturing local correlations. We use four convolution layers and each of them is followed by batch normalization [Ioffe and Szegedy, 2015]. Around each of the two convolution layers, we employ a skip connection [He *et al.*, 2016]. The second is the DAB which focuses on capturing long-range correlations and we will describe it in Section 3.2. The output features of each learning block are denoted by S^i for $i \in \{1, 2, ..., N\}$ where N is an adjustable hyper-parameter. They have the same dimension as S^0.

After S^N is produced by the last learning block, it is fed into the head layer to predict mask $M \in T \times F \times C_m$. Here, C_m is the number of channels. The enhanced spectrogram S^{out} in the latent domain can be calculated by the following generic function:

$$S^{out} = f(M) \odot g(S^{in}) \tag{1}$$

In our framework, we use the training target in [Choi *et al.*, 2019]. So the head layer is a 1×1 convolution layer with $C_m = 2$. $f(M) = tanh(|M|) * M/|M|$, $g(S_{in}) = S_{in}$ and $\odot$ indicates complex multiplication. Finally, S^{out} is fed into the decoder layer to produce the enhanced time-domain signal $R^{out} \in \mathbb{R}$. Logarithmic MSE loss is computed between the original clean waveform and R^{out}.

In conventional wisdom, the encoder and the decoder need to be paired. Although Conv-TasNet [Luo and Mesgarani, 2019] has shown that the decoder does not have to perform the exact inverse operation of the encoder, we use the following method to ensure the convergence of our network. In short, the network structure of the learnable decoder is designed to be exactly the same as the convolutional implementation of ISTFT. Specifically, the learnable decoder is a 1D transposed convolution layer whose kernel size and stride being the window length and hop size in the STFT. As such, TFT-Net becomes a fully end-to-end learnable system.

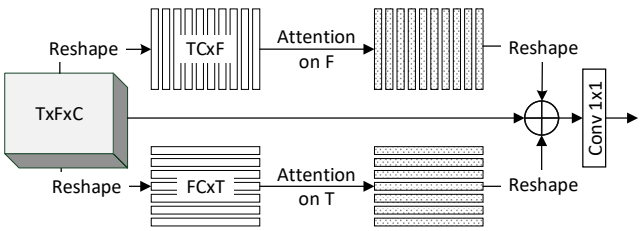


Figure 4: Illustration of DAB. The main idea is to apply attention along the time and the frequency axes in parallel, before combining them with the original features. In SDAB, FC layers are used to learn sample-independent correlations.

3.2 Dual-path Attention Block

Long-range correlations exist along both the time and the frequency axes in a T-F spectrogram. It is quite obvious that audio signals, as time series, contain global correlations along the time axis. Along the frequency axis, there are well-known harmonic correlations. Although it is possible to treat the spectrogram as a 2D image and learn the correlations between every two pixels in the 2-D image, this is computationally too costly. DAB is designed to be a light-weight solution to capture the long-range correlations exhibited in T-F spectrogram.

As Fig.4 shows, we decompose the 2D T-F spectrogram (with C channels) into two stacks of 1D signals, one along the frequency axis and the other along the time axis. When the input feature dimension is $T \times F \times C$, it can be reshaped into TC vectors with dimension $1 \times F$ or FC vectors with dimension $1 \times T$. Then, attention is applied along each axis in parallel. The attended features and the original features are then combined to generate the final output.

Ideally, we can use self-attention [Vaswani *et al.*, 2017] to learn the attention map for each sample. But this might not be necessary. In classic speech enhancement literature [Scalart and FILHO, 1996], a uniform non-linear function is applied to the frequency axis to regenerate harmonics. It also matches with the intuition that harmonic correlations are sample-independent. On the time axis, when calculating Signal-to-Noise Ratio (SNR), the same set of parameters in recursive relation are used, which suggests that temporal correlation is time-invariant. Based on this understanding, we propose a sample-independent DAB, or SDAB. Specifically, the attention layer on F and T are implemented by fully-connected (FC) layers. Along the time path, the input and output dimensions of FC layers are T. Along the frequency path, the input and output dimensions of FC layers are F. FC layer is a kind of attention operation that uses learned weights. In other words, the relationship between every two elements is not a function of the input data. So using a FC layer is equivalent to learning a sample-independent correlations for all the training data. SDAB further reduces the computational complexity of the proposed network. In the experiment section, we will show through ablation study that SDAB performs almost as good as a sample-dependent self-attention.

4 Experiments

4.1 Datasets

AVSpeech+AudioSet

Audios from the AVSpeech dataset are used as clean speech. It is a large dataset proposed by [Ephrat *et al.*, 2018]. 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: $Mix_i = Speech_j + 0.3 \times Noise_k$, where $Speech_j$ and $Noise_k$ are 3-second segments randomly sampled from the speech and noise dataset. In our experiments, 100k segments are randomly sampled from the 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 proposed by [Valentini-Botinhao *et al.*, 2016]. Speeches of 30 speakers selected from Voice Bank corpus [Veaux *et al.*, 2013] are used 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 the DEMAND dataset [Thiemann *et al.*, 2013]. A total of 40 different noise conditions are considered in the training set and 20 different conditions are considered in the 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.

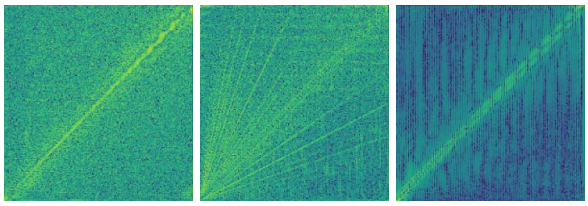
4.2 Evaluation Metrics

We mainly use SDR [Vincent *et al.*, 2006] and PESQ (perceptual evaluation of speech quality) in the ablation study. We further adopt CSIG [Hu and Loizou, 2007], CBAK [Hu and Loizou, 2007], COVL [Hu and Loizou, 2007], and SSNR (Segmental SNR) in the system comparison. For all the metrics, a higher value means a better result.

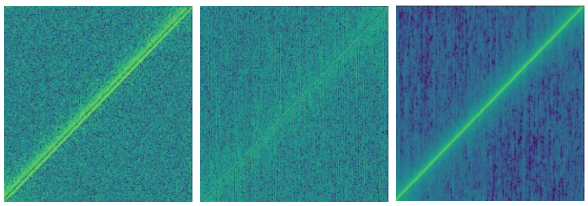
4.3 Implementation Details

Our method is implemented in Pytorch. All audios are resampled to 16kHz. STFT is computed using a Hann window of length 25ms, hop length of 10ms, and FFT size of 512, resulting in an input audio feature of $301 \times 257 \times 2$ scalars. Convolution operation with zero padding, dilation=1 and stride=1 is used, making sure the input and output of the features are the same size. The number of channels for each convolution layer is 96. ReLU activations follow all network layers except for head layer (mask). Batch normalization [Ioffe and Szegedy, 2015] is performed after all convolutional layers.

When training model with the T-F domain framework, we use the training target and loss function in [Yin *et al.*, 2019]. When training model with the cross-domain framework (TFT-Net), the training target in [Choi *et al.*, 2019] is used and the loss function is logarithmic MSE.



(a). Learned weights of SDAB along frequency path.



(b). Learned weights of SDAB along time path.

Figure 5: Illustrations of learned weights of SDAB along frequency path and time path. From left to right is the 1st, 4th and 6th learned weights of SDAB. There are six SDABs in total.

Method	Attn_F	Attn_T	SDR(dB)	PESQ
TF-2stream			16.10	3.31
	√(*)		16.84	3.40
	√	√	17.41	3.45
TF-1stream			13.88	2.99
	√		15.57	3.27
		√	16.15	3.22
	√	√	17.56	3.46

Table 1: Ablation study on AVSpeech + AudioSet. TF-2stream represents PHASEN without FTB [Yin *et al.*, 2019]. TF-1stream represents TFT-Net method without decoder layer and six learning blocks are used. * represents applying attention along frequency by using FTB. Without *, means applying attention by using fc layer.

Method	Sample-dependent		Sample-independent	
	SDR	PESQ	SDR	PESQ
TF-2stream	17.45	3.41	17.41	3.41
TF-1stream	17.16	3.42	17.56	3.46

Table 2: Ablation study on AVSpeech + AudioSet. Sample-dependent attention is implemented by using self-attention [Vaswani *et al.*, 2017]

4.4 Ablation Study

In the ablation study, all the networks are trained with the same random seed. Adam optimizer with a fixed learning rate of 0.0002 is used and the batch size is 8. Mean SDR and PESQ are used on the test dataset as the evaluation metric.

One-path Attention versus Dual-path Attention

SDAB operates on the T-F domain spectrogram, so the T-F domain framework is used when we study SDAB. Besides, two architectures are used to demonstrate the effectiveness of SDAB including TF-2stream and TF-1stream. Here, TF-2stream represents PHASEN without FTB [Yin *et al.*, 2019]. TF-1stream represents our TFT-Net without decoder and six learning blocks are used. Training target and loss function are the same as PHASEN.

Table 1 aims to demonstrate the effectiveness of dual-path attention. Based on TF-2stream, when dual-path attention is applied, 0.57dB and 0.05 gain on SDR and PESQ respectively are observed, compared to one-path attention only along frequency by using FTB in the original PHASEN. Based on TF-1stream, significant gains on SDR and PESQ are observed when dual-path attention is applied. This large gain indicates the necessity of attention along both axes.

Sample-dependent versus Sample-independent

Next, we explore the effectiveness of sample-independent dual-path attention. Sample-dependent attention is implemented by using the self-attention mechanism [Vaswani *et al.*, 2017]. The results are shown in Table 2. Based on the TF-2stream architecture, we find that sample-independent attention achieves almost the same results as sample-dependent attention in terms of SDR and PESQ. Based on the TF-1stream architecture, the results of sample-independent dual-path attention are a little better than the sample-dependent one. Therefore, we can conclude that sample-independent attention will not negatively influence the accuracy of the model. On the other hand, the sample-independent attention mechanism within the SDAB is highly computationally efficient. Specifically, given an input feature map with a size of $301 \times 257 \times 96$, the FLOPs of SDAB is 6.28B which is significantly smaller than the DAB's 14.7B FLOPs. Besides, the memory cost of SDAB is also less than DAB, since no big matrix multiplication is needed.

In order to better understand the attention mechanism, we visualize the learned weights of SDABs. The network we used for analysis has six learning blocks, so there are six pairs of attention weights. Fig.5 shows three pairs of them, taken from the first, the fourth, and the sixth SDABs. The attention weights along both frequency and time axis show a local-global-local pattern. In the first SDAB, we observe large weights on the diagonal, showing that more attention is paid to local components. Large weights also appear in other locations, but in a more random way. In the fourth DAB, large weights are scattered around in both frequency and time attention maps. But divergent lines can be observed in the frequency attention map, which are consistent with the harmonic correlations observed across the frequencies. In the last SDAB, weights are more concentrated on the diagonal than in the first SDAB, showing that global correlations have been successfully utilized.

Method	SDR(dB)	PESQ
TF-1stream	17.56	3.46
TFT-Net*	18.14	3.40
TFT-Net	18.40	3.41

Table 3: Ablation study on AVSpeech + AudioSet. TFT-Net* represents replacing learnable decoder layer by ISTFT

# of learning blocks	SDR(dB)	PESQ
3	17.63	3.30
6	18.40	3.41
9	18.54	3.42

Table 4: Ablation study on AVSpeech + AudioSet.

Cross-domain Framework

Table 3 shows the results of training the network with different frameworks. TF-1stream represents training the network with the T-F domain framework. TFT-Net represents training the network with the proposed cross domain framework. Compared to TF-1stream, TFT-Net obtains 0.84dB gain on SDR and a decrease of 0.05 on PESQ which are consistent with our expectation, since logarithmic MSE is used as the training criteria which is equivalent to optimizing the SDR to a certain extent. This verifies the effectiveness of the cross domain framework. We also try replacing the learnable decoder layer by ISTFT (TFT-Net*). When comparing it to TFT-Net, there are 0.26dB and 0.01 loss on SDR and PESQ respectively which also verifies the effectiveness of the proposed framework.

The Number of Learning Blocks

In order to trade-off accuracy and complexity, we also train the model with different number of learning blocks. Table 4 shows the results. It is consistent with our expectation that the larger the model is, the better the result is. Compared to the case of three learning blocks, when the block number is six, we can obtain 0.77dB and 0.11 gain on SDR and PESQ respectively. When compared to six learning blocks, nine learning blocks only obtain slight improvement. Therefore, for a good trade-off between speed and accuracy, we use six learning blocks as our final proposed model to do comparisions to other state-of-the-art works.

4.5 Comparision to State-of-the-Arts

AVSpeech + AudioSet

We compare our method with three other recent methods, Conv-TasNet [Luo and Mesgarani, 2019], "Google" [Ephrat *et al.*, 2018], and PHASEN [Yin *et al.*, 2019]. Our networks are trained with Adam optimizer. Learning rate is 0.0002 and batch size is 8. The results in Table 5 show that our method outperforms all these three methods. Note that we use the same training data as PHASEN which is only a small fraction (100k/2.4M) of that used by "Google". Such superior performance on a large dataset demonstrates that our method can be generalized to various speakers and various kinds of noisy environments.

Voice Bank + DEMAND

In order to fairly compare the proposed method with many other methods, we also train our model on this small but commonly-used dataset. In this experiment, our networks are trained for 40 epochs, with Adam optimizer. Learning rate is 0.0005 and batch size is 8.

Table 6 shows the comparision results. Our TFT-Net achieves the best results on SSNR among all the methods listed. This is consistent with our expectation since our method uses logarithmic MSE as the training criteria which is equivalent to optimizing the SSNR to a certain extent. Compared to time-domain methods like SEGAN [Pascual *et al.*, 2017], Wavenet [Rethage *et al.*, 2018], and DFL [Germain *et al.*, 2018], our method outperforms them on all the five metrics. This demonstrates the advantages of using information of the T-F domain. Compared to T-F domain methods like MMSE-GAN [Soni *et al.*, 2018], PHASEN [Yin *et al.*, 2019], our method does not get best results on the other four metrics. This maybe because these metrics depend on the magnitude spectrogram of speech and TF-domain methods optimize the spectrogram directly while our method uses time-domain signal as supervision. Besides, authors of Conv-TasNet have given very detailed explanation of why time-domain methods get a lower PESQ score but may still excel in (real subjective) MOS evaluation. MDPhD [Kim *et al.*, 2018] performs processing the two domains in a cascaded way and the input signal is separately processed by two networks. Our method outperforms it on all the metrics indicating the superiority of our fully end-to-end learnable system in taking advantages of both domains.

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
TFT-Net(1.5M step, 100k speech)	18.40	3.41

Table 5: System comparison on AVSpeech + AudioSet

Method	SSNR	PESQ	CSIG	CBAK	COVL
Noise	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
PHASEN	10.18	2.99	4.21	3.55	3.62
MDPhD	10.22	2.70	3.85	3.39	3.27
TFT-Net	10.63	2.75	3.93	3.44	3.34

Table 6: System comparison on Voice Bank + DEMAND

5 Conclusion

In this paper, we propose a cross-domain framework TFT-Net which takes advantage of the knowledge we have about spectrogram and avoids some of the drawbacks that T-F-domain methods have been suffering from. We also propose a novel long-range correlations learning module SDAB which is very lightweight and effective. Through experiments on two datasets, we show the superiority of the proposed method over prior arts.

In the near future, we plan to speed up our model and apply it to low-latency applications. Another important direction is to study how different training targets and loss functions will influence the proposed method. Finally, we plan to extend the proposed method to various audio-related tasks such as dereverberation.

References

[Choi *et al.*, 2019] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex u-net. *arXiv:1903.03107*, 2019.

[Ephrat *et al.*, 2018] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv:1804.03619*, 2018.

[Erdogan *et al.*, 2015] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *ICASSP*, pages 708–712, 2015.

[Gemmeke *et al.*, 2017] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780, 2017.

[Germain *et al.*, 2018] Francois G Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. *arXiv:1806.10522*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu and Loizou, 2007] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *TASLP*, 16(1):229–238, 2007.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.

[Kim *et al.*, 2018] Jang-Hyun Kim, Jaejun Yoo, Sanghyuk Chun, Adrian Kim, and Jung-Woo Ha. Multi-domain processing via hybrid denoising networks for speech enhancement. *arXiv:1812.08914*, 2018.

[Kim *et al.*, 2019] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. Transformer with gaussian weighted self-attention for speech enhancement. *arXiv:1910.06762*, 2019.

[Krawczyk and Gerkmann, 2014] Martin Krawczyk and Timo Gerkmann. Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement. *TASLP*, 22(12):1931–1940, 2014.

[Luo and Mesgarani, 2019] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *TASLP*, 27(8), 2019.

[Paliwal *et al.*, 2011] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. The importance of phase in speech enhancement. *speech communication*, 53(4):465–494, 2011.

[Pandey and Wang, 2019] Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP*, pages 6875–6879, 2019.

[Pascual *et al.*, 2017] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv:1703.09452*, 2017.

[Plapous *et al.*, 2005] Cyril Plapous, Claude Marro, and Pascal Scalart. Speech enhancement using harmonic regeneration. In *ICASSP*, volume 1, pages I–157, 2005.

[Rethage *et al.*, 2018] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *ICASSP*, pages 5069–5073, 2018.

[Scalart and FILHO, 1996] Pascal Scalart and Jozue VIEIRA FILHO. Speech enhancement based on a priori signal to noise estimation. In *ICASSP*, volume 2, pages 629–632, 1996.

[Soni *et al.*, 2018] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *ICASSP*, pages 5039–5043, 2018.

[Thiemann *et al.*, 2013] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 035081. ASA, 2013.

[Valentini-Botinhao *et al.*, 2016] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pages 146–152, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Veaux *et al.*, 2013] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *O-COCOSDA/CASLRE*, pages 1–4, 2013.

[Vincent *et al.*, 2006] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *TASLP*, 14(4):1462–1469, 2006.

[Wakabayashi *et al.*, 2018] Yukoh Wakabayashi, Takahiro Fukumori, Masato Nakayama, Takanobu Nishiura, and Yoichi Yamashita. Single-channel speech enhancement with phase reconstruction based on phase distortion averaging. *TASLP*, 26(9):1559–1569, 2018.

[Wang and Chen, 2018] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *TASLP*, 26(10):1702–1726, 2018.

[Williamson *et al.*, 2016] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *TASLP*, 24(3):483–492, 2016.

[Yin *et al.*, 2019] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Phasen: A phase-and-harmonics-aware speech enhancement network. *arXiv:1911.04697*, 2019.