Joint time-domain and frequency-domain progressive learning for single-channel speech enhancement and recognition

Gongzhen Zou[1], Jun Du[1], Shutong Niu[1], Hang Chen[1], Yuling Ren[2], Qinglong Li[2], Ruibo Liu[2], and Chin-Hui Lee[3]

[1] University of Science and Technology of China, China
[2] China Mobile Online Services Company Limited, China
[3] Georgia Institute of Technology, USA
jundu@ustc.edu.cn

Abstract. Single-channel speech enhancement for automatic speech recognition (ASR) has been extensively researched. Traditional methods usually directly learn clean target, which may introduce speech distortions and limit ASR performance. Meanwhile, these methods usually focus on either the time or frequency domain, ignoring their potential connections. To tackle these problems, we propose a joint time and frequency domain progressive learning (TFDPL) method for speech enhancement and recognition. TFDPL leverages information from both domains to estimate frequency masks and waveforms, gradually predicting less-noisy and cleaner targets. Experimental results show that TFDPL outperforms traditional methods in ASR and perceptual metrics. TFDPL achieves relative reductions of 43.83% and 36.03% in word error rate for its intermediate outputs on the CHiME-4 real test set using two different acoustic models and certain improvements in PESQ and STOI metrics for clean output on the simulated test set.

Keywords: automatic speech recognition · speech enhancement · joint time and frequency domain · progressive learning.

1 Introduction

With the advancement of deep learning, automatic speech recognition (ASR) [1] has made significant progress and has been widely applied in our daily lives [2]. However, in complex acoustic environments, speech may be interfered with by various sources of noise, leading to degradation in ASR performance.

Speech enhancement (SE) is a critical technology in speech processing that aims to improve the quality and intelligibility of corrupted speech [3]. Moreover, it can be utilized as a front-end system to enhance the robustness of ASR systems [4]. In recent years, supervised speech enhancement techniques based on deep neural networks have been widely studied and established as the mainstream approach [5]. These methods can be categorized into two classes based on the domain: frequency-domain methods and time-domain methods.

Frequency-domain methods often utilize the Short-Time Fourier Transform (STFT) to convert the original waveform into a time-frequency spectrogram, which serves as the input to the neural network. These methods aim to predict frequency-domain masks or features, such as Ideal Ratio Mask (IRM) [6] or log-power spectra (LPS) [7]. However, prediction of these methods often discards the clean phase information, which can be detrimental for speech recognition [8]. Additionally, certain prediction targets like IRM are limited by their own assumptions and cannot perfectly reconstruct the clean speech, thus limiting the performance of frequency-domain methods. In recent years, there has been increasing interest in time-domain-based speech enhancement methods [9–11]. These methods directly process the raw waveform to overcome challenges associated with phase estimation and have a higher theoretical performance ceiling.

Most speech enhancement methods aim to improve the quality and intelligibility of corrupted speech. During training, these methods often utilize targets such as IRM, clean LPS, or clean waveform. However, using these targets can sometimes result in excessive suppression and distortion, which may have a negative impact on ASR performance [12]. To address this issue, researchers have proposed methods to mitigate over-suppression and improve ASR performance. [12] introduced an asymmetric loss function to improve speech preservation, while [13] introduced a progressive learning-based speech enhancement network that gradually improves the Signal-to-Noise Ratio (SNR) until learning clean spectral features, where the intermediate target can effectively preserve speech information. [14] and [15] proposed progressive learning (PL) methods in the frequency and time domains, respectively, and demonstrated the effectiveness of intermediate targets in improving ASR performance. In [16], the authors demonstrated that all audio-only object-oriented progressive learning (AOPL) models outperform their audio-only counterparts (AODL) in speech enhancement. These findings highlight the advantages of progressive learning methods in ASR back-end and speech perceptual quality. On the other hand, previous research in speech enhancement has primarily focused on separate modeling of either time-domain or frequency-domain information. However, due to the complementary nature of the latent information in these two domains, integrating them can enhance the performance of the models [17, 18].

In this paper, we propose a novel joint time-domain and frequency-domain progressive learning approach (TFDPL) for single-channel speech enhancement and recognition. TFDPL progressively predicts less-noisy and clearer speech, simultaneously estimates time-frequency masks and waveform using information from both the time and frequency domains, and further combines these two prediction targets through a fusion loss. Experimental results demonstrate that TFDPL outperforms traditional methods in ASR and perceptual metrics. On the CHiME-4 real test set, TFDPL's intermediate output achieves relative word error rate (WER) reductions of 43.83% and 36.03% compared to the untreated noisy speech, respectively, using two different acoustic models without retraining. The final results also demonstrate the best PESQ and STOI scores on the simulated test set.

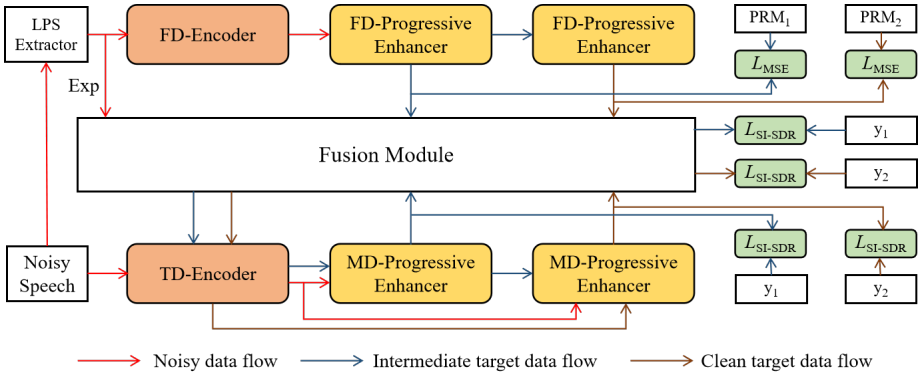


Fig. 1. The overview of proposed joint time-domain and frequency-domain progressive learning network.

2 Progressive learning with joint time domain and frequency domain

In this section, we will provide a detailed description of the TFDPL model. We divide the prediction of clean speech into two stages, aiming to progressively predict intermediate target speech with a 10dB improvement in SNR relative to the noisy speech and the final clean speech. The overview of the TFDPL model is shown in Fig. 1.

TFDPL model consists of three modules: the progressive frequency-domain masking module, the progressive mix-domain module, and the fusion module. The TFDPL model takes noisy time-domain signals as input. First, the LPS features of the signal are extracted and normalized [19]. Then, the normalized features are fed into the progressive frequency-domain masking module to estimate progressive masks. The estimated masks are multiplied with the original spectrogram and reconstructed back to waveform signals using inverse STFT (ISTFT). The reconstructed signals, along with the original noisy speech, are then fed into the progressive mix-domain module. In addition, we propose a novel fusion strategy, where the fusion module extracts the corresponding LPS features from the estimated targets of the two modules at the same stage of progressive learning, and weights and reconstructs them into fused speech to better utilize frequency and time domain information.

2.1 Problem formulation

For single-channel speech enhancement, we have a noisy speech signal denoted as y, which consists of a combination of the clean target speech signal s and background noise signal n.

$$y(t) = s(t) + n(t) \tag{1}$$

where t indicates a time index. The model progressively predicts the corresponding frequency-domain mask and time-domain waveform for each stage. In [20] and [15], progressive ratio masks (PRM) and low-noise speech were used as estimation targets for each stage, demonstrating their effectiveness in improving ASR systems. Therefore, the predicted frequency-domain mask and time-domain waveform for the i-th stage of the TFDPL model are defined as follows:

$$M_{\mathrm{PRM}_i}(k,l) = \frac{|S(k,l)|^2 + |N_i(k,l)|^2}{|S(k,l)|^2 + |N(k,l)|^2} \tag{2}$$

$$y_i(t) = s(t) + n_i(t) \tag{3}$$

where $M_{\mathrm{PRM}_i}(k,l)$ and $y_i(t)$ represent the predicted targets of the progressive frequency-domain masking module and the progressive mix-domain module in the i-th stage, respectively. $n_i(t)$ represents the residual noise in the i-th stage, while $S(k,l)$, $N_i(k,l)$ and $N(k,l)$ represent the STFT of the clean speech, residual noise in the i-th stage, and input noise, respectively. k and l are indices representing frames and frequency bins, respectively.

Indeed, it can be observed that when $N_i(k,l)$ and $n_i(t)$ are both equal to 0, M_{PRM_i} and y_i correspond to the traditional IRM and the clean speech s, respectively. They serve as the clean targets for the progressive frequency-domain masking module and the progressive mix-domain module.

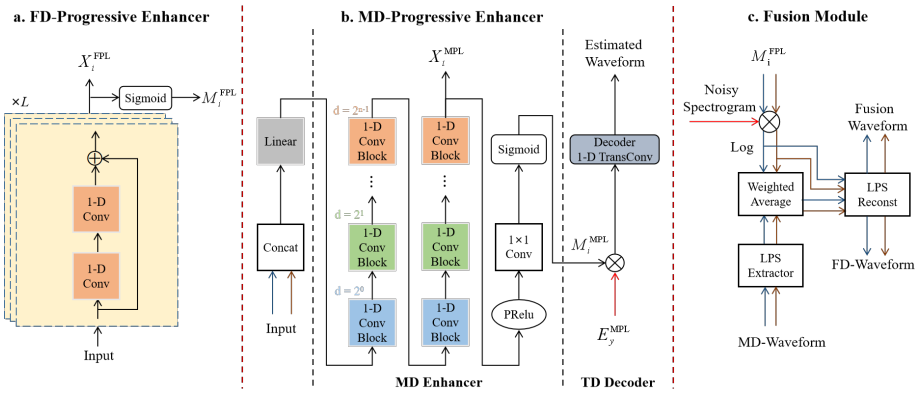


Fig. 2. The detailed design of the components in the TFDPL model. (a) Frequency-domain progressive enhancer. (b) Mix-domain progressive enhancer. (c) Fusion module.

2.2 Joint time and frequency domain progressive learning

Progressive frequency-domain masking module The progressive frequency-domain masking module consists of two components: the frequency-domain (FD) encoder and the frequency-domain (FD) progressive enhancer (shown in Fig.

2(a)). The data flow and parameters of this module are denoted as FPL. In [16], each stage of progressive learning is composed of L Blocks, each consisting of a 1D convolutional layer with residual connection, a ReLU activation, and a batch normalization. We employ 5 Blocks and 2 conformer layers as the FD encoder, denoted as $\mathcal{F}^{\mathrm{FPL}}_{\mathrm{encoder}}(\cdot)$.

First, the LPS features of the input signal y are extracted and then fed into the FD encoder:

$$E^{\mathrm{FPL}} = \mathcal{F}^{\mathrm{FPL}}_{\mathrm{encoder}}(\mathcal{F}^{\mathrm{LPS}}(y), \Lambda^{\mathrm{FPL}}_{\mathrm{encoder}}) \tag{4}$$

where $\mathcal{F}^{\mathrm{LPS}}(\cdot)$ represents LPS extractor. $\Lambda^{\mathrm{FPL}}_{\mathrm{encoder}}$ and E^{FPL} represent the parameter sets of the FD encoder and the encoded features.

$\mathcal{F}^{\mathrm{FPL}}_{\mathrm{enhancer}_i}(\cdot)$ denotes subsequent FD progressive enhancer composed of 5 Blocks, which is used to predict the intermediate representation of that stage:

$$X^{\mathrm{FPL}}_i = \mathcal{F}^{\mathrm{FPL}}_{\mathrm{enhancer}_i}(X^{\mathrm{FPL}}_{i-1}, \Lambda^{\mathrm{FPL}}_{\mathrm{enhancer}_i}) \tag{5}$$

where $\Lambda^{\mathrm{FPL}}_{\mathrm{enhancer}_i}$ and X^{FPL}_i represent the parameter set and the intermediate representation in the i-th stage, respectively. In the case of the first stage, X^{FPL}_0 is equivalent to E^{FPL}.

Then, the mask M^{FPL}_i for the i-th stage is obtained by applying a sigmoid activation.

$$M^{\mathrm{FPL}}_i = \sigma(X^{\mathrm{FPL}}_i) \tag{6}$$

where M^{FPL}_i is a mask with values ranging from 0 to 1, we expect to use the mask to obtain enhanced speech in the i-th stage with a specified relative increase in SNR compared to the noisy input:

$$\hat{y}^{\mathrm{FPL}}_i = \mathcal{F}^{\mathrm{Reconst}}_{\mathrm{PRM}}(y, M^{\mathrm{FPL}}_i, W_{\mathrm{istft}}) \tag{7}$$

$\mathcal{F}^{\mathrm{Reconst}}_{\mathrm{PRM}}(\cdot)$ represents the waveform reconstruction module, which involves multiplying the extracted spectrum of the input noisy speech y with the predicted mask, and then using the y phase information to reconstruct the waveform using ISTFT. $\hat{y}^{\mathrm{FPL}}_i$ represents the waveform reconstructed by the progressive frequency-domain masking module in the i-th stage.

Progressive mix-domain module The progressive mix-domain module consists of two components: a time-domain (TD) encoder, and a mix-domain (MD) progressive enhancer (shown in Fig. 2(b)). The data flow and parameters of this module are denoted as MPL. The MD progressive enhancer consists of three sub-parts: a linear layer, a mix-domain (MD) enhancer, and a time-domain (TD) decoder. We introduced a linear layer to fuse the encoded features of the time-domain waveform and the waveform reconstructed based on the progressive frequency-domain masking module. Each MD enhancer consists of two stacks of n 1D-ConvBlocks, with increasing dilation factors. The structure of the 1D-ConvBlocks is the same as in Conv-TasNet [21], but without skip connections.

The trainable TD encoder takes noisy speech y and progressive frequency-domain masking module estimated speech $\hat{y}_i^{\mathrm{FPL}}$ as inputs, where i takes a value of 1 or 2, resulting in high-dimensional encoded features.

$$(E_y^{\mathrm{MPL}}, E_{\hat{y}_1^{\mathrm{FPL}}}^{\mathrm{MPL}}, E_{\hat{y}_2^{\mathrm{FPL}}}^{\mathrm{MPL}}) = \mathcal{F}_{\mathrm{encoder}}^{\mathrm{MPL}}((y, \hat{y}_1^{\mathrm{FPL}}, \hat{y}_2^{\mathrm{FPL}}), \Lambda_{\mathrm{encoder}}^{\mathrm{MPL}}) \tag{8}$$

where $\mathcal{F}_{\mathrm{encoder}}^{\mathrm{MPL}}(\cdot)$ represents TD encoder, which consists of a 1D convolutional layer. E_y^{MPL}, $E_{\hat{y}_1^{\mathrm{FPL}}}^{\mathrm{MPL}}$, and $E_{\hat{y}_2^{\mathrm{FPL}}}^{\mathrm{MPL}}$ denote the encoded high-level features of the input noisy speech y, $\hat{y}_1^{\mathrm{FPL}}$, and $\hat{y}_2^{\mathrm{FPL}}$, respectively. $\Lambda_{\mathrm{encoder}}^{\mathrm{MPL}}$ represents the parameter set of the TD encoder. Next, we concatenate the features E_y^{MPL} and $E_{\hat{y}_1^{\mathrm{FPL}}}^{\mathrm{MPL}}$, and pass the concatenated feature through a linear layer for information fusion and dimensionality reduction. This allows us to obtain a feature that contains both time-domain and frequency-domain information.

$$E_1^{\mathrm{MPL}} = \mathcal{F}_{\mathrm{linear}_1}(Concat(E_y^{\mathrm{MPL}}, E_{\hat{y}_1^{\mathrm{FPL}}}^{\mathrm{MPL}}), \Lambda_{\mathrm{linear}_1}) \tag{9}$$

where $Concat$ represents the concatenation of two vectors along the feature dimension. E_1^{MPL} represents the fused features in the first stage, which is then fed into the first MD enhancer.

$$[X_1^{\mathrm{MPL}}, M_1^{\mathrm{MPL}}] = \mathcal{F}_{\mathrm{enhancer}_1}^{\mathrm{MPL}}(E_1^{\mathrm{MPL}}, \Lambda_{\mathrm{enhancer}_1}^{\mathrm{MPL}}) \tag{10}$$

where $\mathcal{F}_{\mathrm{enhancer}_1}^{\mathrm{MPL}}(\cdot)$ and $\Lambda_{\mathrm{enhancer}_1}^{\mathrm{MPL}}$ represent the first MD enhancer and the corresponding parameter set. X_1^{MPL} and M_1^{MPL} denote the mix-domain module intermediate representation and masks in the first stage. We concatenate X_1^{MPL} with $X_{\hat{y}_2^{\mathrm{FPL}}}^{\mathrm{MPL}}$ and feed it into the second linear layer to extract fusion information for the second stage, which is then passed to the second MD progressive enhancer.

$$E_2^{\mathrm{MPL}} = \mathcal{F}_{\mathrm{linear}_2}(Concat(X_1^{\mathrm{MPL}}, E_{\hat{y}_2^{\mathrm{FPL}}}^{\mathrm{MPL}}), \Lambda_{\mathrm{linear}_2}) \tag{11}$$

$$[\_, M_2^{\mathrm{MPL}}] = \mathcal{F}_{\mathrm{enhancer}_2}^{\mathrm{MPL}}(E_2^{\mathrm{MPL}}, \Lambda_{\mathrm{enhancer}_2}^{\mathrm{MPL}}) \tag{12}$$

where E_2^{MPL} and M_2^{MPL} represent the fusion features and masks for the second stage. Finally, we perform element-wise multiplication between the obtained masks (M_1^{MPL} and M_2^{MPL}) and the E_y^{MPL} separately, and pass each result through their respective TD decoder to obtain the estimated intermediate targets and clean targets from the mix-domain module.

$$\hat{y}_i^{\mathrm{MPL}} = \mathcal{F}_{\mathrm{decoder}_i}^{\mathrm{MPL}}(E_y^{\mathrm{MPL}}, M_i^{\mathrm{MPL}}, \Lambda_{\mathrm{decoder}_i}^{\mathrm{MPL}}) \tag{13}$$

where $\mathcal{F}_{\mathrm{decoder}_i}^{\mathrm{MPL}}(\cdot)$ and $\Lambda_{\mathrm{decoder}_i}^{\mathrm{MPL}}$ represent the i-th TD decoder and its corresponding parameter set, and $\hat{y}_i^{\mathrm{MPL}}$ represents the waveform estimated by the progressive mix-domain module in the i-th stage.

Fusion module This module connects the progressive frequency-domain masking module and progressive mix-domain module, as shown in Fig. 2(c) for more details. First, the fusion module receives the masking M_i^{FPL} estimated by the progressive frequency-domain masking module and the spectrum and phase of the noisy speech y to obtain the reconstructed speech $\hat{y}_i^{\mathrm{FPL}}$ through ISTFT. Afterward, we can input $\hat{y}_i^{\mathrm{FPL}}$ into the progressive mix-domain module to fuse the information from both the time-domain and frequency-domain, and obtain its predicted waveform $\hat{y}_i^{\mathrm{MPL}}$.

Finally, we propose a novel fusion strategy that combines the outputs of the progressive frequency-domain masking module and progressive mix-domain module, aiming to further exploit the complementarity of different domain information:

$$\mathrm{LPS}_i^{\mathrm{fusion}} = \lambda * \mathcal{F}^{\mathrm{LPS}}(\hat{y}_i^{\mathrm{MPL}}) + (1-\lambda) * \mathcal{F}^{\mathrm{LPS}}(\hat{y}_i^{\mathrm{FPL}}) \tag{14}$$

$$\hat{y}_i^{\mathrm{fusion}} = \mathcal{F}_{\mathrm{LPS}}^{\mathrm{Reconst}}(\hat{y}_i^{\mathrm{MPL}}, \mathrm{LPS}_i^{\mathrm{fusion}}, W_{\mathrm{istft}}) \tag{15}$$

By employing a weighted fusion approach, we acquire the fused LPS feature $\mathrm{LPS}_i^{\mathrm{fusion}}$, with a weight parameter λ ranging from 0 to 1. The function $\mathcal{F}_{\mathrm{LPS}}^{\mathrm{Reconst}}(\cdot)$ denotes the waveform reconstruction based on the LPS features and the phase of $\hat{y}_i^{\mathrm{MPL}}$. $\hat{y}_i^{\mathrm{fusion}}$ represents the waveform obtained through the fusion strategy.

2.3 Multi-target loss

We propose a multi-task learning approach to train TFDPL. In this approach, we utilize a multi-scale scale-invariant signal-to-distortion ratio (SI-SDR) loss [22] for the progressive mix-domain module and fusion module. Additionally, we use a minimum mean squared error (MSE) loss for the progressive frequency-domain masking module.

$$\mathcal{L}_i^{\mathrm{FPL}} = \mathcal{L}_{\mathrm{MSE}}(M_i^{\mathrm{FPL}}, M_{\mathrm{PRM}_i}) \tag{16}$$

$$\mathcal{L}_i^{\mathrm{MPL}} = \mathcal{L}_{\text{SI-SDR}}(\hat{y}_i^{\mathrm{MPL}}, y_i) \tag{17}$$

$$\mathcal{L}_i^{\mathrm{fusion}} = \mathcal{L}_{\text{SI-SDR}}(\hat{y}_i^{\mathrm{fusion}}, y_i) \tag{18}$$

The final optimization objective of TFDPL is a linear combination of the three mentioned losses.

$$\mathcal{L}^{\mathrm{TFDPL}} = \sum_i \alpha_i * \mathcal{L}_i^{\mathrm{FPL}} + \beta_i * \mathcal{L}_i^{\mathrm{MPL}} + \gamma_i * \mathcal{L}_i^{\mathrm{fusion}} \tag{19}$$

3 Experiments and Analysis

3.1 Data corpus

Clean speech is obtained from the WSJ0 SI-84 dataset [23], which consists of 7,138 utterances from different speakers. We randomly selected 7,000 utterances for training, 65 for validation, and 73 for testing. The noise data used to generate

noisy-clean pairs is sourced from the CHIME-4 noise dataset [24]. During the training process, we employed an online augmentation strategy where speech and noise pairs were randomly selected and randomly segmented into durations ranging from 4 to 6 seconds. The SNR was varied between -5dB and 5dB to generate noisy speech. We simulated a test set to evaluate the perceived quality of the enhanced speech at five different SNR levels (-10dB, -5dB, 0dB, 5dB, 10dB) using 73 clean speech utterances from the test set and three training-time unseen noises from the NOISEX-92 corpus [25]: Destroyer Engine, Factory1, and Speech Babble. Additionally, we conducted an ASR performance evaluation of our framework on the CHiME-4 real test set, which includes 1,320 real recordings in four different conditions: bus (BUS), cafe (CAF), pedestrian area (PED), and street (STR).

3.2 Implementation details

The speech waveform was sampled at a frequency of 16kHz. We applied a 32-ms Hanning window with 16-ms overlap to extract audio frames. Then, a 512-point STFT was used to compute the spectrum of each frame, resulting in 257-dimensional LPS features. Before feeding the features into the neural network, they were normalized using global mean and variance [19].

For the TFDPL model, each MD enhancer is composed of two stacks, each containing 8 1D-ConvBlocks. The remaining hyper-parameters setting is similar to the original Conv-TasNet [21]. Besides, we do not use skip connection in MD enhancer. We used PyTorch to train the model, with an initial learning rate set to 5e-4. The batch size was 12. If the loss on the validation set did not decrease after an epoch, the learning rate was halved. Adam [26] is used as the optimizer. For the loss configuration, we use $\alpha_1 = 2.3$, $\alpha_2 = 1.5$, and β_i and γ_i are set to 1 to balance the training loss. For the fusion configuration, we set $\lambda = 0.5$.

We trained four baseline models for comparison with our proposed method. The first model has the same network structure as the progressive frequency-domain masking module of TFDPL, denoted as FDPL. The second model is denoted as TDPL, we set the stack of the progressive enhancer in TDPL to contain 8 1D-ConvBlocks, because we found that this has better performance, and the rest of the hyperparameters are set similarly to [15]. The third model, denoted PL-ANSE [14], combines progressive learning with the traditional IMCRA [27] algorithm. The fourth is the traditional speech enhancement Conv-TasNet, denoted as TDSE.

We evaluated our method on two different ASR systems. The first one is an official ASR system [24], referred to as $ASR(1)$, where the acoustic model is trained using the DNN-HMM architecture with sMBR criteria [28]. The second system is trained with LF-MMI using TDNN and is referred to as $ASR(2)$. Both ASR systems used a 5-gram Kneser-Ney (KN) smoothed language model for the first-pass decoding [29], and scoring was performed using an RNN-based language model.

3.3 Evaluation metrics

To evaluate the perceived quality of the enhanced speech, we employed the perceptual evaluation of speech quality (PESQ) [30] and short-time objective intelligibility (STOI) [31] metrics. Higher values in both metrics indicate better performance. Additionally, we employed word error rate (WER) to assess the model's improvement on the ASR system, where lower values are better. The intermediate and clean outputs of the model render great service to ASR back-end and human listener, respectively.

3.4 Results and analysis

In this section, we will compare the performance of the proposed model with other models in terms of ASR back-end and perceived quality of the enhanced speech. Explanation of terms in Tables 1, 2, 3, 4 and 5: Noisy represents the noisy speech without any enhancement. D-Fusion represents divide fusion, which is the fusion output obtained by applying the fusion strategy proposed in Section 2.2 to separately trained FDPL and TDPL models. J-Fusion represents joint fusion, which indicates the output of the fusion module in TFDPL.

Table 1. WER(%) comparison of different targets for different methods on CHiME-4 real test set with $ASR(1)$ and $ASR(2)$.

Target	Model	$ASR(1)$	$ASR(2)$
Noisy	-	23.84	13.46
-	PL-ANSE	18.57	12.48
Clean	TDSE	24.55	21.69
+10dB	FDPL	16.26	10.38
	TDPL	15.18	9.74
	D-Fusion	14.41	9.12
Clean	FDPL	17.76	11.54
	TDPL	26.43	25.16
	D-Fusion	14.83	10.64

Analysis on recognition performance Table 1 presents the results of different speech enhancement methods in ASR systems. PL-ANSE [14] combines multiple targets of progressive learning with IMCRA [27] algorithm to estimate speech. We can observe that the intermediate target (+10dB) of progressive learning can always improve the performance of ASR, while its clean target or TDSE directly estimating the clean target may degrade the performance of ASR. Studies [20] and [15] have demonstrated separately that the intermediate targets in progressive frequency-domain and time-domain models effectively enhance the performance of ASR systems. Our experiments validate this finding. Additionally, the ASR performance of the intermediate targets of FDPL and TDPL is better than that of PL-ANSE and TDSE, so we choose FDPL and TDPL for subsequent fusion experiments. When FDPL and TDPL results are fused, the performance of intermediate target fusion still exceeds the performance of clean

target fusion on two different ASR backends and improves ASR performance. Due to the superior performance of the intermediate targets of the progressive learning method in ASR, subsequent ASR experiments are based on the intermediate targets. Furthermore, the fused results show improvements compared to the best results of the individual models before fusion, providing strong evidence for the complementary nature of the frequency and time domains and indicating the effectiveness of our fusion strategy. This is also the motivation behind our joint training.

Table 2. WER(%) comparison of different speech enhancement methods in various environments on CHiME-4 real test set with $ASR(1)$.

Model	Domain	BUS	CAF	PED	STR	AVG
Noisy	-	36.55	24.73	19.92	14.16	23.84
FDPL	*Freq.*	24.36	17.69	13.30	9.69	16.26
TDPL	*Time*	21.18	16.16	13.32	10.05	15.18
D-Fusion	*Time & Freq.*	21.05	15.39	12.01	9.19	14.41
TFDPL	*Freq.*	21.63	14.94	11.79	9.28	14.41
	Time & Freq.	20.21	14.59	12.05	9.19	14.01
J-Fusion	*Time & Freq.*	**19.42**	**14.40**	**11.04**	**8.70**	**13.39**

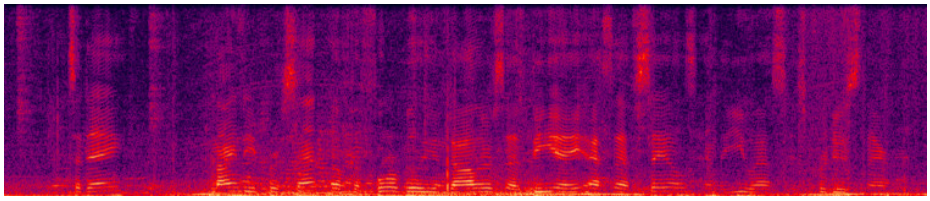
Table 3. WER(%) comparison of different speech enhancement methods in various environments on CHiME-4 real test set with $ASR(2)$.

Model	Domain	BUS	CAF	PED	STR	AVG
Noisy	-	21.16	13.39	10.41	8.87	13.46
FDPL	*Freq.*	16.41	10.44	8.48	6.20	10.38
TDPL	*Time*	14.41	9.60	8.13	6.84	9.74
D-Fusion	*Time & Freq.*	13.74	9.10	7.60	6.05	9.12
TFDPL	*Freq.*	13.98	9.66	7.94	6.05	9.41
	Time & Freq.	12.99	9.54	7.75	6.54	9.21
J-Fusion	*Time & Freq.*	**12.75**	**8.52**	**7.21**	**5.98**	**8.61**

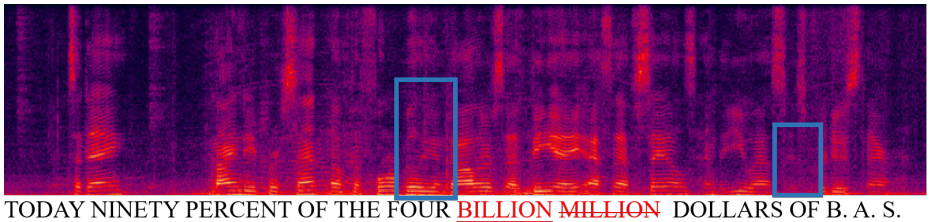
Tables 2 and 3 display the WER for different speech enhancement methods under various environmental conditions on the CHiME-4 real test set, using $ASR(1)$ and $ASR(2)$. The observation reveals that the intermediate targets of progressive learning have improved ASR performance in various environmental conditions. By comparing the WER in various environments, FDPL and TDPL have demonstrated their advantages in low-noise and challenging conditions, respectively, highlighting the complementarity of frequency-domain and time-domain information. We further fused the prediction results of TDPL and FDPL, and it led to improved ASR performance in all environments, highlighting the robustness of time and frequency domain information fusion.

Regarding the TFDPL model, its progressive frequency-domain masking module shows significant improvements compared to FDPL, achieving relative improvements of 11.38% and 9.34% on the two acoustic models, respectively, with

no additional parameters compared to FDPL. As for the progressive mix-domain module, it incorporates information from the frequency-domain leading to performance improvements relative to TDPL in all environments for two acoustic models, with average relative improvements of 7.71% and 5.44%, respectively. The performance of both the progressive frequency-domain masking module and progressive mix-domain module in TFDPL has improved, indicating that there is a mutually reinforcing effect between the time-domain and frequency-domain in TFDPL, thus confirming the effectiveness of TFDPL.

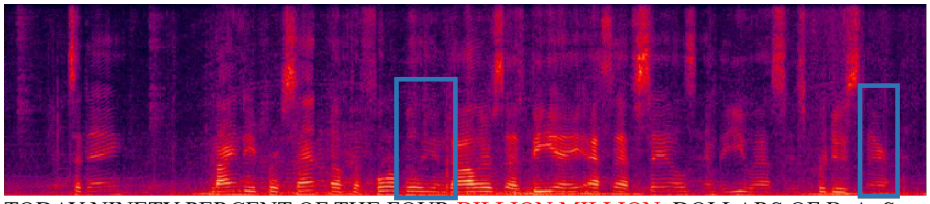


(a) Noisy



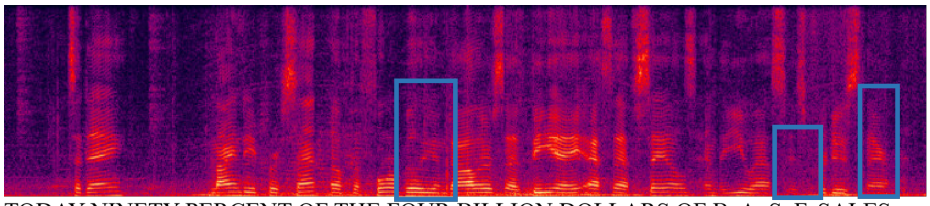
TODAY NINETY PERCENT OF THE FOUR BILLION ~~MILLION~~ DOLLARS OF B. A. S. F. SALES IN ~~AND~~ THE U. S. IS ~~HAS~~ PRODUCED THERE

(b) FDPL



TODAY NINETY PERCENT OF THE FOUR BILLION ~~MILLION~~ DOLLARS OF B. A. S. F. SALES IN ~~AND~~ THE U. S. IS PRODUCED THERE ~~AND~~

(c) TDPL



TODAY NINETY PERCENT OF THE FOUR BILLION DOLLARS OF B. A. S. F. SALES IN ~~AND~~ THE U. S. IS PRODUCED THERE

(d) TFDPL

Fig. 3. Spectrograms and ASR Results of FDPL, TDPL, and TFDPL. (a) Noisy speech, (b) Intermediate output of the FDPL model, (c) Intermediate output of the TDPL model, (d) Intermediate output of the TFDPL model fusion module.

Finally, the fusion of TFDPL model results achieved relative improvements of 43.83% and 36.03% compared to noisy speech in two acoustic models, reaching the best performance. Compared to the fusion results of separately trained FDPL and TDPL, TFDPL achieved relative improvements of 7.08% and 5.59% in two acoustic models, and relative improvements of 11.79% and 11.60% compared to TDPL. In Fig. 3, we selected a representative utterance from a real BUS environment to visually compare the performance of FDPL, TDPL, and TFDPL. In the blue boxed regions in Fig. 3(b) and Fig. 3(c), the target speech of FDPL and TDPL respectively experienced excessive suppression and distortion, resulting in substitution errors in the corresponding ASR results. TFDPL effectively combines information from both the time and frequency domain, thereby avoiding such errors. The experimental results strongly demonstrate the effectiveness of TFDPL's intermediate targets in ASR.

Table 4. PESQ comparison on different speech enhancement methods at several SNRs.

Metrics		PESQ					
Model	Domain	-10	-5	0	5	10	*avg.*
Noisy	-	1.38	1.52	1.79	2.11	2.44	1.85
TDSE	*Time*	1.57	2.27	2.76	3.11	3.39	2.62
FDPL	*Freq.*	1.48	1.99	2.42	2.75	3.02	2.33
TDPL	*Time*	1.59	2.28	2.78	3.14	3.41	2.64
D-Fusion	*Time & Freq.*	1.66	2.33	2.84	3.20	3.45	2.70
TFDPL	*Freq.*	1.54	2.06	2.47	2.79	3.05	2.38
	Time & Freq.	1.41	2.11	2.67	3.04	3.36	2.52
J-Fusion	*Time & Freq.*	**1.75**	**2.42**	**2.94**	**3.27**	**3.50**	**2.78**

Table 5. STOI(%) comparison on different speech enhancement methods at several SNRs.

Metrics		STOI(%)					
Model	Domain	-10	-5	0	5	10	*avg.*
Noisy	-	49.53	60.20	72.38	83.10	90.46	71.13
TDSE	*Time*	62.40	81.77	90.19	93.99	95.99	84.87
FDPL	*Freq.*	58.35	74.10	84.76	90.99	94.53	80.55
TDPL	*Time*	62.66	82.21	90.57	94.25	96.17	85.17
D-Fusion	*Time & Freq.*	64.77	82.26	90.47	94.30	96.29	85.62
TFDPL	*Freq.*	59.63	75.22	85.65	91.48	94.75	81.35
	Time & Freq.	62.84	82.46	90.99	94.62	96.43	85.47
J-Fusion	*Time & Freq.*	**65.53**	**82.99**	**91.12**	**94.69**	**96.46**	**86.16**

Analysis on perceptual quality metrics Table 4 and 5 presents the average PESQ and STOI comparisons of different models' clean targets across five SNR levels and three types of unknown noise. The PESQ and STOI of TDPL's clean target are superior to TDSE's clean target, demonstrating the effectiveness of progressive learning methods in perceptual quality. The fusion results of TFDPL

achieved the best PESQ and STOI scores across all SNRs. The PESQ and STOI of TDPL's clean target are superior to TDSE's clean target, demonstrating the effectiveness of progressive learning methods in perception. The fusion results of FDPL and TDPL outperformed their individual models, indicating the complementary nature of time and frequency domain information in improving perceptual quality. Particularly, at SNR of -10dB, the STOI improvement relative to TDPL was 2.11, demonstrating the robustness of the time and frequency domain fusion in challenging environments.

For the progressive frequency-domain masking module of TFDPL, it still maintains improvement over FDPL. In the progressive mix-domain module, it shows an improvement in STOI compared to TDPL but a decrease in PESQ. However, the fusion of the two yields the best performance, with an improvement of 0.08 in PESQ and 0.76 in STOI compared to the fusion results of separately trained models, demonstrating its effectiveness in perceptual quality.

4 Conclusion

In this paper, we propose a TFDPL method for speech enhancement and recognition. TFDPL progressively predicts less-noisy and clearer speech, while estimating time-frequency masks and waveforms using information in the time and frequency domains, and further combines these two prediction targets through a fusion loss. Finally, the mutually beneficial effect of time-domain and frequency-domain is achieved. The experimental results demonstrate that TFDPL outperforms both time-domain and frequency-domain progressive learning methods, as well as their fusion results, in both ASR and human listener tasks. On the CHiME-4 real test set, TFDPL's intermediate output achieves relative WER reductions of 43.83% and 36.03% compared to the untreated noisy speech, under two different acoustic models. The clean output also demonstrates the best PESQ and STOI scores on the simulated test set. The positive experimental results demonstrate the effectiveness of the TFDPL approach.

References

1. Stephen E Levinson, Lawrence R Rabiner, and M Mohan Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.
2. Yu Zhang, William Chan, and Navdeep Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4845–4849.
3. Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.
4. Yanhui Tu, Jun Du, Lei Sun, Feng Ma, and Chin-Hui Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones.," in *INTERSPEECH*, 2017, pp. 394–398.
5. DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

6. DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.
7. Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.
8. Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.
9. Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7009–7013.
10. Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.
11. Ashutosh Pandey and DeLiang Wang, "Dense cnn with self-attention for time-domain speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 1270–1279, 2021.
12. Quan Wang, Ignacio Lopez Moreno, Mert Saglam, Kevin Wilson, Alan Chiao, Renjie Liu, Yanzhang He, Wei Li, Jason Pelecanos, Marily Nika, et al., "Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition," *arXiv preprint arXiv:2009.04323*, 2020.
13. Yan-Hui Tu, Jun Du, Tian Gao, and Chin-Hui Lee, "A multi-target snr-progressive learning approach to regression based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1608–1619, 2020.
14. Zhaoxu Nian, Yan-Hui Tu, Jun Du, and Chin-Hui Lee, "A progressive learning approach to adaptive noise and speech estimation for speech enhancement and noisy speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6913–6917.
15. Zhaoxu Nian, Jun Du, Yu Ting Yeung, and Renyu Wang, "A time domain progressive learning approach with snr constriction for single-channel speech enhancement and recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6277–6281.
16. Chen-Yue Zhang, Hang Chen, Jun Du, Bao-Cai Yin, Jia Pan, and Chin-Hui Lee, "Incorporating visual information reconstruction into progressive learning for optimizing audio-visual speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
17. Feng Dang, Qi Hu, Pengyuan Zhang, and Yonghong Yan, "Forknet: Simultaneous time and time-frequency domain modeling for speech enhancement," *arXiv preprint arXiv:2305.08292*, 2023.
18. Chuanxin Tang, Chong Luo, Zhiyuan Zhao, Wenxuan Xie, and Wenjun Zeng, "Joint time-frequency and time domain learning for speech enhancement," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3816–3822.
19. Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

20. Nan Zhou, Jun Du, Yan-Hui Tu, Tian Gao, and Chin-Hui Lee, "A speech enhancement neural network architecture with snr-progressive multi-target learning for robust speech recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 873–877.
21. Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
22. Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.
23. Douglas B Paul and Janet Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.
24. Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.
25. Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.
26. Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
27. Israel Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.
28. George Saon and Hagen Soltau, "A comparison of two optimization techniques for sequence discriminative training of deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5567–5571.
29. Reinhard Kneser and Hermann Ney, "Improved backing-off for m-gram language modeling," in *1995 international conference on acoustics, speech, and signal processing*. IEEE, 1995, vol. 1, pp. 181–184.
30. Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.
31. Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.