Audio-only Target Speech Extraction with Dual-Path Band-Split RNN

1st *Anonymous*
Faculty of Electronic and Information Engineering
Xi'an Jiaotong University
Xi'an, China

***Abstract*—Even if more and more deep learning models in the field of target speech extraction (TSE) [1] achieved better performance on certain datasets by continuously refining modules and experimenting with new algorithms, they still remain constrained by generic frameworks and have not been able to propose new task decomposition mechanisms or utilize new information. In this paper, we propose a novel model architecture that focuses on extracting the target speaker and suppressing interfering noise simultaneously, acknowledging the intrinsic similarity in the nature of these tasks. The model is divided into two branches, one for extracting the target speaker's speech and the other for computing the speech of the interferer, thus controlling the shallow latent features to learn the essence of TSE task. Additionally, we adopt a mechanism similar to self-enrollment, where the latent features of the two branches are cross-fused at each stage of the extraction process, in order to further leverage the results obtained by both branches.**

***Index Terms*—target speech extraction, dual-path mechanism, information disentanglement, multi-task learning**

I. INTRODUCTION

The human ear possesses selective auditory perception, where individuals selectively focus on sounds of interest while ignoring other ambient noises in many real-life scenarios. For instance, in a concert, individuals can concentrate on listening to the singer's voice while occasionally engaging in conversations with friends nearby. The objective of Target Speech Extraction (TSE) is to mimic this selective auditory perception of the human ear by separating the speech of the target person of interest from the mixed audio. It is an important research topic under the cocktail party [2] problem in speech processing. This task has various applications such as robust speech interfaces, auditory assistance systems, and so forth.

To specify the task of target speech extraction, the system has two input, one is the mixed audio wave, which contains the voice of target speaker. Another is a clue, which in our work is a segment of clean voice of the target speaker and named as the enrollment speech, to inform the model of the target speaker's identity. Then the model does the extraction process, and the output is the clean voice of the target speaker, whose content is consistent with that in the mixed audio.

The earliest methods for target speech extraction can be traced back to beamformers based on Singular Value Decomposition (SVD) [3]. Some methods also reference techniques from Blind Source Separation (BSS) tasks [4] [5] [7] [8]. However, these methods often rely on the access to a large amount of data from the target speaker's voice beforehand and training models specifically for the one target speakers [4].

In recent years, with the advancement of deep neural networks, various models have emerged in the field of target speech extraction, achieving increasingly better performance on various datasets. The majority of these models follow a common algorithmic framework. It has four parts, a clue encoder, a mixture encoder, a fusion module, and a extraction module. The clue encoder processes the clue, to get an auxiliary clue embedding, which contains fruitful identity information of the target speaker. The mixture encoder processes the mixed audio, to get a time series embedding, which contains information of the target speaker's voice, the interfering speaker's voice, and the ambient noise. Then the fusion module combines both the embeddings mentioned above, and input the refined embedding into the extraction module, which generate a mask to apply on the mixture embedding and get the final extracted clean voice by an inverse transformation operation of the mixture encoder. To be specific, the clue in our work is a segment of clean voice of the target speaker, which is called the enrollment speech. And consequently, the clue embedding can be viewed as the vocal print of the target speaker.

Following this framework, A variety of TSE models conducted research and implemented improvements on all of these modules. Some works view the clue embedding as a single vector [6] [8] [9] [10] and do some vector-level operation between the clue embedding and the mixture embedding as the fusion part, such as concatenation, addition or multiply. Some treat the clue embedding as a time series embedding [15] [16] just the same as the mixture and then apply time-sequence-level operation like cross attention. Some researchers focus on the encoder itself. For example, the enrollment speech can be processes to I-vector [11], D-vector [6] or X-vector [12]. All kinds of DNN-based encoder also keep emerging. SpEx [8] and MC-SpEx [13] study the feature space's consistency of the clue encoder and the mixture encoder. The different characteristics and advantages and disadvantages between frequency domain methods and time domain methods are still under intense debate. More works focus on the fusion module to seek for making the most use of the clue. Among them, SEF-Net [15] and AV-Sepformer [16] utilize the attention mechanism to mimic the reference effect of the clue. MC-

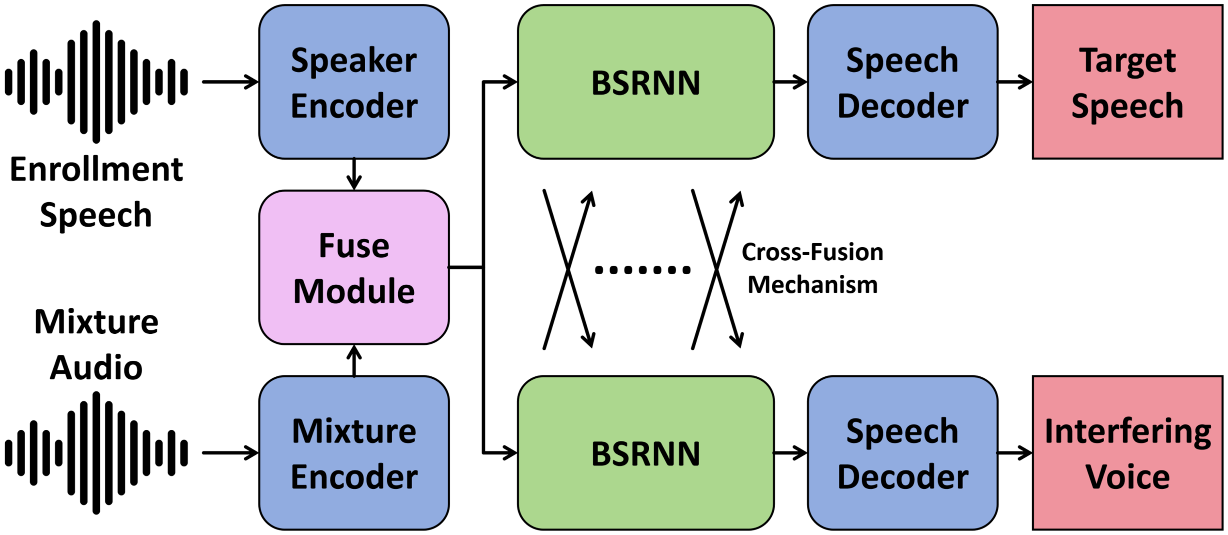

Fig. 1. An overall structure of Dual-Path Band-Split RNN (DP-BSRNN), the model contains four general steps and two branches.

SpEx [13] and X-TF-GridNet [18] treat different scales as channels and apply 2D convolution layers to mine cross-scale information. VEVEN [14] model changes the fusion layer into the delivery of RNN state to reduce the parameter count of the model. What's more, each work typically proposes a new backbone network to accommodate their improvements on the aforementioned modules, aiming to achieve the best performance in experiments.

Even though these works did make significant improvement by new module devises, they still share a common drawback, that they seldom break the aforementioned framework and seek for other promising mechanisms. For one simple example, those models never make use of the interfering speaker's voice, though extracting the target speaker's voice can be viewed as ridding the interfering voice as well. X-TasNet [19] first mentioned the feasibility of simultaneously outputting speech signals of both the target speaker and interfering speakers, which is also validated through experiments conducted by us in the recent past. Then a model in the field of speech enhancement utilized the dual-path processing and disentangled feature learning to successfully achieve state-of-the-art performance. Inspired by this, we propose our Dual-Path Band-Split RNN (DP-BSRNN), which uses the newly proposed BSRNN [17] in Music Source Separation task as the backbone, and extracts both the target speaker's voice and the interfering speaker's voice simultaneously.

II. Methods

The architecture of Dual-Path Band-Split RNN (DP-BSRNN) is shown in Fig.1. Note that BSRNN [17] has achieve an outstanding and stable performance in our former experiments of TSE task, we choose it as the backbone model. The DP-BSRNN has a clue encoder, a mixture encoder, a band split module at the beginning. Then the model bifurcate into two branches, one is to extract the target speech, another extracting the interfering voice. Each branch still follows the common framework of a extraction module, a mask generator and finally a voice decoder. We will introduce the details in this section.

A. Clue Encoder

We choose the simple ResNet34 followed by a mean-pooling layer as clue encoder here, since clue encoder is not the point in our work. The enrollment speech is extracted into one single feature vector, which is treated as the vocal print of the target speaker ideally.

B. Band Split Module

The band split module follows the same design as that in BSRNN. The model first takes the pure mixed audio signal as input and get the complex-valued spectrogram $\mathbf{X} \in \mathbb{C}^{F \times T}$, where F and T are the frequency and temporal dimensions respectively. The fullband spectrogram is then split into K subband spectrograms $\mathbf{B}_i \in \mathbb{C}^{G_i \times T}, i = 1, ..., K$ with pre-defined bandwidth $\{G_i\}_{i=1}^{K}$ satisfying $\sum_{i=1}^{K} G_i = F$. The real and imaginary parts of each subband spectrogram $\mathbf{B}_i$ are then concatenated and passed to a layer normalization module [20] and a fully-connected (FC) layer to generated an N-dimensional real-valued subband feature $\mathbf{Z}_i \in \mathbb{R}^{N \times T}$ Note that each subband spectrogram has its own normalization module and FC layer. All K subband features$\{Z_i\}_{i=1}^{K}$ are then merged to generated a transform feature tensor $\mathbf{Z} \in \mathbb{R}^{N \times K \times T}$.

C. Fusion Module

Our work aims to exploit the possibility to utilize the interfering voice for information disentanglement and seek for more comprehensive and powerful framework. So the fusion module is only a variable that need to be controlled in the ablation study instead of a significant proposal. Note that the enrollment speech is processed to one single feature vector, based on our previous experimental results, vector dot product generally achieves a better performance than addition and concatenation. So we choose dot product as the fusion method here. Note that the mixture embedding has been split into subbands in this step, we conduct the fusion operation on each subband instead of treating the mixture as an integral whole and fuse only once. The cross-fusion operation won't change the dimension of the latent features (if we apply concatenation, a linear layer will be following to rectify the output dimension).

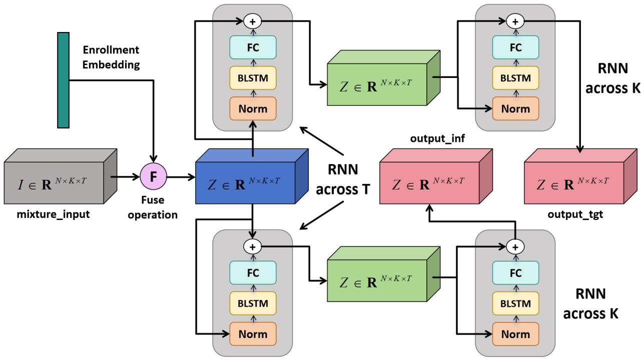

Fig. 2. The beginning of DP-BSRNN where the mixture embedding is input in two branches for further process. In the first block, we do not adopt information cross-fusion operations to ensure the distinctiveness of the two tasks.

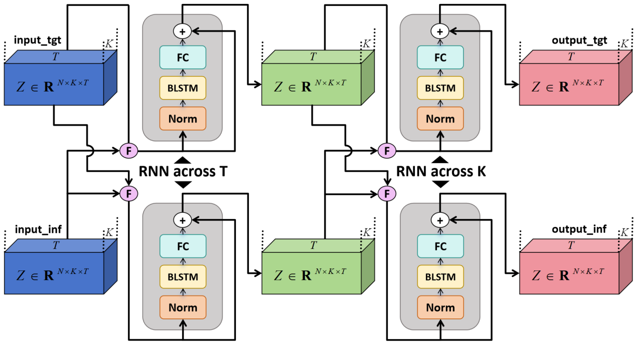

Fig. 3. The structure of intermediate block of the DP-BSRNN, before the model inputs the latent features into the RNN block each time, the information from both branches will undergo a cross-fusion operation.

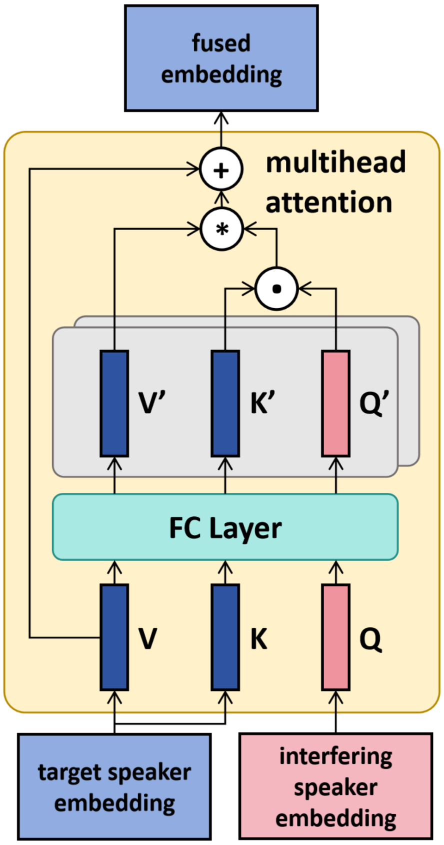

Fig. 4. A given example of the cross-fusion module between the two branches, using cross-attention mechanism.

D. Dual-Path Band and Sequence Modeling Module

The general design of dual-path band and sequence modeling module in each branch is the same as that in BSRNN [17]. Our work's innovation is to split the model into two branches of BSRNN processing, one to extraction the target speech, another to calculate the interfering voice. In this way, we can reinforce the model's understanding of the task based on the intermediate features obtained after fusion operations in shallow layers. The operation of extracting the target speaker's voice is equivalent to suppressing interference after extracting the interfering speaker's voice. Therefore, in this mode of information decoupling and multi-task training, if the shallow features of the model can simultaneously extract the voices of the target speaker and the interfering speaker with good performance, we can consider that the model has thoroughly understood the essence of the extraction task. What's more, in our previous study, we have already verified that when given the enrollment speech of the target speaker, the model can extract both the target speaker's voice and the interfering speaker's voice. Also, some works like SpEx++ [10] show that utilizing the periodical extraction results by the model itself is able to favour the following extracting process, which is called self-enrollment. Inspired by these, in the design of the two branches, we can also introduce the same fusion operation to enhance the extraction performance stage by stage.

Specifically, as depicted in Fig.2, These two submodels will share and merge latent features at each node before the band or sequence RNN block, and the fused features will be used as input in the subsequent stage of processing. What is mentionable here is that only blocks except the first one follow the design above. The only function of the first dual-path block, as depicted in Fig.3, is to split the outputs from one identical input. And we don't introduce cross-fusion layer in it. Regarding the fusion part, due to the isomorphic nature of the tasks in the two branches, simple vector-level fusion operations such as addition, multiplication, and concatenation lack interpretability. Therefore, inspired by innovative fusion operations which have achieved state-of-the-art performances from models like SEF-Net [15], AV-Sepformer [16], and X-TFGridNet [18], we decided to implement a cross-attention mechanism as the information interaction between the two branches, as illustrated in Fig.4.

Based on the devise, the output of each branch in this module is the same as that in BSRNN, denoted by $\mathbf{Q} \in \mathbb{R}^{N\times K\times T}$.

E. Dual-Path Mask generator

There is a mask generator following the sequence and band modeling module to calculate the mask, which will be apply on the mixture embedding in the next step. Recently some works like MC-SpEx [13] starts to research on the information mining ability of the mask generation step. Inspired by this, we decide to enable the mask generator in the two branches of the cross fusion mechanism, just like what the model did in the previous module, as is depicted in Fig.5. Unlike the dual-path sequence and band modeling module, the cross information operation is optimal instead of imperative, just like that between the RNN blocks. For the masks will be applied on the embedding of the mixed audio, the output of this module is denoted as $\mathbf{M}_i \in \mathbb{C}^{G_i\times T}, i = 1, ..., K$ with predefined bandwidth $\{G_i\}_{i=1}^{K}$ satisfying $\sum_{i=1}^{K} G_i = F$, then the K subband masks are concatenated to restore the complete mask $\mathbf{M} \in \mathbb{C}^{F\times T}$.

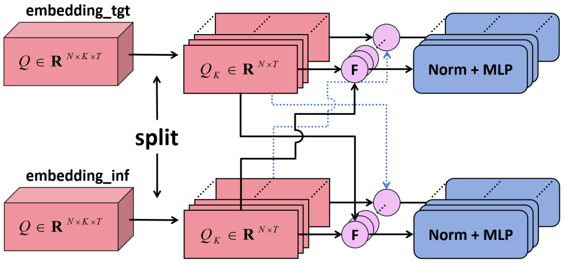
embedding_tgt
$Q \in \mathbf{R}^{N\times K\times T}$
$Q_K \in \mathbf{R}^{N\times T}$
F
Norm + MLP
split
embedding_inf
$Q \in \mathbf{R}^{N\times K\times T}$
$Q_K \in \mathbf{R}^{N\times T}$
F
Norm + MLP

Fig. 5. The structure of the mask generator in DP-BSRNN, the fusion operation is quite similar to the one in the previous module.

III. Experiments

A. Datasets and experiment setup

In line with methodologies from prior studies, our experiments are conducted using the Libri2Mix dataset. All training and testing data were manually synthesized by us. We selected two speakers' voices to synthesize a mixed audio, then extracted a segment of another speech from one speaker as the reference audio. In this way, the two speakers in the same mixed audio would take turns as the target speaker, responsible for two iterations of training or testing. There are a total of 3000 segments of mixed audio in the testing data, which means a total of 6000 tests can be conducted. The remaining synthesized audios are used as training data, totaling approximately 30,000 training samples.

We adopt a two-stage training mode with a pre-training step and a fine-tuning step. First, we remove the cross-fusion module between the two branches in the model, making each branch to handle the extraction task independently without interference, and pre-train the model. Then in the second stage, we add the cross-fusion modules to the model, train the parameters in them, and fine-tune other parameters in the pure BSRNN process, which was trained in the first stage. In this way, we can accelerate the training progress and tackle the problem of chaotic gradient descent generated by direct training on too complex models.

Specifically, In each training iteration, the model's two outputs are separately compared with the clean target speech and the interference speech to calculate the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) loss function. The losses from both outputs are weighted and summed before performing gradient descent. After several attempts, we ultimately set the weights of the two losses to 0.6 and 0.4 respectively, to achieve the optimal performance.

In other training configurations, we set the batch size to 8, and the audio sampling rate to 16,000. Both the enrollment speaker embedding and the feature representations of mixed speech are set to a dimension of 256. The enrollment speaker embedding mixing method is vector dot-product. The information cross-fusion method for the two branches in the model is cross-attention mechanism. The speaker encoder of the model is pre-trained on a speaker recognition task first and then fine-tuned jointly with the TSE task. We also compared the performance of multi-task training, where the feature vectors of registered speech undergo a speaker recognition task simultaneously during training, and the calculated cross-entropy loss is added to the gradient descent process. The learning rate during training is set to 0.01, accompanied by a weight decay of 0.0001.

B. Experiment results

We firstly ran experiments on the basic single-path BSRNN model as the baseline, which only extract the target speech and ignore the interfering voices. Then we tried the dual-path structure to conduct a comparison experiment. We also focused on multiple fusion operations between the two branches and construct a Mask-Only Dual-Path Band-Split BSRNN (MO-DP-BSRNN) to perform as another dual-path baseline. The experimental data is explicitly illustrated in Table I.

Our experimental results show that the performance of the DP-BSRNN model without the information cross-fusion mechanism (underlined) is significantly higher than that of the regular BSRNN, demonstrating the feasibility of our idea.

TABLE I

Results of our proposed Dual-Path Band-Split RNN compared with the original BSRNN, note that MO-DP-BSRNN represents the mask-only dual-path mechanism, which only split the model into two branches in the mask generator.

model	spk-embed fuse	dual-path fuse	weight-sharing	validation(tgt)	validation(inf)	inference	param-quantity(M)
BSRNN	multiply	-	-	12.83	-	12.80	265.3
BSRNN	FiLM	-	-	13.01	-	12.63	265.7
MO-DP-BSRNN	multiply	-	×	12.46	12.35	12.31	448.8
DP-BSRNN	multiply	multiply	×	14.00	13.17	13.55	566.5
DP-BSRNN	multiply	attention	×	12.96	12.05	12.78	589.1

The performance of the model using the information cross-fusion mechanism is slightly lower than that of the model without it. We speculate that this is because the cross-fusion method is too simple (in order to reduce the training cost, we only adopted the basic single-layer cross-attention model, which is much simplified compared to previous works using similar mechanisms [15] [16]), we leave this part as a future work. The performance of the model with a dual-path mechanism only when generating masks (labeled as MO-DP-BSRNN) is slightly lower than that of the single-path BSRNN. We consider that this is due to the influence of the multi-task mechanism. The single-path BSRNN is overfit on the extraction work.

IV. Conclusion and future work

We decompose the model into two branches to extract both the speech of the target speaker and the interfering speaker simultaneously, with reference to the registered speech. During this process, we cross-fuse the latent features computed by the two branches of the model. In this way, We view the subsequent computation as a refined further extraction task, aiming to enhance the model's understanding of the task of extracting the target speaker and making the shallow features more aligned with the essence of the extraction operation. Ultimately, our model achieved significant improvements compared to the previous single-branch BSRNN on Libri2Mix dataset.

During the research process, we also identified some problems that deserve further investigation. We leave these questions to future work. They mainly focus on the following aspects:

- In this work, we expected that the cross-fusion of information from two branches could enable the two sub-models to utilize more information, thereby enhancing the overall performance. However, experimental results indicate that although the DP-BSRNN with the cross-fusion mechanism performs slightly better than the regular BSRNN, it is significantly inferior to the model without the cross-fusion mechanism. We speculate that this is due to the too simple cross-fusion module. The information is inadequately refined before fusion, thus disrupt the latent features instead. We will continue to explore different configurations for the cross-fusion module in subsequent research.
- Although we did achieve a noticeable performance improvement in the experiments, our model design also resulted in a doubling of the parameter count. We wish to balance the cost performance by implementing a weight-sharing mechanism for the two branches. We believe that within the deep BSRNN Block, latent features have already been refined, making the tasks of the two branches essentially isomorphic—both are a reverse-extraction operations (given information of one speaker, extract another speaker's voice). This provides feasibility for implementing a weight-sharing mechanism.
- In our model, there are two instances of information fusion. One is between the enrollment speech embedding and the mixed speech embedding, while the other is the fusion of latent features from the two branches. However, the features of the clean enrollment speech, which is under our supervise, are fused with the mixed speech only once at the beginning of the BSRNN, whereas the lack of interpretable intermediate layer outputs from the two branches undergoes multiple cross-fusions. This raises further consideration, which is whether it is feasible to introduce multiple fusions of the enrollment speech while fusing between the two branches, thereby enhancing the model's utilization of the crucial cue.
- One significant starting point in our work is to make the model understand the essence of the target speaker extraction task. We achieve this by letting the model simultaneously perform extraction and denoising tasks on the same input. However, we haven't investigated this aspect in the fusion module (the module where enrollment speech is fused with mixed speech). There is still much research to be done on how to enable the model to correctly utilize registered speech, akin to humans referencing enrollment speech's voice and then extracting content with consistent timbre. There is still much room for exploration in incorporating this operation into the fusion module.

References

[1] Zmolikova K, Delcroix M, Ochiai T, et al. Neural Target Speech Extraction: An overview[J]. IEEE Signal Processing Magazine, 2023, 40(3): 8-29.

[2] Bronkhorst A W. The cocktail-party problem revisited: early processing and selection of multi-talker speech[J]. Attention, Perception, Psychophysics, 2015, 77(5): 1465-1487.

[3] Flanagan J L, Johnston J D, Zahn R, et al. Computer-steered microphone arrays for sound transduction in large rooms[J]. The Journal of the Acoustical Society of America, 1985, 78(5): 1508-1518.

[4] Hershey J, Casey M. Audio-visual sound separation via hidden markov models[J]. Advances in Neural Information Processing Systems, 2001, 14.

[5] Rivet B, Wang W, Naqvi S M, et al. Audiovisual speech source separation: An overview of key methodologies[J]. IEEE Signal Processing Magazine, 2014, 31(3): 125-134.

[6] Wang Q, Muckenhirn H, Wilson K, et al. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking[J]. arXiv preprint arXiv:1810.04826, 2018.

[7] Janský J, Málek J, Čmejla J, et al. Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 676-680.

[8] Xu C, Rao W, Chng E S, et al. Spex: Multi-scale time domain speaker extraction network[J]. IEEE/ACM transactions on audio, speech, and language processing, 2020, 28: 1370-1384.

[9] Ge M, Xu C, Wang L, et al. Spex+: A complete time domain speaker extraction network[J]. arXiv preprint arXiv:2005.04686, 2020.

[10] Ge M, Xu C, Wang L, et al. Multi-stage speaker extraction with utterance and frame-level reference signals[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6109-6113.

[11] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[12] Snyder D, Ghahremani P, Povey D, et al. Deep neural network-based speaker embeddings for end-to-end speaker verification[C]//2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016: 165-170.

[13] Chen J, Rao W, Wang Z, et al. Mc-spex: Towards effective speaker extraction with multi-scale interfusion and conditional speaker modulation[J]. arXiv preprint arXiv:2306.16250, 2023.

[14] Yang L, Liu W, Tan L, et al. Target Speaker Extraction with Ultra-Short Reference Speech by VE-VE Framework[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[15] Zeng B, Suo H, Wan Y, et al. SEF-Net: Speaker Embedding Free Target Speaker Extraction Network[J].

[16] Lin J, Cai X, Dinkel H, et al. Av-Sepformer: Cross-Attention Sepformer for Audio-Visual Target Speaker Extraction[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[17] Luo Y, Yu J. Music Source Separation With Band-Split RNN[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[18] Hao F, Li X, Zheng C. X-Tf-Gridnet: A Time-Frequency Domain Target Speaker Extraction Network with Adaptive Speaker Embedding Fusion[J]. Available at SSRN 4611108.

[19] Zhang Z, He B, Zhang Z. X-tasnet: Robust and accurate time-domain speaker extraction network[J]. arXiv preprint arXiv:2010.12766, 2020.

[20] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.