Target Speaker Extraction Method by Emphasizing the Active Speech with an Additional Enhancer

Xue Yang, Changchun Bao*, Xu Zhang and Xianhong Chen
Institute of Speech and Audio Information Processing, School of Information Science and Technology,
Beijing University of Technology, Beijing 100124, China
E-mail: yangx11@emails.bjut.edu.cn, baochch@bjut.edu.cn, xuzhang223@emails.bjut.edu.cn, chenxianhong@bjut.edu.cn

Abstract—**Target speaker extraction (TSE) is a practical solution to the cocktail party problem. Recently, a novel embedding-free TSE method was proposed. In this method, the enrollment and the mixed signal are directly interacted to exploit the contextual information within the enrollment. In the absence of noise, the derived guidance exhibits the onset, offset and voice activity similar to the mixed signal. However, in the presence of noise, such similarity may be destroyed since the enrollment is interacted with both speech and noise signals in the mixture. If the noise (e.g., babble noise) contains components that resemble the enrollment to some extent, the misleading guidance may be generated after the direct interaction. To tackle this issue, an additional enhancer is designed in this paper to derive an auxiliary guidance that emphasizes the active speech. Specifically, this enhancer consists of a processing block and an interaction block. The processing block mainly utilizes the recurrent layers to model the temporal dynamics of the enrollment and mixed signal. In this block, the speech and noise signals are modeled in different manners and the similarity between the enrollment and noise can be reduced. Afterwards, the processed representations of the enrollment and mixed signal are utilized to derive an enhanced representation in the interaction block. This enhanced representation emphasizes the active speech and is employed as an auxiliary guidance for the extraction. Experimental results demonstrate the effectiveness of our proposed method in complex acoustic environments.**

I. Introduction

In complex acoustic environments, the mixed signal captured by the microphone may comprise speech signals from multiple speakers, noise and reverberation. While human beings can effectively focus on the speaker of interest in such environment, it is hard for the machine to possess this capability. This challenge is known as the cocktail party problem [1]. To tackle this problem, the speech separation (SS) and target speaker extraction (TSE) are primarily studied. The SS aims to estimate all speech signals within the mixed signal [2] and has recently achieved outstanding performance [3-6]. However, the SS assumes to know the number of speakers and faces the global permutation ambiguity [7-8]. These issues make the SS less practical in real-world scenarios. In contrast, the TSE focuses on isolating the target speaker's speech, guided by the auxiliary information of the target speaker [9]. Hence, the TSE is a more practical solution since it does not require to know the number of speakers and avoids the global permutation ambiguity. In this paper, the reference utterance of the target speaker, known as the enrollment, is employed as the auxiliary information due to its effectiveness and easy accessibility.

Depending on how enrollment is leveraged, the TSE can be divided into two categories: embedding-based TSE method and embedding-free TSE method. Typically, the embedding-based TSE method derives a speaker embedding from the enrollment and employs this embedding as guidance. In [10-11], the speaker embedder pretrained on the speaker verification or recognition task was employed to derive the embedding. In [12-13], the speaker embedder was jointly trained with the extraction network to address the sub-optimization issue. Besides, techniques such as speaker representation loss [14] or self-supervised disentangled representation [15] were adopted to improve the embedding discriminability. Following the derivation of the speaker embedding, the extraction process involves fusing the embedding with the features of the mixed signal and various fusion methods have been proposed [16-19]. Additionally, the multi-stage framework [20], onset/offset information [21] or iterative refined adaptation [22] have been used to boost the extraction performance. Nevertheless, being a compact vector, speaker embedding mainly summarizes the speaker characteristics and discards the content details. Thus, these embedding-based TSE methods may not fully leverage the contextual information within the enrollment.

Recently, the embedding-free TSE methods have gained increasing attention since both the speaker characteristics and the content details are leveraged. In [23-25], high-dimensional feature sequences derived from the enrollment and the mixed signal were fused through various attention mechanisms. In [26], the hidden states and cell states of the recurrent neural network (RNN) were adopted to summarize the speaker information. In [27], two pooling approaches were introduced to generate speaker representations. In these methods, the processed feature sequences or states were utilized as guidance and the contextual information is partially exploited. In [28], the enrollment and mixed signal were directly interacted through an attention mechanism. In the absence of noise, the derived guidance of this contextual information exploitation network (CIENet) exhibits the onset, offset and voice activity similar to the mixed signal. However, in the presence of noise, such similarity may be destroyed since the enrollment is interacted with both speech and noise signals in the mixture. If the noise (e.g., babble noise) contains components that resemble the enrollment to some extent, the misleading guidance may be generated after the direct interaction. Therefore, it is necessary to explore more appropriate guidance in complex acoustic environments.

In this paper, we extend the CIENet with an additional enhancer to derive an auxiliary guidance that emphasizes the active speech. Specifically, this additional enhancer consists of a processing block and an interaction block. The processing block mainly utilizes the recurrent layers to model the temporal dynamics of the enrollment and mixed signal. In this block, the speech and noise signals are modeled in different manners and the similarity between the enrollment and noise can be reduced. Afterwards, the processed representations of the enrollment and the mixed signal are interacted to derive an enhanced representation in the interaction block. This enhanced representation emphasizes the active speech and is used as an auxiliary guidance for the extraction. This enhanced contextual information exploitation network (CIENet-Enh) achieves better performance in complex acoustic environments.

The rest of this paper is organized as follows. The proposed TSE method is first detailed in Section 2. Afterwards, the experimental setup is outlined in Section 3. The results and discussions are presented in Section 4. Finally, the conclusions are drawn in Section 5.

II. Proposed Method

A. Problem Formulation

Guided by the target speaker's enrollment $\boldsymbol{e}\in\mathcal{R}^{1\times L_e}$, the TSE aims to isolate the target speaker's speech from the mixed signal $\boldsymbol{y}\in\mathcal{R}^{1\times L_y}$, namely:

$$\hat{x}_{tgt}=\mathcal{M}\left(\boldsymbol{y}\middle|\boldsymbol{e}\right)=\mathcal{M}\left(\left(\sum_{i=1}^{S}\boldsymbol{z}_i+\boldsymbol{n}\right)\middle|\boldsymbol{e}\right) \tag{1}$$

where L_e is the length of the enrollment, L_y is the length of the mixed signal, $\hat{x}_{tgt}$ is the estimated signal of the target speaker, $\mathcal{M}(\boldsymbol{a}|\boldsymbol{b})$ is the mapping function of $\boldsymbol{a}$ given $\boldsymbol{b}$, S is the number of speakers, the vector $\boldsymbol{z}_i$ denotes the signal coming from the i^{th} speaker and the vector $\boldsymbol{n}$ represents the additive noise.

B. Proposed Network Architecture

As illustrated in Fig. 1, our proposed CIENet-Enh extends the original CIENet with an additional enhancer.

1) *Original CIENet*

The original CIENet takes the mixed signal $\boldsymbol{y}$ and the enrollment $\boldsymbol{e}$ as inputs to isolate the target speaker's speech through mask estimation. As given in the lower part of Fig. 1, the mixed signal $\boldsymbol{y}$ and the enrollment $\boldsymbol{e}$ are first transformed into the time-frequency (T-F) domain using short-time Fourier transform (STFT). Their T-F representations $\boldsymbol{Y}\in\mathcal{C}^{T_Y\times F}$ and $\boldsymbol{E}\in\mathcal{C}^{T_E\times F}$ are compressed with the dynamic range compression (DRC) [29] to reduce the dynamic range and emphasize the regions with small values. The compressed T-F representations $\boldsymbol{Y_c}\in\mathcal{R}^{2\times T_Y\times F}$ and $\boldsymbol{E_c}\in\mathcal{R}^{2\times T_E\times F}$ are expressed as

$$\boldsymbol{Y}_c=\mathrm{Concat}\left(\left|\boldsymbol{Y}\right|^{\gamma}\cos\boldsymbol{\theta}_Y,\left|\boldsymbol{Y}\right|^{\gamma}\sin\boldsymbol{\theta}_Y\right) \tag{2}$$

$$\boldsymbol{E}_c=\mathrm{Concat}\left(\left|\boldsymbol{E}\right|^{\gamma}\cos\boldsymbol{\theta}_E,\left|\boldsymbol{E}\right|^{\gamma}\sin\boldsymbol{\theta}_E\right) \tag{3}$$

where T_Y and T_E are the frame numbers of the mixed signal and the enrollment, F is the number of frequencies, $|\boldsymbol{Y}|$ and $|\boldsymbol{E}|$ are the magnitude spectra, $\boldsymbol{\theta}_Y$ and $\boldsymbol{\theta}_E$ are the phase spectra, Concat(·, ·) is the concatenation operation, γ denotes the compression factor ranging from 0 to 1.

Afterward, the T-F representations $\boldsymbol{Y_c}$ and $\boldsymbol{E_c}$ are interacted in the interaction block to derive the guidance $\boldsymbol{F_c}\in\mathcal{R}^{2\times T_Y\times F}$. For clarity, the magnitude spectrum comparison is depicted in Fig. 2. In the absence of noise, the direct interaction between the enrollment in Fig. 2(a) and the clean mixture in Fig. 2(b) leads to the guidance in Fig. 2(c). This derived guidance not only retains target speaker's characteristics to some extent but also exhibits the onset, offset and voice activity similar to the clean mixture. In the presence of noise, the enrollment is interacted with both speech and noise signals in the mixture. If the noise contains components that resemble the enrollment to some extent, the misleading guidance may be generated. For instance, the direct interaction between the enrollment in Fig. 2(d) and the noise in Fig. 2(e) is depicted in Fig. 2(f). Besides, the direct interaction between the enrollment in Fig. 2(g) and the noisy mixture in Fig. 2(h) is shown in Fig. 2(i). Compared with the guidance in Fig. 2(c), misleading guidance in Fig. 2(i) no longer exhibits the onset, offset and voice activity similar to the clean mixture. Thus, more appropriate guidance needs to be explored.

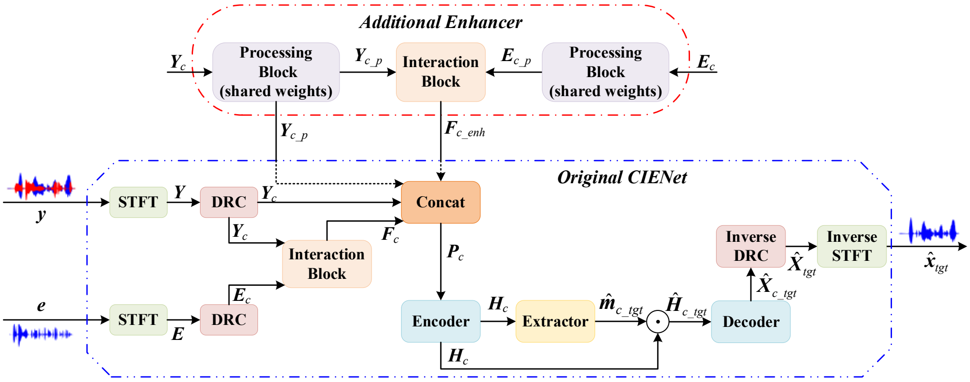

Fig.1 The general framework of CIENet-Enh.

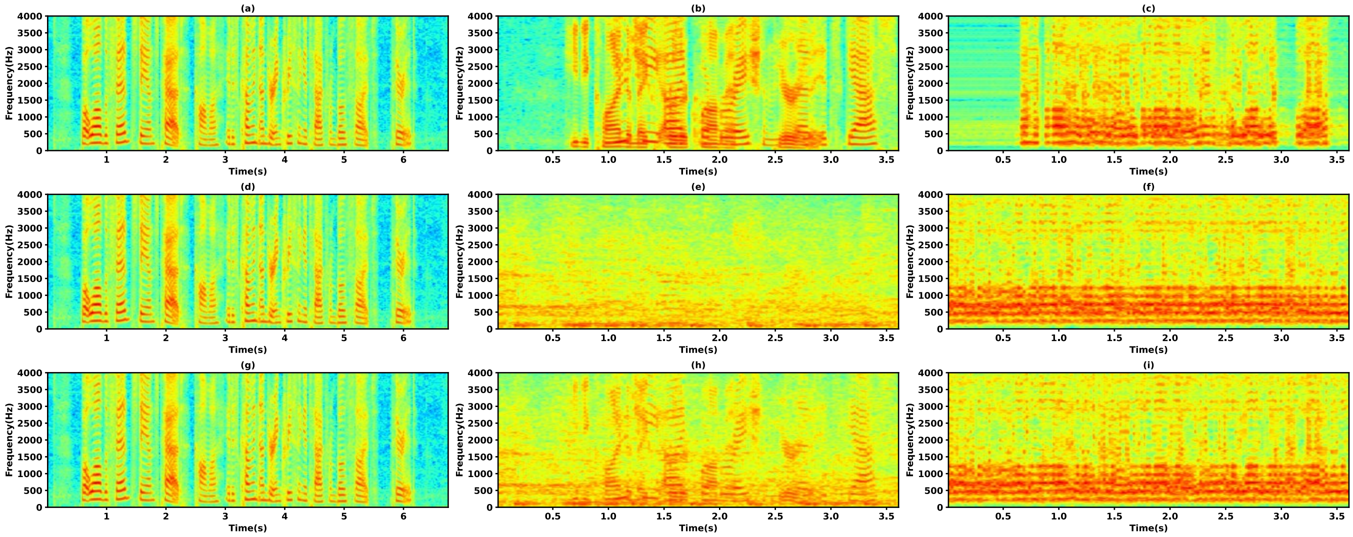

Fig. 2 The magnitude spectrum comparison in the original CIENet. (a) Enrollment; (b) Clean mixture; (c) Interaction between the enrollment and clean mixture; (d) Enrollment; (e) Noise; (f) Interaction between the enrollment and noise; (g) Enrollment; (h) Noisy mixture; (i) Interaction between the enrollment and noisy mixture.

In the original CIENet, the guidance $\boldsymbol{F}_c$ is concatenated with the T-F representation $\boldsymbol{Y}_c$ of the mixed signal. The stacked feature tensor of shape $4 \times T_Y \times F$ is further processed by the encoder and the extractor to estimate the mask $\hat{\boldsymbol{m}}_{c\_tgt} \in \mathbb{R}^{K \times T_Y \times F}$. As illustrated in Fig. 1, this mask is element-wise multiplied with the output $\boldsymbol{H}_c$ of the encoder to derive the feature tensor $\hat{\boldsymbol{H}}_{c\_tgt}$ of the target speaker, namely:

$$\hat{\boldsymbol{H}}_{c\_tgt} = \boldsymbol{H}_c \odot \hat{\boldsymbol{m}}_{c\_tgt} \tag{4}$$

where K is the channel dimension and the symbol "$\odot$" denotes the element-wise multiplication. The derived feature tensor $\hat{\boldsymbol{H}}_{c\_tgt}$ is processed by the decoder to obtain the estimated T-F representation $\hat{X}_{c\_tgt} \in \mathbb{R}^{2 \times T_Y \times F}$. Finally, the inverse DRC and inverse STFT are applied to recover the estimated signal $\hat{\boldsymbol{x}}_{tgt}$. In practice, the STFT and inverse STFT are implemented with convolutional layer and transposed convolutional layer.

2) *Additional Enhancer*

To derive more appropriate guidance for complex acoustic environments, an additional enhancer is introduced. As shown in the upper part of Fig. 1, this enhancer consists of a processing block and an interaction block.

The T-F representation $\boldsymbol{Y}_c$ and $\boldsymbol{E}_c$ are individually processed by the processing block to derive the processed representations $\boldsymbol{Y}_{c\_p}$ and $\boldsymbol{E}_{c\_p}$. As depicted in Fig. 3(a), the channel dimension of the representation $\boldsymbol{Y}_c$ is first expanded to D using a two-dimensional convolutional layer. Then, the expanded feature sequence undergoes reshaping and modeling by two bi-directional long short-term memory (BLSTM) [30] layers. The resulting feature is reshaped and further transformed with a convolutional layer to obtain the processed representation $\boldsymbol{Y}_{c\_p} \in \mathbb{R}^{2 \times T_Y \times F}$. Similarly, the T-F representation $\boldsymbol{E}_c$ of the enrollment is processed to derive the processed representation $\boldsymbol{E}_{c\_p} \in \mathbb{R}^{2 \times T_E \times F}$. Note that the BLSTM layers are utilized to model the temporal dynamics of the enrollment and mixed signal. Since the temporal dynamics of speech and noise signals are distinct, they can be modeled in different manners. In this way, the similarity between the enrollment and noise can be reduced.

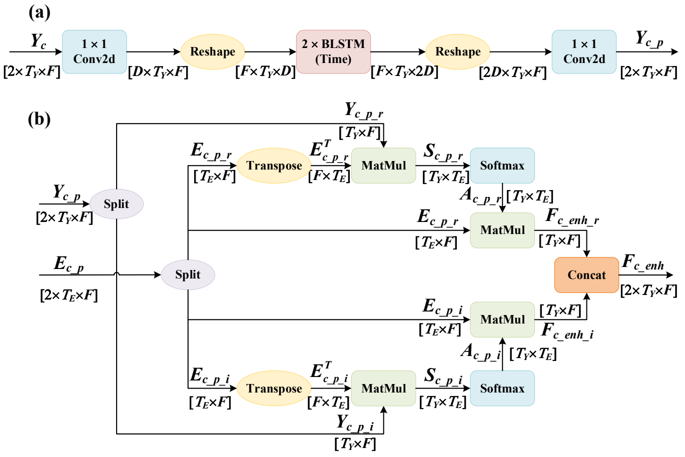

Fig.3 (a) Processing block; (b) Interaction block.

Afterwards, these processed representations $\boldsymbol{Y}_{c\_p}$ and $\boldsymbol{E}_{c\_p}$ are interacted in the interaction block to derive the enhanced representation $\boldsymbol{F}_{c\_enh} \in \mathbb{R}^{2 \times T_Y \times F}$. As illustrated in Fig. 3(b), these representations are split into the real-part features ($\boldsymbol{Y}_{c\_p\_r}$ and $\boldsymbol{E}_{c\_p\_r}$) and imaginary-part features ($\boldsymbol{Y}_{c\_p\_i}$ and $\boldsymbol{E}_{c\_p\_i}$). The real-part feature $\boldsymbol{Y}_{c\_p\_r}$ is matrix-multiplied with the transposed real-part feature $\boldsymbol{E}^T_{c\_p\_r}$ to calculate similarity matrix $\boldsymbol{S}_{c\_p\_r}$ for different frames. Subsequently, the Softmax function is applied on the last dimension of this similarity matrix to derive the attention matrix $\boldsymbol{A}_{c\_p\_r}$. This attention matrix is further matrix-multiplied with the real-part feature $\boldsymbol{E}_{c\_p\_r}$ to obtain the enhanced real-part feature $\boldsymbol{F}_{c\_enh\_r}$. Similarly, the enhanced imaginary-part feature $\boldsymbol{F}_{c\_enh\_i}$ can be derived, namely:

$$\boldsymbol{F}_{c\_enh\_r} = \boldsymbol{A}_{c\_p\_r}\boldsymbol{E}_{c\_p\_r} = \text{Softmax}\left(\boldsymbol{Y}_{c\_p\_r}\boldsymbol{E}^T_{c\_p\_r}\right)\boldsymbol{E}_{c\_p\_r} \tag{5}$$

$$\boldsymbol{F}_{c\_enh\_i} = \boldsymbol{A}_{c\_p\_i}\boldsymbol{E}_{c\_p\_i} = \text{Softmax}\left(\boldsymbol{Y}_{c\_p\_i}\boldsymbol{E}^T_{c\_p\_i}\right)\boldsymbol{E}_{c\_p\_i} \tag{6}$$

where the superscript "T" is the transpose operation. These enhanced features are concatenated to produce the enhanced representation $\boldsymbol{F}_{c\_enh} \in \mathbb{R}^{2 \times T_Y \times F}$, which is further employed as an auxiliary guidance. As shown in Fig. 1, the representations $\boldsymbol{Y}_c$,

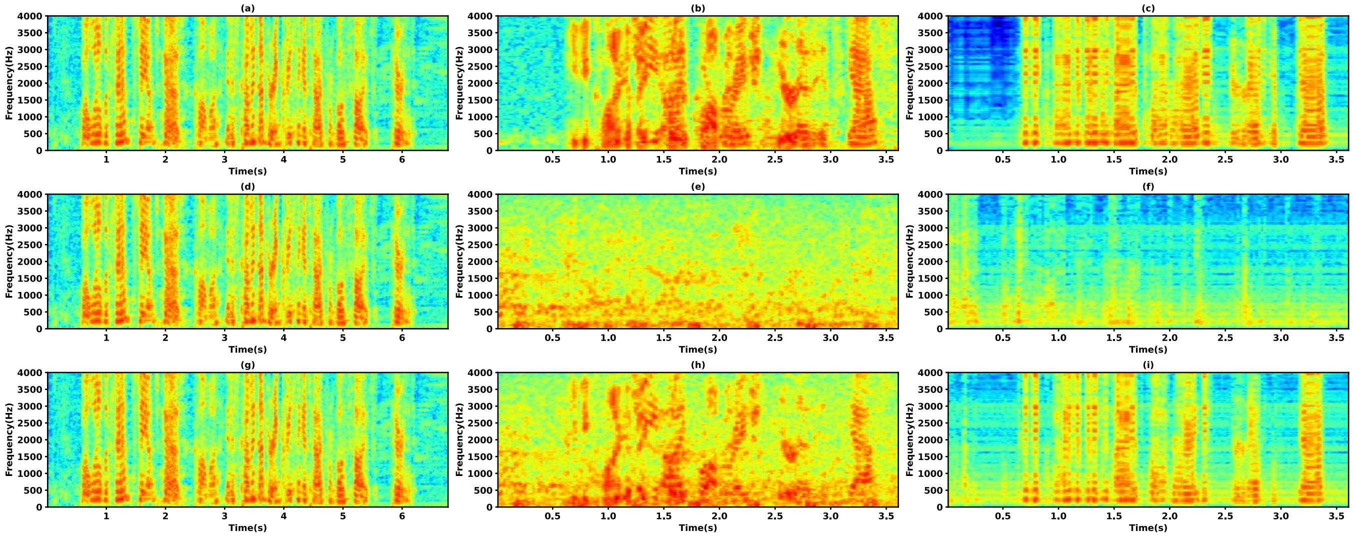

Fig. 4 The magnitude spectrum comparison in the additional enhancer.
(a) Processed enrollment; (b) Processed clean mixture; (c) Interaction between the processed enrollment and processed clean mixture;
(d) Processed enrollment; (e) Processed noise; (f) Interaction between the processed enrollment and processed noise;
(g) Processed enrollment; (h) Processed noisy mixture; (i) Interaction between the processed enrollment and processed noisy mixture.

$\boldsymbol{F}_c$, $\boldsymbol{Y}_{c\_p}$ and $\boldsymbol{F}_{c\_enh}$ are concatenated to form the stacked feature tensor $\boldsymbol{P}_c \in R^{8\times T_Y \times F}$. Note that the same interaction block is used in the original CIENet.

For clarity, the magnitude spectrum comparison in this additional enhancer is depicted in Fig. 4. After modeling in the processing block, the harmonic structures of the processed enrollment in Fig. 4(a) and the processed clean mixture in Fig. 4(b) are somewhat distorted. However, their interaction in Fig. 4(c) still exhibits certain voice activity of the clean mixture and target speaker's characteristics. Additionally, the interaction between the processed enrollment in Fig. 4(d) and processed noise in Fig. 4(e) is given in Fig. 4(f). Compared with the direct interaction between the enrollment and noise in Fig. 2(f), the similarity between the processed enrollment and the processed noise is effectively reduced. This may be due to the different modeling manners for the speech and noise signals in the processing block. Furthermore, the interaction between the processed enrollment in Fig. 4(g) and the processed noisy mixture in Fig. 4(h) is shown in Fig. 4(i). It can be observed that this enhanced representation emphasizes the active speech in complex acoustics environments. Therefore, this enhanced representation is employed as an auxiliary guidance, although the over-suppression may occur.

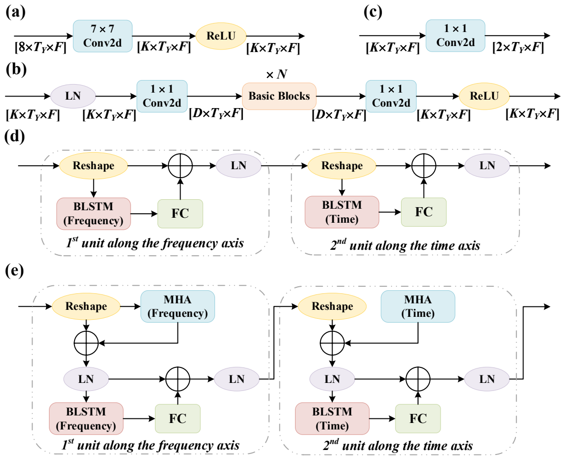

Fig. 5 (a) Encoder; (b) Extractor; (c) Decoder;
(d) mDPRNN block; (e) mDPTNet block.

3) *Details of Different Modules*

The details of the encoder, extractor and decoder are illustrated in Fig. 5. The encoder consists of a convolutional layer and a rectified linear unit (ReLU). The extractor comprises layer normalization (LN), *N* basic blocks positioned between two convolutional layers and ReLU. The decoder is a convolutional layer. For fair comparison with the original CIENet, the same types of basic blocks are used. The first type is the modified dual-path RNN (mDPRNN) block derived from [3]. This block includes two units to model dependencies along both the frequency and time axes. Each unit consists of BLSTM, fully connected (FC) layer, skip connection and LN. The second type is the modified dual-path transformer (mDPTNet) block derived from [4], which uses multi-head attention (MHA).

III. Experimental Setup

Three benchmark datasets derived from the Wall Street Journal (WSJ0) corpus are utilized in this paper. The first one is WSJ0 Hipster Ambient Mixtures (WHAM!) [31], which includes noises collected from the real-world scenarios. The second dataset WHAMR! [32] is the reverberant version of the WHAM! dataset. The third dataset is WSJ0-2Mix [33], where neither noise nor reverberation is introduced. Note that each speaker in the mixed signal is considered as the target speaker in turn. The sampling rate is 8kHz for all three datasets.

To perform TSE in the T-F domain, the Hanning window is utilized to split waveforms. The window length and the hop size are set to 32ms and 16ms, respectively. Besides, the number of frequencies *F* is 129. The compression factor γ in DRC is set to 0.5. The hyper-parameters *K* and *D* are 256 and 64, respectively. In the basic blocks, the number of hidden units for each direction of BLSTM is 128 and the number of attention heads is 4. Besides, the number of basic blocks N is set to 6.

All models are trained with 4s long speech signals for 120 epochs. The Adam [34] optimizer is utilized and the initial learning rate is set to 0.0003. In the first 100 epochs, the learning rate is multiplied with a factor of 0.99 for ever two epochs. In the last 20 epochs, this factor is reduced to 0.9. The gradient clipping is adopted to limit the maximum L_2 norm to 1. The training objective is to maximize the scale-invariant signal-to-distortion ratio (SI-SDR) [35] between the estimated signal and the ground-truth signal of the target speaker.

IV. Results and Discussions

The SI-SDR improvement (SI-SDRi) and signal-to-distortion ratio [36] improvement (SDRi) are employed to evaluate the extraction accuracy.

A. Performance Comparison on the Complex Datasets

In Tab. 1, the CIENet-Enh is compared with several methods on the WHAM! and WHAMR! datasets. The embedding-based methods and embedding-free methods are separated with a double line. The first four methods conduct the TSE in the time-domain, while others perform the TSE in the T-F domain.

Tab. 1 Comparison on the complex datasets.

Methods	WHAM!		WHAMR!	
	SI-SDRi (dB)	SDRi (dB)	SI-SDRi (dB)	SDRi (dB)
SpEx [8]	12.2†	13.0†	10.3†	9.5†
SpEx+ [13]	13.1†	13.6†	10.9†	10.0†
DPRNN-Spe-IRA [22]	14.2	14.6	–	–
SpEx++ [20]	14.3	14.7	11.7	10.7
X-TF-GridNet [19]	15.7	16.1	15.3	14.2
Enhance-CIENet-mDPTNet	3.5	4.8	3.8	3.8
CIENet-mDPRNN [28]	15.7	16.1	15.5	14.1
CIENet-mDPTNet [28]	16.6	17.0	15.7	14.3
CIENet-Enh-mDPRNN	16.1	16.4	16.4	15.0
CIENet-Enh-mDPTNet	**17.2**	**17.5**	**17.2**	**15.8**

-Results with superscript "†" are given in [20].

On both datasets, the X-TF-GridNet [19] outperforms the time-domain approach using advanced network architecture [5] and fusion method. All five methods utilize the embedding as guidance and do not fully leverage the contextual information within the enrollment. In contrast, the CIENet achieves higher performance by fully leveraging the contextual information. However, in the presence of noise, misleading guidance may be generated. By introducing an additional enhancer, an auxiliary guidance emphasizing the active speech is leveraged and the CIENet-Enh can achieve better performance. Additionally, the Enhance-CIENet-mDPTNet is designed to show that CIENet-Enh is more effective than the enhance-then-extract approach. In this reference method, the mixed signal is first denoised with a pretrained model [37] and then the extraction is conducted with the pretrained CIENet-mDPTNet model. As shown in Tab. 1, this reference method performs the worst on both datasets. This is due to the sub-optimization issue of the cascaded system and the distortions of speech signals introduced during the enhancement process. In constrast, the CIENet-Enh can be optimized end-to-end, utilizing an enhanced representaion that emphasizes the active speech as an auxiliary guidance.

B. Performance Comparison on the WSJ0-2Mix Dataset

In Tab. 2, the CIENet-Enh is compared with a baseline method and CIENet on the WSJ0-2Mix dataset. The baseline method is similar to the CIENet-Enh, except that the direct interaction $\boldsymbol{F}_c$ is not used as guidance. Compared with CIENet-mDPRNN, the performance of the baseline method degrades, indicating that the direct interaction $\boldsymbol{F}_c$ is more effective on the clean dataset. Therefore, both the direct interaction $\boldsymbol{F}_c$ and the enhanced representation $\boldsymbol{F}_{c\_enh}$ are utilized as guidance in the CIENet-Enh. Additionally, the CIENet-Enh does not exhibit a significant improvement over CIENet. This is reasonable since the direct interaction $\boldsymbol{F}_c$ can already provide effective guidance on this clean dataset. Furthermore, only a marginal increase is observed in the number of parameters (N_p) for CIENet-Enh.

Tab. 2 Comparison on the WSJ0-2Mix dataset.

Methods	N_p ($\times 10^6$)	WSJ0-2Mix	
		SI-SDRi (dB)	SDRi (dB)
Baseline-mDPRNN	2.9	20.4	20.6
CIENet-mDPRNN [28]	2.7	20.7	21.0
CIENet-mDPTNet [28]	2.9	21.4	21.6
CIENet-Enh-mDPRNN	2.9	20.9	21.1
CIENet-Enh-mDPTNet	3.1	**21.5**	**21.8**

V. Conclusions

In this paper, an additional enhancer was introduced to extend CIENet for better accommodation of complex acoustic environments. Specifically, this enhancer includes a processed block and an interaction block. In the processing block, the temporal dynamics of the enrollment and mixed signal were modeled. By modeling the speech and noise signals in different manners, the similarity between the enrollment and noise was reduced. As a results, an enhanced representation emphasizing the active speech was derived in the interaction block. This enhanced representation was employed as an auxilary guidance for the extraction. Our proposed CIENet-Enh achieved superior performance on both WHAM! and WHAMR! datasets.

VI. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 61831019 and 62006010), R&D Program of Beijing Municipal Education Commission (KM202210005029).

References

[1] E. C. Cherry, On Human Communication, Cambridge, MA, USA: MIT Press, 1957.

[2] JD. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 46-50.

[4] 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.

[5] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221-3236, 2023.

[6] X. Yang, C. Bao, and X. Chen, "Coarse-to-fine speech separation in the time-frequency domain," in *Speech Communication*, vol. 155, 103003, 2023.

[7] M. Delcroix *et al.*, "Improving speaker discrimination of target speech extraction with time-domain Speakerbeam," in *IEEE ICASSP*, 2020, pp. 691-695.

[8] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370-1384, 2020.

[9] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Cernocky, and D. Yu, "Neural target speech extraction: An overview," in *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8-29, 2023.

[10] Q. Wang *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728-2732.

[11] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *IEEE ASRU*, 2019, pp. 327-334.

[12] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE ICASSP*, 2018, pp. 5554-5558.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1406-1410.

[14] S. Mun, S. Choe, J. Huh, and J. S. Chung, "The sound of my voice: Speaker representation loss for target voice separation," in *IEEE ICASSP*, 2020, pp. 7289-7293.

[15] Z. Mu, X. Yang, S. Sun, and Q. Yang, "Self-supervised disentangled representation learning for robust target speech extraction," in *Proc. AAAI*, 2024, pp. 18815-18823.

[16] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for Speakerbeam target speaker extraction," in *IEEE ICASSP*, 2019, pp. 6965-6969.

[17] W. Wang, C. Xu, M. Ge, and H. Li, "Neural speaker extraction with speaker-speech cross-attention network," in *Proc. Interspeech*, 2021, pp. 3535-3539.

[18] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-SEPFORMER: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *IEEE ICASSP*, 2023, pp. 1-5.

[19] F. Hao, X. Li, and C. Zheng, "X-TF-GridNet: A time-frequency domain target speaker extraction network with adaptive speaker embedding fusion," in *Information Fusion*, 102550, 2024.

[20] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *IEEE ICASSP*, 2021, pp. 6109-6113.

[21] Y. Hao, J. Xu, P. Zhang, and B. Xu, "Wase: Learning when to attend for speaker extraction in cocktail party environments," in *IEEE ICASSP*, 2021, pp. 6104-6108.

[22] C. Deng *et al.*, "Robust speaker extraction network based on iterative refined adaptation," in *Proc. Interspeech*, 2021, pp. 3530-3534.

[23] X. Xiao *et al.*, "Single-channel speech extraction using speaker inventory and attention network," in *IEEE ICASSP*, 2019, pp. 86-90.

[24] B. Zeng, S. Hongbin, Y. Wan, and M. Li, "SEF-Net: Speaker embedding free target speaker extraction network," in *Proc. Interspeech*, 2023, pp. 3452-3456.

[25] Y. Hu, H. Xu, Z. Guo, H. Huang, and L. He, "SMMA-Net: An audio clue-based target speaker extraction network with spectrogram matching and mutual attention," in *IEEE ICASSP*, 2024, pp. 1496-1500.

[26] L. Yang, W. Liu, L. Tan, J. Yang, and H.-G. Moon, "Target speaker extraction with ultra-short reference speech by VE-VE framework," in *IEEE ICASSP*, 2023, pp. 1-5.

[27] W. Zhang, L. Yang, and Y. Qian, "Exploring time-frequency domain target speaker extraction for causal and non-causal processing," in *IEEE ASRU*, 2023, pp. 1-6.

[28] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target speaker extraction by directly exploiting contextual information in the time-frequency domain," in *IEEE ICASSP*, 2024, pp. 10476-10480.

[29] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," in *JASA Express Letters*, vol. 1, no. 1, pp. 014802, 2021.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[31] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368-1372.

[32] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *IEEE ICASSP*, 2020, pp. 696-700.

[33] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE ICASSP*, 2016, pp. 31-35.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[35] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?," in *IEEE ICASSP*, 2019, pp. 626-630.

[36] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.

[37] Y.-X. Lu, Y. Ai, and Z.-H, Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proc. Interspeech*, 2023, pp. 3834-3838.