2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389752

EXPLORING TIME-FREQUENCY DOMAIN TARGET SPEAKER EXTRACTION FOR CAUSAL AND NON-CAUSAL PROCESSING

Wangyou Zhang[1], *Lei Yang*[2], *Yanmin Qian*[1†]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]Samsung Research China - Beijing (SRC-B)

ABSTRACT

In recent years, target speaker extraction (TSE) has drawn increasing interest as an alternative to speech separation in realistic applications. While time-domain methods have been widely used in recent studies to achieve high performance, the potential of time-frequency (T-F) domain methods have been less explored. In this paper, we try to fill this gap and propose to incorporate the top-performing T-F domain speech separation method into the TSE framework. We first explore different speaker information fusion methods for the proposed model. In addition to the commonly-used concatenation-based fusion, we propose a novel speaker token-based fusion method to fuse the target speaker information. Second, we show that the proposed model can be easily extended for causal processing with strong performance. Experiments on the WSJ0-2mix and LibriMix benchmarks show that our proposed model outperforms the widely-used time-domain models in both causal and non-causal settings by a large margin.

Index Terms— Target speaker extraction, speech separation, time-frequency domain, dual-path modeling

1. INTRODUCTION

Humans are known to have the capability of listening to, following and recognizing one target speaker in the presence of interference speakers and background noise. This process is usually termed the cocktail party problem [1, 2], a well-known and important problem proposed by Cherry in his famous paper [1]. More specifically, the task of separating all speech sources in a mixed signal is known as speech separation. With the advances in deep learning, two representative methods have been developed and widely used to solve the permutation problem, i.e., deep clustering [3] and permutation invariant training [4]. On the other hand, we may be interested in only extracting a specific speaker in the mixture instead of separating all speech signals. This task is called target speaker extraction (TSE), where an additional clue indicating the identity of the target speaker is provided. In most existing works, the clue is often a reference speech signal (or the enrollment) spoken by the target speaker.

In recent years, more and more TSE methods have been proposed, starting with time-frequency (T-F) domain methods [5–10]. Later, inspired by the success of time-domain models in speech separation [11–14], similar structures are also adopted in time-domain TSE methods [15–22], which have shown superior performance over T-F domain methods due to their strong modeling capability. While time-domain models are favored in most recent TSE works, it has been shown [23] that the small kernel size in the time-domain encoder may be detrimental to the performance in some realistic conditions with reverberation. On the other hand, recently proposed T-F domain models [24, 25] have shown very promising performance on the speech separation task, but their effectiveness on the target speaker extraction task has not been explored.

In this paper, we aim to revisit the T-F domain model for TSE and show that it can achieve state-of-the-art performance in both non-causal and causal settings. We propose a dual-path target speaker extraction network that operates in the T-F domain. Different from the dual-path time-domain models [12–14] with inter-chunk and intra-chunk modeling, our proposed model defines the two paths along the temporal and frequency dimensions in the short-time Fourier transform (STFT) spectrum. We further explore two different fusion approaches to incorporate the target speaker information into the model. One straightforward method is to concatenate the speaker features in a frame-wise manner. In addition, we also propose a novel speaker token-based fusion method that naturally fits our proposed model. Furthermore, we show that the proposed model can be easily configured to operate in either causal or non-causal mode. To evaluate the performance of the proposed model, we conduct experiments on two widely-used benchmarks, i.e., WSJ0-2mix [3] and LibriMix [26]. The experimental results show that our proposed model surpasses the state-of-the-art time-domain TSE approaches in both causal and non-causal settings, with over 1 dB signal-to-distortion ratio (SDR) improvement in all settings.

[†] Yanmin Qian is the corresponding author.

979-8-3503-0689-7/23/$31.00 ©2023 IEEE

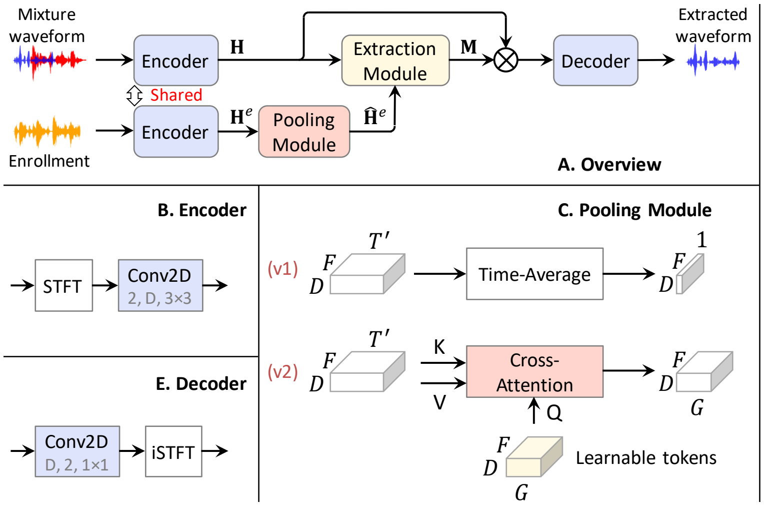


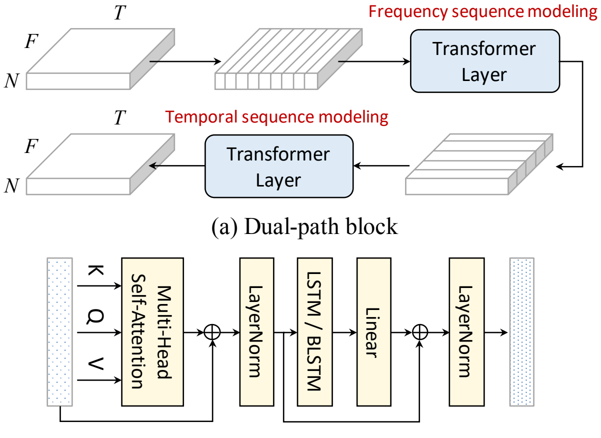
Fig. 1: Overview of the proposed T-F domain target speaker extraction model.

2. T-F DOMAIN TARGET SPEAKER EXTRACTION

2.1. Overview

We base our proposed approach on a recently proposed dual-path network called time-frequency domain path scanning network (TFPSNet) [24], which is one of the top-performing speech separation models. The overview of the proposed model is depicted in Fig. 1, where the feature maps and kernel sizes of convolutional layers are annotated in gray. It consists of an encoder, a pooling module, an extraction module, and a decoder.

The encoder first extracts the complex-valued STFT spectrum from the input waveform, where the real and imaginary parts are stacked as the channel dimension. Each T-F bin in the spectrum is then projected to a D-dimensional embedding via a subsequent 2D convolutional layer with a 3×3 kernel[1]. The resultant representations are more flexible than the original STFT spectrum, while enjoying the robustness of STFT in different conditions. This encoder is shared among the mixture and enrollment waveforms, and the generated representations are denoted as $\mathbf{H} \in \mathbb{R}^{D\times F\times T}$ and $\mathbf{H}^e \in \mathbb{R}^{D\times F\times T'}$, respectively. Here, T and T' are the number of frames, and F is the number of frequency bins. **The pooling module** generates a fixed-length speaker representation $\hat{\mathbf{H}}^e$ from the variable-length enrollment feature $\mathbf{H}^e$, which will be introduced in Section 2.3. **The extraction module** takes as input the mixture feature $\mathbf{H}$ and the speaker representation $\hat{\mathbf{H}}^e$, and produces a high-dimensional mask $\mathbf{M}^{D\times F\times T}$. The estimated mask is then element-wisely multiplied with the mixture feature $\mathbf{H}$ to extract the target speaker's speech. The extracted feature is finally processed by **the decoder**. It first projects the feature to the STFT spectrum via a point-wise convolutional layer, and then transforms the spectrum to the waveform via inverse STFT (iSTFT).


T
F
N
Frequency sequence modeling
Transformer Layer
Temporal sequence modeling
Transformer Layer
K
Q
V
Multi-Head Self-Attention
LayerNorm
LSTM / BLSTM
Linear
LayerNorm

(a) Dual-path block

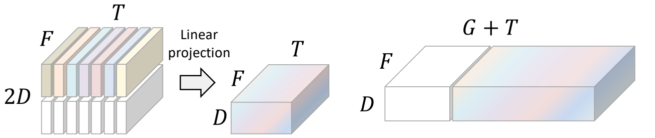
(b) Transformer layer

Fig. 2: Architecture of the dual-path block.

2.2. T-F domain dual-path modeling

The extraction module is the core component of the proposed model, which extracts the target speech via dual-path modeling in the T-F domain. Different from the original dual-path modeling in the time-domain approaches [12–14], we consider the frequency and temporal dimensions in the encoded feature as two different paths. The architecture of the extraction module is illustrated in Fig. 1-D. The input feature is first processed by channel-wise layer normalization and projected to a bottleneck dimension N via a point-wise convolutional layer. The bottleneck feature is then processed by K consecutive dual-path blocks for fine-grained frequency and temporal sequence modeling. As shown in Fig. 2 (a), the dual-path block has a similar structure to [24], consisting of a transformer layer for frequency sequence modeling and another for temporal sequence modeling. However, it should be noted that we replace all T-F path modeling (along the anti-diagonal direction) proposed in [24] with the temporal sequence modeling, which simplifies the structure while preserving strong performance. Each transformer layer has the

[1] We zero-pad the feature in a causal manner (only padding to the left) along the time dimension before convolution.



(v1) Frame-wise concatenation (v2) Speaker token-based fusion

Fig. 3: Two different fusion methods for incorporating the speaker information (denoted by the white blocks).

same architecture as in [13, 24], which is depicted in Fig. 2 (b). The output of the final dual-path block is projected back to D-dimensional by another point-wise convolutional layer. Following [24] and [13], we adopt a two-way mask estimation structure, where two parallel point-wise convolutional layers followed by `Tanh` and `Sigmoid` activations respectively are used. The outputs from the two parallel streams are multiplied and further processed by the `Tanh` activation to produce the final mask $\mathbf{M}$ for target speaker extraction.

2.3. Speaker information fusion

Following the design in SpeakerBeam [7, 16], we adopt a speaker encoder module that is jointly trained with the extraction module. This allows us to train the TSE model without using external data for speaker recognition. To extract the speaker information from the enrollment, we first use the same STFT-based encoder to extract hidden features $\mathbf{H}^e$. In practice, the enrollment length is often variable and may not be equal to the length of the mixture waveform. As a result, the corresponding hidden feature $\mathbf{H}^e$ will also have a different number of frames. In order to obtain a fixed-dimension representation for speaker information fusion, we introduce a pooling module after the encoder.

As shown in Fig. 1-C, we consider two kinds of pooling methods, each corresponding to a different fusion method. The first pooling approach (denoted as `v1`) is the commonly-used averaging operation along the time dimension. Thus, the resultant representation $\hat{\mathbf{H}}^e$ has a fixed shape of $D \times F \times 1$. The second pooling approach (denoted as `v2`) is inspired by the Perceiver IO [27], where an arbitrary input dimension can be mapped to a fixed dimension via cross-attention with a latent query. We devise a group of learnable speaker tokens of shape $D \times F \times G$ in the latent space, where G is the group size. These speaker tokens are used as a query to conduct cross-attention with the input speaker feature $\mathbf{H}^e$. The output speaker tokens are then of a fixed shape $D \times F \times G$. The intuition of the second approach is that more fine-grained speaker representations can be learned compared to simply averaging along the time dimension.

Subsequently, the pooled speaker representation will be incorporated in the extraction module using the corresponding fusion method. For the first method (`v1`), we follow the widely-used frame-wise concatenation-based fusion approach. The time-averaged representation $\hat{\mathbf{H}}^e$ is repeated T times to match the length of the mixture feature and concatenated frame-wisely to the inputs of the first $K-1$ dual-path modules. As shown in Fig. 3 (v1), the concatenated feature is projected back to D-dimensional via a linear layer. For the second method (`v2`), we take inspiration from the memory transformer [28] and rely on the transformer layers in the dual-path blocks to exploit the speaker information implicitly. We concatenate the speaker tokens and the mixture feature along the time dimension right before the first dual-path blocks, resulting in a feature of shape $D \times F \times (G+T)$. This feature is then fed into K dual-path blocks for processing. The first G frames in the output will be discarded, and the rest T frames correspond to the extracted target-speaker representation. This novel fusion approach has not been explored before and it fully leverages the sequence modeling capability of transformer layers to utilize the speaker information.

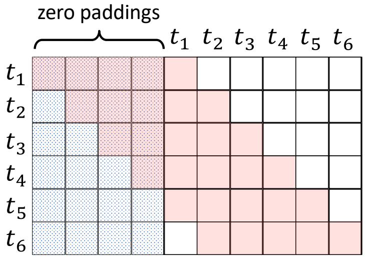


Fig. 4: Time-restricted attention with only lookbacks.

2.4. Extension for causal processing

In order to extend the proposed model for causal processing, we make two modifications only to the transformer layers (in each dual-path block) for temporal sequence modeling. (1) We replace all bidirectional long short-term memory (BLSTM) with unidirectional LSTM. (2) We replace the global multi-head self-attention (MHA) with time-restricted MHA [29, 30], where only a fixed number of history samples are attended to, as shown in Fig. 4.

3. EXPERIMENTAL SETUP

3.1. Data preparation

To evaluate the TSE performance of our proposed model, we conduct experiments on the commonly-used WSJ0-2mix [3] and LibriMix [26] two-speaker mixture datasets. We adopt the `min` versions of both datasets for training. In WSJ0-2mix, the training, development, and evaluation subsets contain 20000, 5000, and 3000 mixture samples, respectively. The signal-to-interference ratio (SIR) in each clean mixture ranges from -10dB to 10dB. In LibriMix, the training, development, and evaluation subsets contain 64700[2], 3000, and 3000 mixture samples, respectively. The signal-to-interference ratios (SIRs) in the mixtures are normally distributed with a mean of 0 dB and a standard deviation of 7 dB. A random noise from the WHAM! [33] corpus is added to each sample. The signal-to-noise ratios (SNRs) are normally distributed with a mean of -2 dB and a standard deviation of 4 dB. For experiments on WSJ0-2mix, we use the exiting enrollment lists[3] for all subsets. For experiments on LibriMix, we randomly select

[2] The `train-100` subset only contains 13900 mixture samples.

[3] https://github.com/gemengtju/SpEx_Plus

Table 1: Evaluation of the proposed methods on the WSJ0-2mix test set (`min` version). WER is evaluated on the `max` version. * denotes the reproduced result.

Model	Att. Lookback	PESQ ↑	STOI (×100) ↑	SI-SNR (dB) ↑	SDR (dB) ↑	WER (%) ↓	#Params (M)	#MACs (G/s)
Original mixture	-	2.01	73.80	0.0	0.2	62.4	-	-
				Non-causal setting				
TD-SpeakerBeam [16]*	-	3.46	96.35	17.1	17.5	10.4	16.21	13.52
SpEx+ [18]	-	-	-	17.4	17.4	-	11.1	-
SpEx$_{\text{pc}}$ [31]	-	-	-	19.0	19.0	-	28.4	-
X-SepFormer [21]	-	3.75	-	19.1	19.7	-	-	-
+ Data augment. [21]	-	3.80	-	19.5	20.1	-	-	-
VEVEN [22]	-	-	-	19.0	19.0	-	2.6	-
Proposed (`v1`)	-	**3.89**	**98.07**	**20.7**	**21.1**	10.3	3.48	27.11
Proposed (`v2`, $G = 1$)	-			cannot converge			3.46	26.93
Proposed (`v2`, $G = 5$)	-	3.86	97.70	20.5	21.0	10.6	3.50	27.33
Proposed (`v2`, $G = 10$)	-	3.86	97.79	20.5	20.9	10.6	3.54	27.84
Proposed (`v2`, $G = 20$)	-	3.87	97.80	20.6	21.0	**9.8**	3.62	28.85
				Causal setting				
SkiM + LCC SISO [32]	-	3.13	95.19	15.5	15.9	-	9.3	9.6
Proposed (`v1`)	5 frames	**3.57**	**96.71**	**17.5**	**17.9**	11.9	2.84	20.78
Proposed (`v1`)	10 frames	3.56	96.64	17.3	17.7	**11.6**	2.84	20.81
Proposed (`v1`)	20 frames	**3.57**	**96.71**	**17.5**	**17.9**	12.4	2.84	20.87
→ $D = 64, N = 32$	20 frames	3.43	95.89	16.2	16.7	15.0	2.00	14.59
Proposed (`v1`)	40 frames	**3.57**	96.59	**17.5**	**17.9**	**11.6**	2.84	20.98

an utterance from Librispeech [34] that is spoken by the same speaker as the enrollment for each speaker in each mixture sample. The enrollment selection is fixed for the development and evaluation subsets[4], while it is dynamically sampled on the fly for the training subset. All mixture samples in the training data are chopped into 4-sec segments, while the enrollments are randomly chopped into 2-sec segments. The sampling rate of all data is 8 kHz in our experiments.

3.2. Model and training configurations

In all experiments, our proposed model consists of $K = 6$ dual-path blocks. The window and hop sizes for STFT/iSTFT are 256 and 128, respectively. The resultant STFT spectrum has 129 frequency bins in each frame. Following TF-PSNet [24], we by default set the embedding and bottleneck dimensions to $D = 256$ and $N = 64$, respectively. The transformer layers have the same configuration as in [13, 24], with 4 attention heads and a cell state dimension of 128 in the BLSTM / LSTM layer. When the first fusion approach (`v1`) in Section 2.3 is used, an additional linear projection layer is added to the first 5 dual-path blocks, as shown in Fig. 3 (`v1`). When the second fusion approach (`v2`) is used, a 4-head cross-attention layer is adopted in the pooling module. We evaluate the effect of different group sizes $G \in \{1, 5, 10, 20\}$ for the learnable speaker tokens. For the causal setting, we compare the performance of using different numbers of lookback frames (5, 10, 20, 40) in the time-restricted attention described in Section 2.4.

Our models are built based on the ESPnet toolkit [35]. All models are trained using the scale-invariant signal-to-noise ratio (SI-SNR) [36] loss. The non-causal (causal) models are trained up to 100 (130) epochs[5] using the Adam optimizer, while an early stop will be triggered if the loss is not reduced for 20 epochs on the development set. The learning rate increases linearly in the first X steps to `4e-4`, and then decreases by a factor of 0.98 after each epoch. We set X to 4000 and 8000 for experiments on WSJ0-2mix and LibriMix, respectively. The batch size is set to 4 and 8 on WSJ0-2mix and LibriMix, respectively.

4. EXPERIMENTAL RESULTS

We evaluate the performance of TSE models with several objective measures, including SDR [37], SI-SNR [36], PESQ [38], and short-time objective intelligibility (STOI) [39]. In addition, we evaluate the ASR performance (word error rate, WER) on the 8kHz[6] `max` version of each dataset using the open-source Whisper Large v2 model[7] [40] without external language models. The number of multiply–accumulate operations (MACs) of our models is computed using the `thop` Python package[8] on a 4-sec mixture sample with a 2-sec enrollment. Note that the original `thop` toolkit does not count the computation in the multi-head self-attention module, so we modify the script to take it into account.

4.1. Performance evaluation on WSJ0-2mix

In this section, we evaluate the performance of the proposed methods on the WSJ0-2mix benchmark, where the average enrollment length is 7.35s. We first compare our models with existing top-performing time-domain TSE methods in the non-causal setting in Table 1. It can be observed that our proposed models outperform existing TSE models by a large margin, with more than 1 dB SI-SNR and SDR improvement. It is worth noting that our models do not apply data augmentation, multi-stage training, or complicated loss combinations. In contrast, the SpEx+ [18] and SpEx$_{\text{pc}}$ [31] models use an additional speaker recognition loss to boost the performance, which requires speaker labels in the training data. The X-SepFormer [21] models adopt a two-stage training strategy to

[4] The enrollment lists are taken from https://github.com/BUTSpeechFIT/speakerbeam.

[5] For the experiment on LibriMix (`train-100` + `train-360`), we only train the model for 40 epochs as it is enough for convergence.

[6] The audios are upsampled to 16kHz before fed into the Whisper model.

[7] https://huggingface.co/openai/whisper-large-v2

[8] https://github.com/Lyken17/pytorch-OpCounter

Table 2: Evaluation of non-causal models on the noisy LibriMix test set (`min` version). WER is evaluated on the `max` version.

Model	**PESQ** ↑	**STOI** (×100) ↑	**SI-SNR (dB)** ↑	**SDR (dB)** ↑	**WER (%)** ↓	**#Params (M)**	**#MACs (G/s)**
Original mixture	1.44	64.48	-2.0	-1.8	73.9	-	-
TSE-V (`train-100`) [20]	-	-	-	10.6	-	-	-
TSE-V (`train-360`) [20]	-	-	-	11.8	-	-	-
Proposed (`v1`, `train-100`)	2.86	88.76	11.6	12.8	14.3	3.48	27.11
+ `train-360`	**2.93**	**90.06**	**12.5**	**14.4**	**12.0**	3.48	27.11

apply multiple loss schemes. The VEVEN [22] model needs to apply a multi-stage training procedure to make the model converge. Comparing different fusion methods in the proposed method, we can see that both fusion methods can work well with similar performance. For the second fusion method (`v2`), it is shown that more speaker tokens generally lead to better performance, although the relative gain is not large. It is worth noting that this novel method has the potential for further extension to achieve more functions at the same time, such as long sequence modeling as proposed in [28]. We leave the exploration of this possibility for future work.

Then, we evaluate the performance of the proposed models in the causal setting. Since it is not straightforward to use the second fusion approach (`v2`) for causal processing, we only evaluate the first fusion approach (`v1`) here. We compare the performance with the state-of-the-art causal speech separation model [32]. We can observe that the proposed models with different lookback lengths in the attention module all show better performance than the existing causal model. The overall computational cost (#MACs) is also much lower than the non-causal setting. When decreasing the lookback length, the TSE performance does not change much. This is attributed to the $K = 6$ stacked dual-path blocks that increases the effective reception field in the attention module by 6 times. The top layer can thus have a reception field of at least 30 frames (~500 ms). And the LSTM layer after each attention module (Fig. 1) further increases the history information the model can access. We further investigate the performance of reducing hidden dimensions in the model with the lookback length of 20 frames. It is not surprising to see that the performance degrades compared to the default configuration, while both model size and computational costs are greatly reduced. Nevertheless, the performance is still better than the existing causal method.

Finally, we evaluate the ASR performance for both settings using the same Whisper model[9]. All our models achieve good ASR performance, while the model with $G = 20$ speaker tokens achieves the lowest WER. We can also see that a better TSE performance does not always lead to better ASR results. This is a commonly observed phenomenon [41] caused by the artifacts introduced by the TSE frontend.

4.2. Performance evaluation on LibriMix

In this section, we evaluate the performance of the proposed model with the first fusion method (`v1`) on the noisy LibriMix data, which is more difficult due to the background noise. Compared to the reported results of the time-domain TSE method in [20], our proposed models achieves much better SDR performance, with over 2 dB improvement when trained with the same amount of data. Further increasing the training data by including the `train-360` subset results in better TSE and ASR performances, which demonstrates the capacity of the proposed model.

5. CONCLUSION

In this paper, we revisit the time-frequency domain approach for target speaker extraction. We propose a dual-path T-F domain TSE model inspired by the recent advances in speech separation. In addition, we explore two different fusion approaches to incorporate the target speaker information, e.g., concatenation-based fusion and speaker token-based fusion. Finally, we show that the proposed model can be easily extended to a causal setting with competitive performance. We evaluate the proposed model in the widely-used WSJ0-2mix and LibriMix benchmarks. Experimental results show that our proposed model outperforms the state-of-the-art time-domain method in both non-causal and causal settings. In future work, we would like to investigate the performance of the proposed model on more realistic conditions.

ACKNOWLEDGMENTS: This work was supported in part by China STI 2030-Major Projects under Grant No. 2021ZD0201500, in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102. The experiments were done using the PI supercomputer at Shanghai Jiao Tong University.

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953. 1

[2] Y. Qian *et al.*, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018. 1

[3] J. R. Hershey *et al.*, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, 2016, pp. 31–35. 1, 3

[4] D. Yu *et al.*, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017, pp. 241–245. 1

[9] We apply text normalization [40] to both decoding outputs and reference labels before calculating the WER.

[5] J. Wang *et al.*, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Interspeech*, 2018, pp. 307–311. 1

[6] X. Xiao *et al.*, "Single-channel speech extraction using speaker inventory and attention network," in *ICASSP*, 2019, pp. 86–90. 1

[7] K. Žmolíková *et al.*, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019. 1, 3

[8] Q. Wang *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Interspeech*, 2019, pp. 2728–2732. 1

[9] S. He *et al.*, "Speakerfilter: Deep learning-based target speaker extraction using anchor speech," in *ICASSP*, 2020, pp. 376–380. 1

[10] R. Giri *et al.*, "Personalized PercepNet: Real-time, low-complexity target voice separation and enhancement," in *Interspeech*, 2021, pp. 1124–1128. 1

[11] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. ASLP.*, vol. 27, no. 8, pp. 1256–1266, 2019. 1

[12] Y. Luo *et al.*, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*, 2020, pp. 46–50. 1, 2

[13] J. Chen *et al.*, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Interspeech*, 2020, pp. 2642–2646. 1, 2, 3, 4

[14] C. Subakan *et al.*, "Attention is all you need in speech separation," in *ICASSP*, 2021, pp. 21–25. 1, 2

[15] C. Xu *et al.*, "Time-domain speaker extraction network," in *Proc. IEEE ASRU*, 2019, pp. 327–334. 1

[16] M. Delcroix *et al.*, "Improving speaker discrimination of target speech extraction with time-domain Speaker-Beam," in *ICASSP*, 2020, pp. 691–695. 1, 3, 4

[17] C. Xu *et al.*, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1370–1384, 2020. 1

[18] M. Ge *et al.*, "SpEx+: A complete time domain speaker extraction network," in *Interspeech*, 2020, pp. 1406–1410. 1, 4

[19] Y. Hao *et al.*, "A unified framework for low-latency speaker extraction in cocktail party environments," in *Interspeech*, 2020, pp. 1431–1435. 1

[20] M. Delcroix *et al.*, "Listen only to me! how well can target speech extraction handle false alarms?" in *Interspeech*, 2022, pp. 216–220. 1, 5

[21] K. Liu *et al.*, "X-SepFormer: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *ICASSP*, 2023, pp. 1–5. 1, 4

[22] L. Yang *et al.*, "Target speaker extraction with ultra-short reference speech by VE-VE framework," in *ICASSP*, 2023, pp. 1–5. 1, 4, 5

[23] T. Cord-Landwehr *et al.*, "Monaural source separation: From anechoic to reverberant environments," in *Proc. IWAENC*, 2022, pp. 1–5. 1

[24] L. Yang *et al.*, "TFPSNet: Time-frequency domain path scanning network for speech separation," in *ICASSP*, 2022, pp. 6842–6846. 1, 2, 3, 4

[25] Z.-Q. Wang *et al.*, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *ICASSP*, 2023, pp. 1–5. 1

[26] J. Cosentino *et al.*, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020. 1, 3

[27] A. Jaegle *et al.*, "Perceiver IO: A general architecture for structured inputs & outputs," in *Proc. ICLR*, 2022. 3

[28] M. S. Burtsev *et al.*, "Memory transformer," *arXiv preprint arXiv:2006.11527*, 2020. 3, 5

[29] D. Povey *et al.*, "A time-restricted self-attention layer for ASR," in *ICASSP*, 2018, pp. 5874–5878. 3

[30] X. Chang *et al.*, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP*, 2020, pp. 6129–6133. 3

[31] W. Wang *et al.*, "Neural speaker extraction with speaker-speech cross-attention network," in *Interspeech*, 2021, pp. 3535–3539. 4

[32] C. Li *et al.*, "Predictive SkiM: Contrastive predictive coding for low-latency online speech separation," in *ICASSP*, 2023, pp. 1–5. 4, 5

[33] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Interspeech*, 2019, pp. 1368–1372. 3

[34] V. Panayotov *et al.*, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210. 4

[35] C. Li *et al.*, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE SLT*, 2021, pp. 785–792. 4

[36] J. Le Roux *et al.*, "SDR—half-baked or well done?" in *ICASSP*, 2019, pp. 626–630. 4

[37] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP.*, vol. 14, no. 4, pp. 1462–1469, 2006. 4

[38] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, 2001, pp. 749–752. 4

[39] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. ASLP.*, vol. 19, no. 7, pp. 2125–2136, 2011. 4

[40] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022. 4, 5

[41] H. Sato *et al.*, "Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition," in *Interspeech*, 2021, pp. 1149–1153. 5