Time-Reversal Enhancement Network With Cross-Domain Information for Noise-Robust Speech Recognition

Fu-An Chao, *National Taiwan Normal University, Taipei, 11677, Taiwan*
Jeih-Weih Hung, *National Chi Nan University, Puli, 54516, Taiwan*
Tommy Sheu, *Delta Electronics, Inc., 11491, Taiwan*
Berlin Chen, *National Taiwan Normal University, Taipei, 11677, Taiwan*

Due to the enormous progress in deep learning, speech enhancement (SE) techniques have shown promising efficacy and play a pivotal role prior to an automatic speech recognition (ASR) system to mitigate the noise effects. In this article, we put forward a novel cross-domain time-reversal enhancement network (CD-TENET). CD-TENET leverages the time-reversed version of a speech signal and two effective features that consider the phase information of a speech signal in the time domain and the frequency domain, respectively, to promote SE performance for noise-robust ASR. Extensive experiments demonstrate that CD-TENET can not only recover the original speech effectively but also improve both SE and ASR performance simultaneously. More surprisingly, the proposed CD-TENET method can offer a marked relative word error rate reduction on test utterances of scenarios contaminated with unseen noises when compared to a strong baseline with the multicondition training setting.

With the significant breakthroughs of deep learning with deep neural networks (DNNs) in the recent past, current automatic speech recognition (ASR) systems have exhibited unprecedented performance and reached human parity. However, when deploying a well-trained ASR system into real-world use cases, the performance might be seriously degraded because of environmental interference such as background noise, reverberation, voices from surrounding speakers (babble noise), and others. To mitigate these deteriorating effects so as to improve the real-world ASR performance, researchers and practitioners have developed a great number of techniques. Among these techniques, speech enhancement (SE) has seen widespread adoption as a preprocessing stage that forms the cornerstone to counteracting environmental interference prior to acoustic modeling.

In general, speech signals are usually characterized by long sequences as well as their complicated hierarchical structure in which the relevant information may be distinct at different granularities (phonemes, syllables, words, etc.). Therefore, it still remains challenging to completely eliminate the noise component in speech signals. As for the noisy speech signals received from a microphone array structure, multichannel approaches usually can behave very well and reduce the majority of noise effects on speech. Contrastively, in most of real-world use cases that lack multiple array microphones, single-channel SE techniques are typically less effective and only yield moderate improvements in speech quality and intelligibility metrics like perceptual evaluation of speech quality (PESQ)[1] and short-time objective intelligibility (STOI).[2] Although there is a great deal of effort that has been demonstrated fruitful, it still remains challenging to improve the SE performance under some severe conditions, such as the low signal-to-noise ratio (SNR) and nonstationary noise scenarios.

1070-986X © 2021 IEEE
Digital Object Identifier 10.1109/MMUL.2021.3139302
Date of publication 31 December 2021; date of current version 4 May 2022.

Since the north star of SE is to recover and achieve the high quality and/or intelligibility of speech, a variety of sophisticated methods have been well-practiced to improve the human auditory perception or optimize the corresponding objective metrics, such as PESQ and STOI. For instance, deep feature loss[3] minimizes the distance between clean speech and its enhanced counterpart in latent spaces using a pretrained audio scene classification network, which gains considerable improvements on both the objective quality metrics and the perceptual evaluation of human listeners. Furthermore, phone-fortified perceptual loss (PFPL)[4] was proposed to extend this idea by adopting a pretrained self-supervised model. Beyond minimizing the above perceptual losses for SE, another option is to optimize the target metrics directly by utilizing the generative adversarial network (GAN). Notably, MetricGAN[5] has been demonstrated to achieve effective SE performance by connecting the PESQ metric straightly to the discriminator.

Most advanced SE techniques that optimize the perceptual quality usually can yield improvements in terms of PESQ metrics. However, this does not necessarily transfer well into the downstream ASR performance, especially when an ASR system is trained with a great amount of speech data contaminated by various types of noise (i.e., the so-called multicondition training setting). This may be due to the artifacts introduced by the front-end SE module and the discrepancy of the training objectives between SE and ASR. To tackle this problem, several efforts have been made to develop robust single-channel SE methods that could better benefit the ASR performance. For example, the joint training scheme[6] is established along this direction.

In light of the above-mentioned observations, the main contributions of this article can be summarized as follows.

1) We design a bi-projection fusion (BPF) mechanism to formulate two novel cross-domain modeling frameworks, CD-TCN[7] and CD-DPTNet,[8] for robust single-channel SE, which leverage both time- and frequency-domain features and dramatically reduce undesired noise effects on speech for superior SE and ASR performance.
2) In further conjunction with our recently proposed approach, namely the time-reversal enhancement network (TENET),[9] our framework can obtain remarkably better results on the VoiceBank-DEMAND benchmark dataset in relation to some top-of-the-line methods. Particularly, the framework behaves quite well in the test set of scenarios contaminated with low SNRs and nonstationary noise sources.
3) Through our experimental analysis, we demonstrate that the time-domain features, i.e., wavegrams, are probably the key factor to promote the performance of SE and robust ASR.

RELATED WORK

SE Algorithms

Considering a discrete-time noisy signal [*n*] captured by a single-channel microphone, we can formulate the following equation:

$$y\,[n] = \; h[n]{*}x[n] + d[n] \tag{1}$$

where *x*[*n*] is the target noise-free speech signal, *h*[*n*] is the convolutional noise, "$*$" denotes the convolution operation, *d*[*n*] is the additive noise, and *n* is the time index. In this study, convolutional noise *h*[*n*] like channel mismatch or reverberation is excluded from consideration and assumed to be negligible. We solely focus on the removal of additive noise [*n*] to recover the speech signal *x*[*n*] from the noisy signal *y*[*n*].

Most of the early studies on noise reduction analyze noisy speech signals in the acoustic frequency domain of signals via short-time Fourier transform (STFT). Such techniques employ the statistics drawn from the short-time spectra of speech to suppress noise. Notably, the prevalent DNN approaches have been adopted to develop SE techniques with overwhelming success in recent years. These DNN-based SE methods can be further divided into two broad categories, i.e., mapping-based and masking-based. For instance, Lu *et al.*[10] is the first to employ a deep denoising autoencoder (DDAE) to map the power spectrum of noisy speech to that of its clean counterpart directly. Wang *et al.*[11] proposed to use a DNN-based model to implicitly predict a time-frequency (T-F) mask that applies to a noisy spectrogram for SE. In particular, pursuing an effective mask for separating the speech and noise feature representations has become one of the most predominant directions for the SE research. The primary DNN-enabled masking algorithms, including, but is not limited to, ideal binary mask (IBM) and ideal ratio mask (IRM),[11] are mostly conducted on the magnitude spectrogram of speech.

In the following sub-section, we briefly introduce the concept of the masking algorithm, which will serve as the main component module of our SE framework in the next section.

Masking-Based SE

A masking-based SE is usually composed of an encoder-decoder architecture and a mask estimation network to separate the speech and noise in the feature space. Firstly, the encoder is designed to transform the speech signal into a series of *N*-dimensional representations, which can act as either a conventional STFT operated in the frequency domain or a trainable 1-D convolution operation learned through the network in the time domain. The whole encoding operation can be formulated as follows:

$$\mathbf{W} = \mathcal{F}(\mathbf{U}\mathbf{X}) \tag{2}$$

where $\mathbf{X}$ is a matrix representing K overlapped segments of length L with respect to an input noisy signal $y[n]$, $\mathbf{U} \in \mathbb{R}^{N\times L}$ is a linear transformation which contains *N* basis functions, $\mathcal{F}(\cdot)$ is an optional activation function, and $\mathbf{W} \in \mathbb{R}^{N\times K}$ is the feature representation for subsequent mask prediction.

To effectively capture the temporal information and consider the long-term dependency of frames in the noisy signal, the mask estimation network can be created by stacking BLSTM or dilated convolution layers such as the temporal convolution network (TCN).[7] However, current state-of-the-art models use the dual-path modeling technique called DPTNet,[8] which significantly improves the accuracy of the mask prediction. In addition, the output of the mask estimation network can directly be a single mask for the target speech, and the other choice is to predict two masks respectively for speech and noise, denoted by

$$[\mathbf{M}_x, \mathbf{M}_d] = \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{W}) \tag{3}$$

where $\mathcal{M}_{\boldsymbol{\theta}}(\cdot)$ represents the mask estimation network, and $\mathbf{M}_x$ and $\mathbf{M}_d$ are the masks for the target speech and noise, respectively.

After that, we can obtain the enhanced representation $\mathbf{D}_{\boldsymbol{x}}$ by applying the speech mask $\mathbf{M}_x$ to the feature representation $\mathbf{W}$, which essentially is a mixture of the target speech and noise:

$$\mathbf{D}_{\boldsymbol{x}} = \mathbf{M}_x \odot \mathbf{W} \tag{4}$$

where $\odot$ is the element-wise multiplication. The enhanced waveform is obtained from $\mathbf{D}_{\boldsymbol{x}}$ using a decoder, which can be formulated as another matrix multiplication:

$$\hat{\mathbf{S}} = \mathbf{V}\mathbf{D}_{\boldsymbol{x}} \tag{5}$$

where $\hat{\mathbf{S}} \in \mathbb{R}^{L\times K}$ contains the reconstructed speech segments with respect to $\mathbf{X}$, and $\mathbf{V} \in \mathbb{R}^{L\times N}$ consists of the basis functions involved in the decoder. For frequency-domain SE, the decoder is usually the inverse STFT (iSTFT), while it can be a 1-D transpose convolution operation for time-domain SE. Finally, we can simply use the overlap-add method to derive the enhanced waveform from $\hat{\mathbf{S}}$.

Speech Feature Representations

To extract the rich and meaningful speech features for masking-based SE, conventional approaches often introduce STFT along with some bits of prior knowledge to obtain the frame-wise features. In this way, we can reduce the computational cost as well as carry out the T-F analysis. To list a few, hand-crafted features such as magnitude spectrograms and logarithmic mel-scaled spectrograms have strong visualization capability that helps to discriminate between speech and noise. Since many researches have shown that the phase information is crucial to the success of SE, complex-valued spectrograms have been well-studied in the past few years.[12,13] To take the phase into account, we can either extract the magnitude and phase parts simultaneously in the (short-time) frequency domain of speech signals, or operate speech signals in time domain directly. Both directions have been proven effective and surpass previous attempts that process only magnitude spectrograms.[7,13] Thanks to the remarkable advances of deep learning, many researches explore the possibilities of extracting features from the raw waveform domain for various tasks via a DNN. The majority of them exploit variants of convolutional neural network (CNN) for feature extraction, and it is believed that the first convolution layer is the most crucial part in waveform-based CNNs.[14] For example, Conv-TasNet[7] proposes to use 1-D convolution layer instead of STFT to extract time-domain features for speech separation. Furthermore, SincNet adopts the parameterized sinc functions into a CNN architecture to simulate the band-pass filter for both speaker and speech recognition.[14] Note here that, the above-mentioned methods all achieve significant results compared to the frequency-domain methods. In particular, such features extracted by a 1-D convolution layer are also employed in the state-of-the-art audio pattern recognition system, PANNs,[15] and they are called "wavegrams" instead.

In contrast to spectrograms, the wavegrams derived from the DNN-based architecture are more like data-driven representations of speech. When only few training samples are available, it might fail to perform well, probably leading to incorrect estimation of the distribution for speech features.[13] Consequently, how to take advantages of both time-domain and

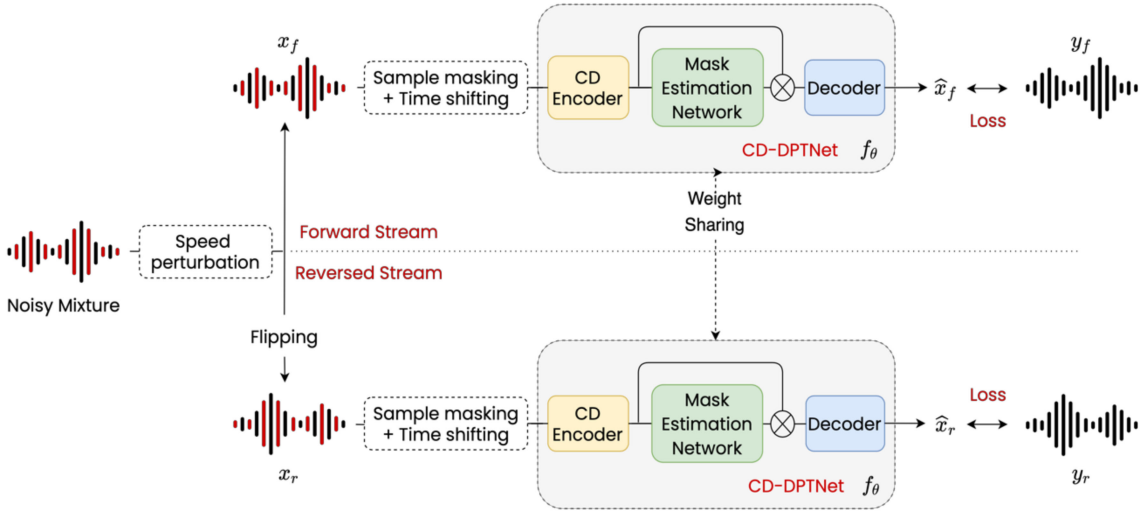

FIGURE 1. A schematic diagram of the CD-TENET framework.

frequency-domain features for DNN-based algorithms to benefit SE is still worth further investigation.

METHODOLOGY

In this section, we introduce a novel cross-domain modeling framework, which consists of a cross-domain (CD) encoder, a mask estimation network, and a decoder. This novel SE framework can harness both time- and frequency-domain features of speech signals to improve SE performance in a synergetic way. Due to the versatility of this framework, our proposed method has a few variants that can be integrated with other approaches. Specifically, we consider three types of cross-domain models in this work. When in conjunction with a TENET, as illustrated in Figure 1, we refer it to as CD-TENET. In the following, we describe the ingredient components of the proposed framework.

Cross-Domain Modeling Framework

As depicted in Figure 2, the input noisy speech signal is converted to time-domain and frequency-domain feature representations, "wavegrams" and "spectrograms," respectively, and both of them retain the phase information of the input signal. The wavegrams are generated from a trainable 1-D convolution layer, while the spectrograms derived from the conventional STFT with discrete Fourier basis functions are non-trainable. Note here that, the "spectrograms" used in this work have real and imaginary parts that are structured as in the article by Koyama *et al.*[13] These two-domain feature representations are calculated concurrently and then spliced together to perform the subsequent mask estimation.

To further extract the shared information across the two distinct features, we have developed a novel fusion scheme called the BPF module in previous work.[16] Conceptually similar to the recent attempt of the computer vision community, the BPF module is designed to learn the relation between the input representations. The corresponding work flow is operationalized as shown in Figure 3, whose details are illustrated as follows.

Given the two branches of feature representation, namely the wavegrams F_c and spectrograms F_s, we first transform them into two new features individually, $F_{\bar{c}}$ and $F_{\bar{s}}$, both of which have the same dimension. Next, in order to exploit both branches and balance the fusion procedure, we use the concatenation of $F_{\bar{c}}$ and $F_{\bar{s}}$ to estimate a ratio mask M. Finally, we apply this mask to both branches and in turn generate the BPF feature. The whole procedure can be formulated as

$$F_{\bar{c}} = \Psi_c\ (F_c, \theta_c) \tag{6}$$
$$F_{\bar{s}} = \Psi_s\ (F_s,\ \theta_s) \tag{7}$$
$$M\ =\ \text{sigmoid}(\Psi_M(\text{concat}(F_{\bar{c}},\ F_{\bar{s}}),\ \theta_M)) \tag{8}$$
$$F_{BPF} =\ M \odot F_{\bar{c}} + (1 - M) \odot F_{\bar{s}} \tag{9}$$

where Ψ_c, Ψ_s, and Ψ_M are projection layer operations with parameters θ_c, θ_s, and θ_M, respectively, M is the estimated ratio mask with the same dimension as $F_{\bar{c}}$ and $F_{\bar{s}}$, and F_{BPF} is the output feature of BPF.

Finally, the obtained BPF feature serves as an auxiliary input concatenated with the wavegrams F_c and

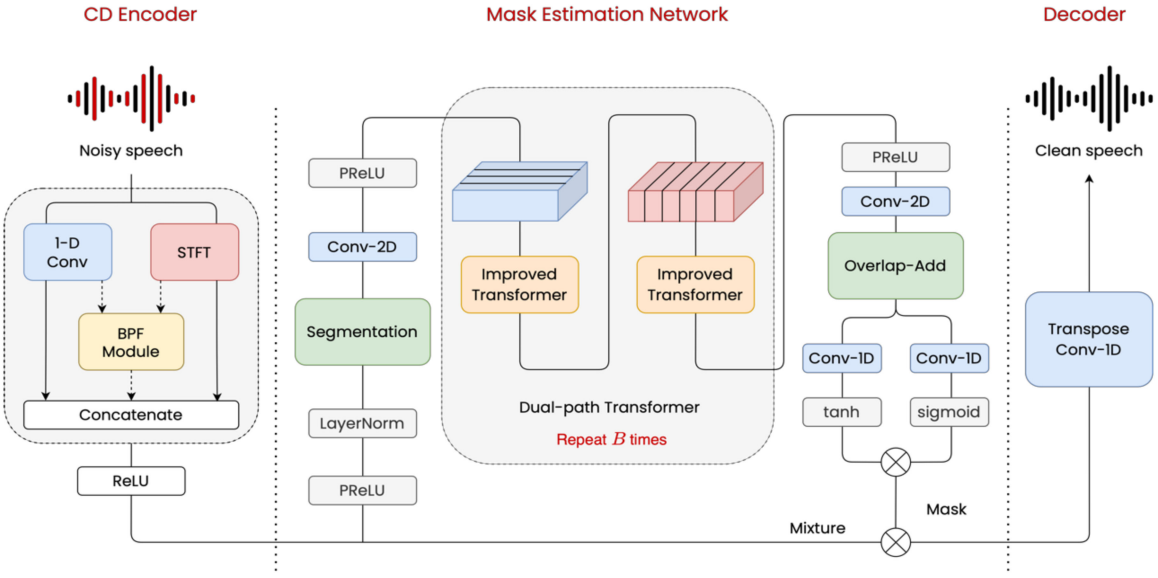

FIGURE 2. Schematic diagram of the CD-DPTNet system.

spectrograms F_s to be passed to the mask estimation network, which can be TCN[7] or DPTNet.[8] For an illustration in Figure 2, we use DPTNet to form the CD-DPTNet system that integrates cross-domain clues of a speech signal, as well as their BPF features, for effective SE.

Time-Reversal Enhancement Network

Apart from the features in the time domain and the frequency domain, we seek to discover other characteristics of the speech signal to boost the SE performance. In our recent work, a TENET,[9] we are the first to propose to utilize the reversed speech on SE, and it has been proven that the reversed speech possess common temporal characteristics as the original (forward) speech due to their spectrograms share the same autocorrelation function. Thus, the original and reversed speech signals might be additive to each other in the learning of an SE model.

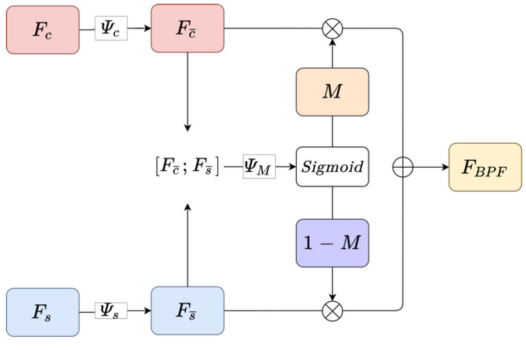

FIGURE 3. Our proposed BPF module.

As an illustration in Figure 1, TENET consists of a time-reversal Siamese network in tandem with a set of data augmentation techniques. The time-reversal Siamese network is equipped with Siamese SE models and forms a two-stream architecture, namely, the forward stream and reversed stream as two different inputs. To be more specific, the forward stream receives the noisy signal $x_f[n]$, which is converted to approximate its clean counterpart $y_f[n]$ as normal, while the reversed stream takes the reversed version of $x_f[n]$ as the input, viz. $x_r\ [n] = x_f\ [L-n]$, to approximate its clean duplicate $y_r\ [n] = y_f\ [L-n]$, where L is the length of $x_f[n]$.

During the training phase, three waveform-based data augmentation schemes are employed: *speech perturbation*, *time shifting,* and *sample masking*, which are all conducted on-the-fly to help the Siamese network to learn the more powerful features. In the following, we briefly introduce these constituent components.

1) Speech perturbation is a simple yet effective technique broadly used for acoustic modeling. It perturbs the speed of the input signal by a given factor f_{sp} and it modifies the pitch of the speech. In this work, f_{sp} is randomly selected from 0.95 to 1.05.
2) Time shifting serves as a straightforward data augmentation approach that shifts the audio samples to the left or right with f_{shift} seconds. In our experiments, we uniformly choose f_{shift} from

0 to 0.625, and we always apply right shift to the audio.

3) Sample masking zeros out a portion of the audio samples, making the masked speech segments silent. Thereby, it encourages the model to predict the clean waveform by considering the context information. There are two hyperparameters in sample masking: the length of each mask, denoted by t, and the maximum number of masks, denoted by m_s. We set t to a fixed value 10, and choose m_s from a uniform distribution in the interval [0, 150].

At inference time, these data augmentation techniques are detached, and we can simply use one of the streams (SE models) for inference without any additional cost.

Loss Function

When adopting an SE technique as the front-end processing, the performance of the downstream ASR is usually degraded. An underlying reason is that the denoising algorithms in SE usually introduce artifacts to the processed utterance. Even though these artifacts do not have much impact on the human perception, they inevitably distort the original speech structure and cause the speech feature mismatch in the subsequent ASR system. To minimize the artifact distortion between the clean speech $\mathbf{s}$ and enhanced speech $\hat{\mathbf{s}}$, we use the negative scale-invariant signal-to-distortion ratio (SI-SDR)[17] as a loss function, which can be formulated as follows:

$$\mathbf{s}_{\text{target}} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2}, \tag{10}$$

$$\mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\text{target}} \tag{11}$$

$$\mathcal{L}_{\text{SI-SDR}}(\mathbf{s}, \hat{\mathbf{s}}) = -10 \cdot \log_{10}\left(\frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2}\right). \tag{12}$$

On the other hand, to consider the perceptual quality for optimization, PFPL[4] is adopted as an auxiliary loss to form a hybrid loss in TENET to minimize the perceptual distance:

$$\mathcal{L}_{\text{PFPL}}(\mathbf{s}, \hat{\mathbf{s}}) = \mathbb{E}_{\mathbf{s},\hat{\mathbf{s}} \sim \mathcal{D}} \left[\|\Phi_{wav2vec}(\hat{\mathbf{s}}) - \Phi_{wav2vec}(\mathbf{s})\|_1 \right] \tag{13}$$

$$\mathcal{L}_{\text{hybrid}}(\mathbf{s}, \hat{\mathbf{s}}) = \mathcal{L}_{\text{SI-SDR}}(\mathbf{s}, \hat{\mathbf{s}}) + \alpha \cdot \mathcal{L}_{\text{PFPL}}(\mathbf{s}, \hat{\mathbf{s}}) \tag{14}$$

where $\mathbf{s}$ and $\hat{\mathbf{s}}$ are sampled from the training dataset $\mathcal{D}$, $\Phi_{wav2vec}$ is the pretrained $wav2vec$ encoder,[18] and α is a tunable weight parameter. Finally, a weighted sum of the hybrid losses for the reversed stream and the forward stream is used to train the component models of the framework

$$\mathcal{L}_{\text{total}}(\mathbf{s}, \hat{\mathbf{s}}) = \beta \cdot \mathcal{L}_{\text{hybrid}}^{\text{forward}}(\mathbf{s}, \hat{\mathbf{s}}) + \gamma \cdot \mathcal{L}_{\text{hybrid}}^{\text{reversed}}(\mathbf{s}, \hat{\mathbf{s}}) \tag{15}$$

where β and γ are tunable weight parameters.

EXPERIMENTS

Corpora

To provide solid evidence of the effectiveness for our proposed methods, we conduct an extensive set of empirical experiments on the VoiceBank-DEMAND[19] dataset, which is a widely-adopted, open-source benchmark corpus for SE. In the training set, 11,572 utterances (from 28 speakers) are presynthesized with 10 types of noise from the DEMAND database at four different SNR values: 0, 5, 10, and 15 dB, while the test set contains 824 utterances (from 2 speakers) contaminated by five types of noise at SNR values of 2.5, 7.5, 12.5, and 17.5 dB. As to the training of the component models of various SE systems, around 200 utterances are set aside for validation, while the others are for training. In the following ASR experiments, we adopt two configurations of acoustic models for comparison, i.e., CCT-AM and MCT-AM. The CCT-AM is trained with all of the *clean-condition* utterances in VoiceBank except the test ones. In contrast, MCT-AM is trained with *noise-corrupted multicondition* data, which are synthesized by adding the clean recordings with 13 types of noise sources (distinct from the test set) from DEMAND with SNR values ranging from 0 to 15 dB. To examine the generalization ability, we also create another test set apart from the original one. Specifically, the noisy data is generated by mixing the noise sources complied from the QUT-NOISE[20] dataset, which contains nonstationary noise at SNR values of -5, 0, 5, and 15 dB, to simulate a more severe test scenario. It is referred to as the VoiceBank-QUT-NOISE test set in the following experiments.

Note that all of the speech signals used for the experiments are resampled to 16-kHz, and the back-end ASR systems, CCT-AM and MCT-AM, are both built on hybrid DNN-HMM acoustic models, which are the factorized time-delay neural networks trained with lattice-free MMI objective function using the Kaldi toolkit.

SE System Configuration

In the settings of SE experiments, we choose three types of models, TCN,[7] DPTNet,[8] and TENET,[9] as the backbones to form the corresponding cross-domain

TABLE 1. PESQ, SI-SDR (dB), WER (%) results of different cross-domain models on voicebank-demand. "f," "t," and "c" denote frequency, time, and cross domains, respectively.

SE Model	Domain	Encoder / Decoder	Mask Estimation Network	VoiceBank-DEMAND		
				PESQ	SI-SDR	WER
No process (CCT-AM)	–	–	–	1.97	8.45	23.76
No process (MCT-AM)	–	–	–	1.97	8.45	8.31
STFT-TCN[13]	F	STFT / ISTFT	TCN	2.48	18.33	8.24
ConvTasNet[7]	T	1-D Conv / 1-D Transpose Conv		2.52	19.29	7.44
CD-TCN[16]	C	CD Encoder / 1-D Transpose Conv		**2.63**	**19.38**	**7.07**
CDPT[9]	F	STFT / ISTFT	DPTNet	3.01	19.06	6.99
DPTNet[8]	T	1-D Conv / 1-D Transpose Conv		2.78	19.34	6.87
CD-DPTNet	C	CD Encoder / 1-D Transpose Conv		**3.01**	**20.00**	**6.81**

modeling frameworks, following their respective best hyperparameters settings for training. In the CD encoder, we take both time-domain (wavegrams) and frequency-domain (spectrograms) features with 256 dimensions, and the window size is set to 16 with 1/2 overlapping. The number of hidden units in the projection layer of the BPF module is set to 128 to obtain 128-dimensional BPF features.

Furthermore, in order to explore the effectiveness of our proposed method, we take several top-of-the-line SE models for comparison: STFT-TCN,[13] Conv-TasNet,[7] DPTNet,[8] DCCRN,[12] PFPL,[4] MetricGAN+,[5] CDPT,[9] and TENET.[9] As for evaluation, PESQ (short for Perceptual Evaluation of Speech Quality; wide-band version)[1] and SI-SDR[17] are adopted as the front-end SE metrics to measure the speech quality and artifact distortion, respectively. On the other hand, we use the word error rate (WER) metric to evaluate the back-end ASR performance and meanwhile represent the speech intelligibility instead of using STOI.[2]

Preliminary Experiment

At the outset, we compare our approach to other methods using different encoder–decoder architectures with either one of two mask estimation networks, viz. TCN and DPTNet, respectively. As presented in Table 1, we can markedly observe that the further utilization of the MCT data enables the resulting MCT-AM to obtain considerable WER reductions in comparison to CCT-AM. Based on MCT-AM, we further build different SE systems and look into their performance. First, it is worth mentioning that all of the methods compared here promote both SE and ASR metrics, especially those methods that employ DPTNet as the mask estimation network. Second, in contrast to frequency-domain approaches, the time-domain approaches can obtain higher SI-SDR scores and deliver more improvements on WER, while losing some quality metrics, i.e., PESQ, in the DPTNet case.

In particular, when adopting our proposed method, models that integrate the cross-domain information can lead to the optimal performance in terms of PESQ, SI-SDR, and WER in both cases (CD-TCN and CD-DPTNet), which also reveals that this framework can be effective for leveraging the advantages of both domain features.

For clarity, we choose the best-behaved model, CD-DPTNet, for comparison in subsequent experiments. Moreover, CD-DPTNet is further employed as the Siamese SE models being integrated with the TENET approach, which in turn create a more robust SE framework referred to as CD-TENET for the following sections.

Comparison With State-of-the-Art Systems

Table 2 lists the experimental results of the presented CD-TENET and a series of top-of-the-line systems. As shown in this table, the proposed CD-TENET, which leverages the time-reversed speech and a set of data augmentations, achieves considerably higher PESQ and ASR performance than CD-DPTNet and the other systems, especially in the case of VoiceBank-QUT-NOISE, where the test set is contaminated with low SNR and nonstationary noise sources.

In contrast to TENET, CD-TENET that introduces the cross-domain information as well as their BPF features also delivers substantial improvements on VoiceBank-QUT-NOISE, which means that this cross-domain modeling framework and the time-reversal approach are additive to each other to exhibit superior

TABLE 2. Results of different se models on voicebank-demand and voicebank-qut-noise.

Acoustic model	SE model	VoiceBank-DEMAND			VoiceBank-QUT-NOISE		
		PESQ	SI-SDR	WER	PESQ	SI-SDR	WER
CCT-AM	No process	1.97	8.45	23.76	1.25	3.88	82.32
MCT-AM	No process	1.97	8.45	8.31	1.25	3.88	38.87
	DCCRN[12]	2.77	18.94	7.31	1.77	11.54	27.56
	PFPL[4]	3.11	17.28	12.78	2.00	9.93	38.67
	MetricGAN+[5]	3.15	8.52	8.32	2.28	2.83	43.63
	TENET[9]	**3.15**	19.12	**6.76**	2.12	12.13	26.50
	CD-DPTNet	3.01	**20.00**	6.81	2.13	13.65	25.17
	CD-TENET	3.12	19.68	6.78	**2.20**	**14.37**	**22.30**

capabilities on noise suppression in unseen noise scenarios. Although CD-TENET gets a slightly worse PESQ score than TENET on VoiceBank-DEMAND test set, the degradation has little effects on the ASR performance, and CD-TENET still outperforms other methods in overall performance. More surprisingly, when adopting CD-TENET as front-end SE prior to the strong baseline, i.e., MCT-AM, we can obtain a significant relative WER reduction of 43% on VoiceBank-QUT-NOISE.

Analysis on BPF Module

In order to have a more comprehensive understanding of the proposed cross-domain modeling framework, we conduct a visual analysis on the BPF module here. Due to that the design of the BPF module is to predict two ratio masks for balancing both branches of feature representation, as shown in (8), we take the weight matrices M and $(1-M)$ on average when inferencing the whole testing set of 824 utterances and investigating the distribution of them in the fusion process. Note that we crop the average masks to 128 frames for simplicity. As depicted in Figure 4, we can see that the BPF features tend to put on more weights to the time-domain features (wavegrams) than the frequency-domain features (spectrograms), we thus deduce that the time-domain features are probably the key component of both SE and ASR. Furthermore, this deduction can also be supported by the results in Table 1. When the backbone of the masking-based SE models is the same, the time-domain models outperform the frequency-domain models. Taking STFT-TCN and Conv-TasNet for instance, both of them are based on the TCN architecture but with

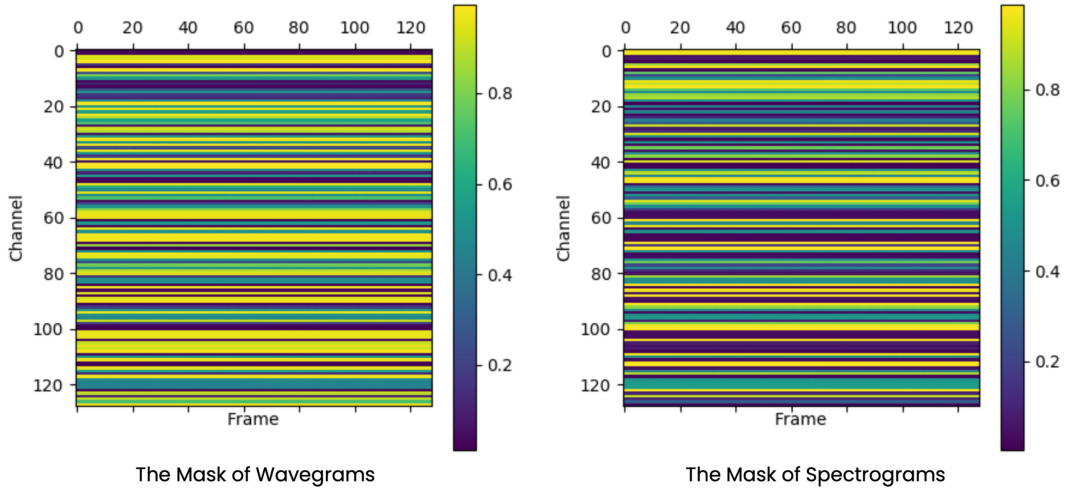

FIGURE 4. Visual Analysis on BPF module.

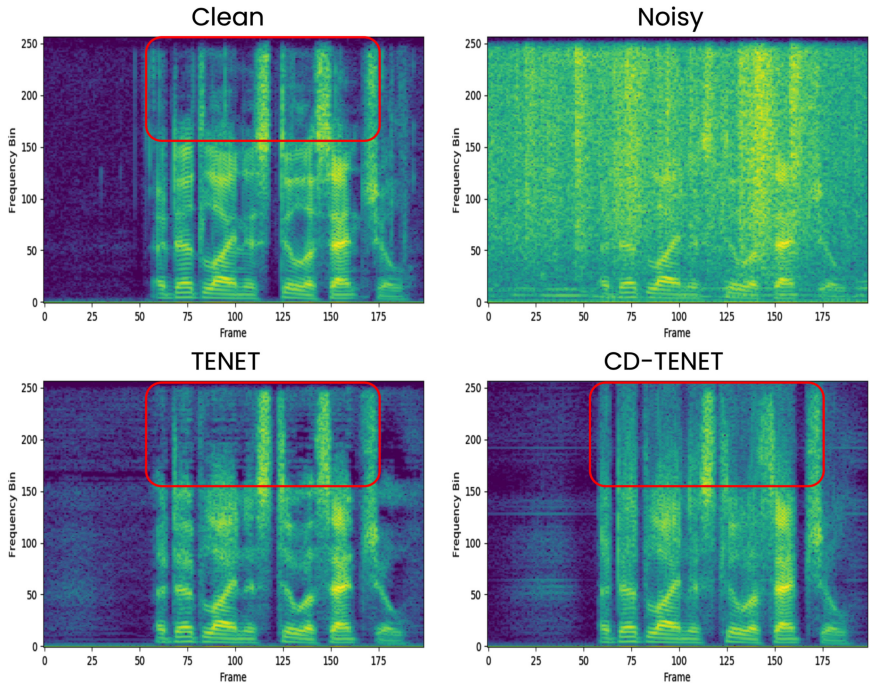

FIGURE 5. Enhanced spectrograms of a noisy speech utterance (p232_203, 2.5dB) from the VoiceBank-DEMAND test set.

different encoder-decoder architectures, and Conv-TasNet can achieve better performance than STFT-TCN in terms of both SI-SDR and WER.

Visualization of Spectrograms

To provide more conclusive evidence on the effectiveness of our proposed method, here we plot the

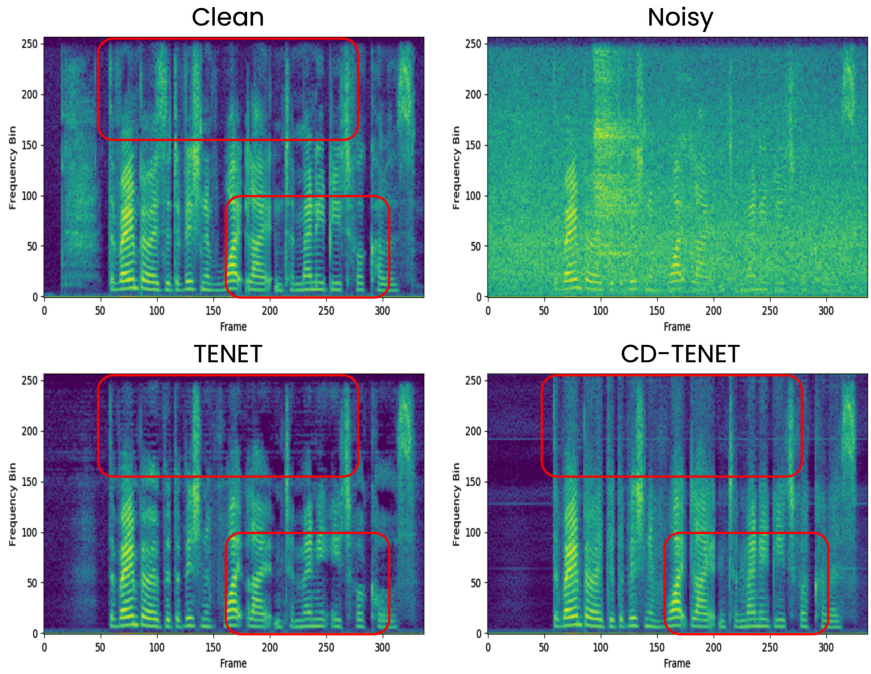

FIGURE 6. Enhanced spectrograms of a noisy speech utterance (p257_007, 0dB) from the VoiceBank-QUT-NOISE test set.

spectrograms of two noisy utterances from VoiceBank-DEMAND and VoiceBank-QUT-NOISE, respectively, along with their clean and enhanced duplicates by using TENET and CD-TENET. As shown in Figures 5 and 6, both TENET and CD-TENET remarkably eliminate the interrupting noise. More impressively, as the area outlined in the red box, CD-TENET that considers the cross-domain information seems have the inpainting capability that can fix up the missing patches on the spectrograms. In Figure 5 (i.e., the VoiceBank-DEMAND case), this inpainting capability mostly recovers the speech in high-frequency bands, which are less significant to speech recognition. Therefore, it only causes a slight decline in WER (*cf.* Table 2). On the other hand, in Figure 6, these inpainted patches exist in both high and low-frequency bands, which helps not only recover the clean speech more completely but also improves the ASR performance greatly (*cf.* Table 2). Moreover, it is observed that this phenomenon mostly occurs in low SNR and non-stationary noise scenarios (i.e., the VoiceBank-QUT-NOISE case). We refer interested readers to https://fuann.github.io/CD-TENET for more exemplars.

CONCLUSION

In this article, we have proposed a novel cross-domain SE model, CD-TENET, which integrates the cross-domain information and the time-reversal SE technique to construct a more robust SE architecture for noise-robust ASR. Compared with the state-of-the-art SE systems, CD-TENET has demonstrated to obtain considerable improvements not only on the front-end SE metrics (PESQ and SI-SDR) but also the back-end ASR (in terms of WER), especially in the test set of scenarios contaminated with low SNRs and nonstationary noise sources. Experimental analysis on the proposed BPF module also suggests that the time-domain features are crucial to elevate the performance of SE and ASR. In the future, we envisage that the proposed framework can offer a promising avenue for SE. Also, we believe that it can be further applied to other developments, such as the audio-visual task, to learn the shared information across image and speech data.

ACKNOWLEDGMENTS

This work was supported in part by Ministry of Science and Technology, Taiwan, under Grant MOST 110-2634-F-008-004- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan, and Grants MOST 108-2221-E-003-005-MY3 and MOST 109-2221-E-003-020-MY3. Any findings and implications in the paper do not necessarily reflect those of the sponsors.

REFERENCES

1. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)- a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.
2. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
3. F. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech*, 2019, pp. 2723–2727.
4. T. Hsieh, C. Yu, S. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using Wasserstein distance for speech enhancement," in *Proc. Interspeech*, 2020, pp. 196–200.
5. S.-W. Fu *et al.*, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.
6. T. Menne, R. Schlüter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2019.
7. Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
8. J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020, pp. 2642–2646.
9. F. -A. Chao, S. -W. Fan Jiang, B. -C. Yan, J. -W. Hung, and B. Chen, "A time-reversal enhancement network for noise-robust ASR," in *Proc. Autom. Speech Recognit. Understanding*, 2021, pp. 55–61.
10. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.
11. Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1381–1390, Jul. 2013.
12. Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.
13. Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks," 2020, *arXiv:2005.11611*.

14. M. Ravanelli and Y. Bengio, "Speech and speaker recognition from raw waveform with sincnet," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018.
15. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, Oct. 2020.
16. F.-A. Chao, J.-W. Hung, and B. Chen, "Cross-domain single-channel speech enhancement model with bi-projection fusion module for noise-robust ASR," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.
17. J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.
18. S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.
19. C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 146–152.
20. D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The qut-noise-timit corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, 2010, pp. 3110–3113.

FU-AN CHAO received the M.S. degree in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2021. His research interests include deep learning, machine learning, and its applications, such as audio and speech processing, speech enhancement, and robust automatic speech recognition. Contact him at fuann@ntnu.edu.tw.

JEIH-WEIH HUNG is currently a Professor with the Department of Electrical Engineering, National Chi Nan University, Puli, Taiwan. His research interests include robust speech recognition, speech enhancement, and machine learning. Hung received the B.S, M.S., and Ph.D. degrees in electrical engineering from the National Taiwan University, in 1994, 1996, and 2001, respectively. Contact him at jwhung@ncnu.edu.tw.

TOMMY SHEU is a Technical Section Manager with the Delta Management System (DMS) Department, Delta Electronics, Inc., Taipei, Taiwan. His research interests include robust speech recognition, live meeting subtitles and speech analytics. Sheu received the B.S and M.S. degrees in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1994 and 1996, respectively. Contact him at tommy.sheu@deltaww.com.

BERLIN CHEN is currently a Professor with the Computer Science and Information Engineering Department, National Taiwan Normal University, Taipei, Taiwan, and the President of the Association for Computational Linguistics and Chinese Language Processing. His research interests include speech recognition and natural language processing. Chen received the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001. He is the corresponding author of this article. Contact him at berlin@ntnu.edu.tw.