ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11465070

LEXTRA: FOLDED PROMPT AND SPLIT-ROLE ATTENTION FOR TARGET SPEAKER EXTRACTION

Pengjie Shen[1,2*], *Shulin He*[2], *Xueliang Zhang*[1], *and Zhong-Qiu Wang*[2]

[1]Department of Computer Science, Inner Mongolia University, Hohhot, China
[2]Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China

shenpengjie@mail.imu.edu.cn, cszxl@imu.edu.cn, {he.shulin96,wang.zhongqiu41}@gmail.com

ABSTRACT

Target speaker extraction (TSE) aims at isolating a desired target speaker from a mixture of multiple speakers, based on a short enrollment utterance of the target speaker. In LExt, a recent onset-prompted TSE method obtaining state-of-the-art performance, it is observed that longer prompts often help but the encoder and separator have to process longer sequences, which increases computation, and the performance often drops when the enrollment is clean while the mixture is noisy-reverberant. To address these issues, we propose LExTra, an extension of LExt which combines folded prompt conditioning with split-role multi-head attention. The folded prompt divides an enrollment into equal-length sub-segments, prepends each sub-segment to a copy of the mixture, and stacks the prepended mixture along a pseudo channel axis. This way, the backbone of the TSE model sees an enrollment utterance with shorter time span while still capable of exploiting the full enrollment. Split-role attention assigns some attention heads to the cross-attention from mixture to enrollment to extract speaker cues, and assigns the remaining ones to mixture self-attention to preserve stream coherence. We find that this mechanism can improve the robustness of TSE when there are acoustic mismatches between the enrollment and mixture. Evaluation results on the WSJ0-2mix and WHAMR! datasets show the effectiveness of LExTra.

***Index Terms*—** Target speaker extraction, onset-prompted speech separation

1. INTRODUCTION

In multi-speaker scenarios, such as meetings, conversational settings, and public venues, it is often needed to isolate the speech of a single targeted speaker from a recording containing concurrent speech by multiple speakers. Target speaker extraction (TSE) [1–5] addresses this task: given a short enrollment utterance from the targeted speaker, the system extracts that speaker's speech from a mixture containing interfering speech signals. Compared to speech separation (SS), which blindly separates all sources [6–13], TSE gets rid of the well-known permutation ambiguity problem [8,14] by directly conditioning on the speaker characteristics of the targeted speaker to extract the speech of the targeted speaker.

However, TSE is typically more challenging than SS, since enrollment utterances and mixtures may come from different acoustic environments (e.g., clean vs. noisy and reverberant), and extracted signals may suffer from speaker confusion errors [15, 16] when the interferers share similar speech characteristics with the target speaker. These difficulties make robust TSE a particularly demanding and challenging research problem, despite its advantages over conventional separation.

A widely-adopted TSE strategy is to encode the enrollment utterance into a fixed-length speaker embedding, most commonly an x-vector [17] or a d-vector [18], and to fuse this embedding with mixture features to steer the TSE model [15,19–22]. The fusion may be performed via early concatenation with mixture features [15], gating or adaptive-filtering mechanisms that modulate mixture features [2,19], or a more recent attention-based scheme [23,24]. Methods driven by speaker embeddings are appealing for their modularity: the speaker embedding is computed once for each speaker and the TSE system can be reused for rapid adaptation to new speakers. Recent overviews [25–27] summarize several weaknesses when the entire enrollment is compressed into a single, fixed-length vector and then fused with mixture features. First, fixed-length embeddings discard temporal structure within the enrollment and can limit the granularity of speaker cues, which motivates stronger feature-level interactions beyond a single vector [25–27]. Second, the effectiveness of the embedding highly depends on the quality and duration of the enrollment. Short or low-SNR enrollments degrade the reliability of the extracted speaker embeddings [25, 26]. Third, when the enrollment and mixture are recorded in acoustically-different conditions, the fusion stage can propagate domain mismatches and even let noise or reverberation leak into the conditioning pathway, which hurts extraction and increases speaker confusion [25]. These observations have motivated designs that keep richer interactions between enrollment and mixture features or that explicitly account for acoustic mismatches.

To address these challenges, a line of research has explored moving away from fixed-length speaker embeddings, instead leveraging variable-length speaker embeddings [28–31]. In particular, Xiao *et al.* [28] proposed a method that first extracts embedding sequences from both the mixture and the enrollment utterance. A cross-attention mechanism is then applied, where each embedding in the mixture sequence serves as a query, and the embeddings from the enrollment sequence act as keys and values. This process produces a new embedding sequence that, by design, has the same length as the original mixture embedding sequence. More recently, LExt [32] introduced a prompt-based approach, which prepends the

*This work was done while P. Shen was a visiting student at SUSTech. This research was supported by Inner Mongolia Natural Science Foundation (Grant No. 2025LHMS06005), CCF-Lenovo Research Fund (Grant No. 20240203), and National Key Research and Development Program of China (Grant No. 2025YFF0518003). *Corresponding author: Zhong-Qiu Wang*.

enrollment to the mixture and trains a strong deep neural network (DNN) to predict the target speech based on the resulting mixture. The prepended enrollment could act as a prompt to help the DNN identify and reconstruct the target speaker. Although LExt is very simple, it achieves strong TSE performance.

Although being simple and effective, LExt has two limitations. First, longer prompts usually carry richer target cues and often improve extraction performance. However, prepending the prompt to mixture forces the encoder and sequence modeling blocks of the DNN to process more frames. This increases computation and memory, and it can raise end-to-end latency because the whole prompt plus mixture must traverse the DNN backbone before decoding. In models with temporal attention, the prepended frames also expand the attention window and further increase the computational load. In short, the gains from longer prompts come with a proportional computational cost. Second, in many applications the enrollment is recorded in a clean condition with little or no reverberation and noise, while the mixture contains background noise, interfering speakers, and room reverberation. Using a clean enrollment segment as the prompt introduces a domain mismatch with the noisy-reverberant mixture, which weakens the reliability of the conditioning signal and may increase speaker confusion. This mismatch is typical in benchmarks that explicitly add noise and reverberation to conversational speech, for example in WHAMR! [33]. Therefore, improving robustness to such mismatches is essential for prompt-based TSE methods.

To address the limitations above, this paper introduces two complementary techniques. First, long prompts improve target conditioning but increase computation because they extend the input sequence. We propose a folded prompt scheme: a long onset prompt is divided into short segments with equal length, each segment is prepended to a copy of the mixture and encoded in parallel, and the resulting segment features are then concatenated along a pseudo channel dimension before the separator. This preserves full enrollment coverage and keeps the temporal length seen by the backbone nearly unchanged, which reduces attention cost and end-to-end latency compared with using one long prompt. To handle acoustic mismatches between enrollment and mixture, we design a split-role attention block with two pathways: a speaker-aware pathway letting mixture features attend to enrollment features to extract target-specific cues and a context-aware pathway performing self-attention within the mixture to model internal structure and temporal dependencies. By separating identity modeling from stream coherence modeling, the block avoids competition between objectives and improves robustness under noisy-reverberant conditions, yielding more consistent TSE.

2. METHOD

In TSE, the objective is to extract the speech of a targeted speaker from a mixture of multiple overlapping speakers. The observed time-domain mixture signal y can be formulated as

$$y = s + v \in \mathbb{R}^N, \tag{1}$$

where N is the number of time-domain samples, and s and v are respectively the target speech and non-target signals. The non-target signals include environmental noise, room reverberation and competing speakers. To assist the extraction of the target speaker's speech, an enrollment utterance $e \in \mathbb{R}^E$ is assumed provided. This utterance is spoken by the same speaker as the target speech s, but contains different content. The goal of TSE is to estimate the clean target speech s based on y by leveraging the speaker-specific information embedded in the enrollment utterance e.

In [32], an onset prompt strategy named LExt is proposed for TSE. Instead of using a separate speaker embedding module, it concatenates enrollment and mixture signals in the time domain, and trains a DNN to jointly model the speaker information in the enrollment and the spectro-temporal patterns in the mixture to perform TSE. Compared with approaches such as USEF-TSE [31] that rely on explicit cross-attention between mixture and enrollment, LExt employs a cleaner design without additional speaker conditioning modules. It has been evaluated on standard datasets including WSJ0-2mix [8], WHAM! [34] and WHAMR! [33], achieving state-of-the-art TSE performance.

Our proposed system builds opon the onset prompt design in LExt [32]. The input consists of the mixture waveform y prepended with an onset prompt sampled from the enrollment utterance e. These signals are concatenated along time axis in the time domain, and then processed by a DNN. The DNN outputs an estimate $\hat{s}$ in the time-frequency (T-F) domain, which is transformed back to time domain to obtain the final extracted signal. This design allows the model to use both the prompt guidance and the detailed speaker information carried by the enrollment. In the rest of this section, we describe two techniques we propose to improve LExt: folded prompt and split-role multi-head attention. See Fig. 1 for an overview.

2.1. Folded Prompt Conditioning

In LExt [32], a fixed enrollment window of length T_0 is taken from e and is prepended to the mixture signal y. In the proposed folded scheme, we keep the same total enrollment window T_0, but divide it into P equal sub-segments and feed them in parallel as pseudo channels to the TSE model:

$$e' = \left[e^{(1)},\, e^{(2)},\, \ldots,\, e^{(P)}\right], \text{ where } e^{(i)} \in \mathbb{R}^T \text{ with } T = \tfrac{T_0}{P}. \tag{2}$$

Each sub-segment is concatenated with the same mixture y (replicated across all the P prompt channels), using the same glue signal g, which marks the boundary between the enrollment and mixture signals, following LExt [32]. That is,

$$x^{(i)} = \left[e^{(i)},\, g,\, y\right] \in \mathbb{R}^{T+G+N}, \text{ for } i = 1, \ldots, P. \tag{3}$$

Notice that the prompt-channels reuse the identical mono mixture y and differ only in the enrollment sub-segment. They are introduced as input to the TSE model for speaker conditioning:

$$X_{\text{in}} = \text{Stack}\left(x^{(1)}, \ldots, x^{(P)}\right) \in \mathbb{R}^{P \times (T+G+N)}. \tag{4}$$

Through this mechanism, the input signal length to the DNN model is reduced from $T_0 + G + N$ in LExt to $T + G + N$ in the proposed method. Although the number of input channels is increased from 1 to P, the increased amount of computation spent on processing multiple input channels can be designed very small. For example, in modern speech separation models such as TF-GridNet [35, 36], the encoder embeds multi-channel input signals to D-dimensional features in its first layer through a 2D convolutional (Conv2D) layer, and the increased amount of computation spent in this layer is negligible compared with that on the other layers.

Compared with prepending a long prompt of length T_0 to the mixture, the folded prompt reduces the length of the prepended signal to $T{=}T_0/P$, but still exposes the full T_0 seconds of enrollment through the P channels. In LExt [32], it is observed that setting the

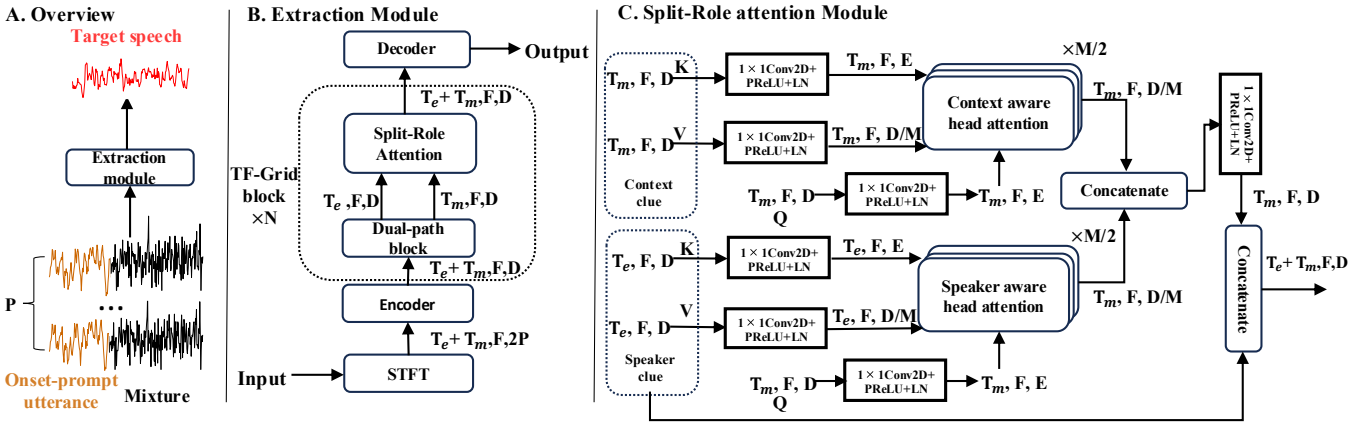

Fig. 1: LExTra with folded prompt and split-role attention. (a) System overview: the onset-prompt (enrollment+glue) is folded into P pseudo-channels and concatenated with the mixture. (b) Extraction module, T_e and T_m denote the STFT frame counts of the onset-prompt and the mixture, respectively. (c) Split-role attention with M heads partitioned equally into speaker-aware and context-aware groups.

enrollment segment length to 4 seconds produces the best TSE performance. With $T_0 = 4$, in the proposed system we experiment with splitting it into two 2-second sub-segments (i.e., P=2), concatenating each with the same mixture, and stacking them along the channel dimension for subsequent processing.

2.2. Split-Role Multi-Head Attention

Following LExt [32], we use TF-GridNet [35, 36] as the DNN backbone for TSE. In TF-GridNet, a Conv2D layer with a kernel size of 1×1 first embeds P prompt channels to D-dimensional embeddings. Next, TF-GridNet alternates temporal and spectral modeling blocks, each augmented with a Transformer-style self-attention layer. We propose to replace the self-attention operation with split-role multi-head attention (SR-MHA) in all blocks (unless otherwise noted), while preserving the other layers. Given that enrollment features X_e and mixture features X_m are obtained by splicing along time, SR-MHA are designed to attend from X_m to X_e via speaker-aware heads, and within X_m via context-aware heads.

To better handle the acoustic mismatches between enrollment and mixture signals, we divide the attention heads into two roles. The speaker-aware heads use queries from the mixture features and keys and values from the enrollment features, so that they focus on extracting speaker cues. The context-aware heads use queries, keys, and values all from the mixture, so that they can model the spectro-temporal coherence of the mixture and compensate for noise and reverberation. The outputs of both roles are concatenated and projected as

$$\text{SR-MHA}(X_m, X_e) = \text{Concat}(H^{spk}, H^{ctx})W^O, \tag{5}$$

with

$$H^{spk} = \text{softmax}\left(\frac{Q_m W^Q (K_e W^K)^\top}{\sqrt{d_k}}\right) V_e W^V, \tag{6}$$

$$H^{ctx} = \text{softmax}\left(\frac{Q_m \widetilde{W}^Q (K_m \widetilde{W}^K)^\top}{\sqrt{d_k}}\right) V_m \widetilde{W}^V, \tag{7}$$

where X_m and X_e denote the mixture and enrollment features, and W^O is the output projection. This split-role design ensures that part of the attention is dedicated to speaker identity while the other part maintains the consistency of the acoustic stream, potentially resulting in more robust TSE under mismatched conditions.

2.3. Loss Functions

Following [32], the loss function is defined on the estimated target speech after discarding the predictions in the time range of the concatenated enrollment and glue signals. We adopt the scale-invariant signal-to-distortion ratio (SI-SDR) loss [37], which is widely-used in TSE research.

3. EXPERIMENTAL SETUP

We evaluate LExTra under various enrollment-segment conditions for TSE, based on the WSJ0-2mix [8] and WHAMR! [33] datasets, both of which have been widely-adopted in prior TSE research. This section describes the datasets, experimental setup, evaluation metrics, and baseline systems used in our study.

WSJ0-2mix [8] is so far the most popular dataset to evaluate monaural talker-independent speaker separation algorithms in anechoic conditions. It consists of 20,000 (∼30.4 h), 5,000 (∼7.7 h) and 3,000 (∼4.8 h) two-speaker mixtures in its training, validation and test sets, respectively. The clean source signals are sampled from the WSJ0 corpus. The speakers for training and validation do not overlap with the speakers for testing. The two utterances in each mixture are fully-overlapped, and their relative energy level is uniformly sampled from the range $[-5, 5]$ dB.

The WHAMR! [33] dataset is used to validate our algorithms in noisy-reverberant conditions. It is based on the two-speaker mixtures in WSJ0-2mix [8] but reverberates each clean speech source and adds non-stationary noises. In our experiments, we focus specifically on evaluating how different enrollment segment conditions (e.g., length, content variation) affect the performance of LExTra.

We employ TF-GridNet [35, 36] as the DNN architecture for LExTra, which operate in the T-F domain. Following in [32], we investigate two TF-GridNet configurations, denoted as *TFGridNetV1* and *TFGridNetV2*. The *V1* version uses less computation for faster experimentation and is adopted in the ablation studies to assess the effectiveness of the proposed modifications under controlled conditions, and the *V2* version is employed to facilitate comparisons with established benchmark systems. Following the symbols defined in Table I of the TF-GridNet paper [35, 36], for TFGridNetV1, we set the hyper-parameters to $D = 128, B = 4, I = 1, J = 1, H = 200, L = 4$ and $E = 16$, and, for TFGridNetV2, we set them to $D = 128, B = 6, I = 1, J = 1, H = 256, L = 4$ and $E = 16$.

For the glue signal g in Eq. (3), which is utilized to prompt the DNN the time ranges of the enrollment utterance and the mixture, we set its length to 32 ms and its values to zero.

We employ the Adam optimizer, starting with a learning rate of 10^{-3}, which is decayed by 0.5 once the performance on the validation set is not improved in 4 consecutive epochs. For STFT, the square root of Hann window is used, and the window and hop sizes are respectively set to 16 and 8 ms. The evaluation metrics include SI-SDR improvement (SI-SDRi), SDR improvement (SDRi), and PESQ.

4. EVALUATION RESULTS

4.1. Results of Folded Prompt (FP)

LExt benefits from longer enrollment but longer prompt increases the sequence length that the DNN backbone needs to process. To inject enrollment information without increasing the processed sequence length, we fold a T-second enrollment segment into P shorter sub-segments (e.g., T=4 seconds, P=2), prepend each sub-segment to a copy of the mixture, and stack the resulting signals along a pseudo-channel dimension before feeding them to DNN for TSE. In Table 1, "+FP" means applying this folded-prompt (FP) scheme on top of LExt while keeping the backbone unchanged (i.e., using standard self-attention, with SR-MHA disabled), and V1/V2 denote TFGridNetV1/TFGridNetV2, respectively. Table 1 shows that the +FP variant keeps the parameter count unchanged (5.04 M for V1, 10.88 M for V2), while substantially reducing computation and runtime: for V1 with T=4 s, +FP (P=2) reduces FLOPs from 45.16 GFlops (LExt) to 29.27 GFlops and improves RTF from 0.0462 to 0.0337. Similar trends hold for V2, where FLOPs drop from 73.29 to 48.04 GFlops and RTF decreases from 0.0774 to 0.0564. The RTF values are measured on an NVIDIA RTX TITAN GPU. Overall, prompt folding preserves the information of a longer enrollment while avoiding a proportional increase in the processed sequence length, thereby achieving comparable SI-SDRi with notably lower computational cost and faster inference.

In Table 2, under the same TFGridNetV2 backbone, LExT(+FP) achieves 24.1 dB SI-SDRi, 24.4 dB SDRi, and 4.10 PESQ, showing competitive performance.

Table 1: Ablation of folded prompt on LExT with standard self-attention

System	DNN arch.	T (s)	P	Params (M)	FLOPs (G)	RTF	SI-SDRi (dB)↑
LExt	V1	4	1	5.0	45.2	0.046	**23.0**
LExT+FP		2	2	5.0	**29.3**	**0.034**	**23.0**
LExt	V1	1	1	5.0	**22.4**	0.029	22.0
LExT+FP		1	2	5.0	**22.4**	**0.028**	**22.7**
LExt	V2	4	1	10.9	73.3	0.077	**24.1**
LExT+FP		2	2	10.9	**48.0**	**0.056**	**24.1**

Table 2: Comparison with existing algorithms based on WSJ0-2mix.

System	Len. of enroll. (s)	SI-SDRi (dB)↑	SDRi (dB)↑	PESQ↑
Mixture	-	0.0	0.0	1.68
SpEx+ [19]	Full	16.9	17.2	3.43
DPRNN-Spe-IRA [38]	Full	17.3	17.6	3.43
SpEx++ [20]	Full	17.9	18.3	3.52
X-TF-GridNet [21]	Full	20.7	21.7	3.77
CIENet-mDPTNet [29]	Full	21.4	21.6	3.91
USEF-TSE [31]	Full	23.3	23.5	-
LExt (TFGridNetV2) [32]	4	**24.1**	24.3	**4.10**
LExt+FP (TFGridNetV2)	2×2	**24.1**	**24.4**	**4.10**

Table 3: Results of using SR-MHA in TF-GridNetV1 on WHAMR!.

System	DNN arch.	Attention-layer type	Context (s)	Heads	SI-SDRi (dB)↑
LExt	TFGridNetV1	Standard self-attention	1	4	16.3
LExt	TFGridNetV1	SR-MHA	1	2 (1+1)	16.3
LExt	TFGridNetV1	SR-MHA	1	4 (2+2)	16.7
LExt	TFGridNetV1	SR-MHA	1	8 (4+4)	**17.1**

Table 4: Comparison with existing algorithms on WHAMR!.

System	Len. of enroll. (s)	SI-SDRi (dB)↑	SDRi (dB)↑	PESQ↑
Mixture	-	0.0	0.0	1.68
SpEx+ [19]	Full	10.9	10.0	-
SpEx++ [20]	Full	11.4	10.4	-
X-TF-GridNet [21]	Full	14.6	13.8	-
CIENet-mDPTNet [29]	Full	15.7	14.3	2.55
USEF-TSE [31]	Full	16.1	14.9	-
LExt (TFGridNetV2) [32]	4	18.3	16.7	2.94
LExTra (TFGridNetV2)	2×2	**18.6**	**16.9**	**2.98**

4.2. Results of Split-Role Multi-Head Attention

We evaluate SR-MHA on WHAMR! using the TF-GridNetV1 backbone. Table 3 shows the results. All systems start from LExt (i.e., no prompt folding) and differ only in the attention block: the baseline uses the standard self-attention layers in TFGridNet, while SR-MHA replaces those layers with the proposed one. The "Heads" column denotes the total number of heads and their role split: $(a+b)$ means that a heads are speaker-aware and b heads are context-aware. We observe that increasing the number of attention heads with a balanced split (e.g., $(2+2)$ and $(4+4)$) improves the TSE performance. We emphasize that SR-MHA builds on the original self-attention module in TFGridNet, and under a fixed hidden dimension and sequence length, the FLOPs that a multi-head block use are largely governed by the hidden dimension, not by the number of attention heads, and hence the computation of the model only increases slightly.

4.3. Comparison with Existing Algorithms on WHAMR!

We evaluate LExTra on WHAMR! and compare it with representative TSE benchmarks. We use TFGridNetV2 as the DNN backbone, fold 4-second onset prompt into two 2-second segments, and use the proposed split-role attention, where the total number of heads is set to 8, among which 4 heads are speaker-aware and 4 are context-aware. Table 4 shows the results. We observe that LExTra delivers strong performance on WHAMR!. It consistently outperforms LExt, although using less amount of computation due to prompt folding. These results show the effectiveness of LExTra at TSE in noisy-reverberant conditions.

5. CONCLUSION

We have proposed LExTra, which tackle two issues of LExt: large amount of computation incurred by prepending enrollment utterance for onset prompting, and the mismatches between enrollment and mixture signals in noisy-reverberant conditions. A folded prompt keeps full enrollment information, while reducing the sequence length to process. Split-role multi-head attention disentangles cross-speaker cues from mixture self-context. Together, they reduce the amount of computation, improve the robustness of LExt, and yield similar or better TSE performance.

6. REFERENCES

[1] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "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.

[2] Q. Wang, R. Skerry-Ryan, Y. Jia *et al.*, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[3] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving Speaker Discrimination of Target Speech Extraction with Time-domain Speakerbeam," in *Proc. ICASSP*, 2020, pp. 691–695.

[4] S. He, H. Li, and X. Zhang, "Speakerfilter: Deep Learning-Based Target Speaker Extraction Using Anchor Speech," in *Proc. ICASSP*, 2020, pp. 376–380.

[5] K. Saijo, W. Zhang, Z.-Q. Wang, S. Watanabe, T. Kobayashi, and T. Ogawa, "A Single Speech Enhancement Model Unifying Dereverberation, Denoising, Speaker Counting, Separation, And Extraction," in *Proc. ASRU*, 2023, pp. 1–6.

[6] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[7] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria," in *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, 2007, pp. 1066–1074.

[8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative Embeddings for Segmentation and Separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[9] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation," in *Proc. ICASSP*, 2018, pp. 1–5.

[10] Z.-Q. Wang and D. Wang, "Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, 2019.

[11] Y. Luo and N. Mesgarani, "TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *Proc. ICASSP*, 2018, pp. 696–700.

[12] ——, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-domain Single-channel Speech Separation," in *Proc. ICASSP*. IEEE, 2020, pp. 46–50.

[14] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-independent Multi-talker Speech Separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[15] J. Vzmolikova, M. Delcroix, K. Kinoshita, and T. Nakatani, "SpeakerBeam: Speaker Adapted Speech Enhancement for Multi-Talker Mixtures," in *Proc. Interspeech*, 2017, pp. 3574–3578.

[16] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target Confusion in End-to-End Speaker Extraction: Analysis and Approaches," in *Proc. Interspeech*, 2022.

[17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[18] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural Networks for Small Footprint Text-dependent Speaker Verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.

[19] J. Xu, J. Shi, M. Zhang, X. Tan, D. Xu, and B. Xu, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[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 *Proc. ICASSP*, 2021, pp. 6109–6113.

[21] F. Hao, X. Li, and C. Zheng, "X-TF-GridNet: A Time–Frequency Domain Target Speaker Extraction Network with Adaptive Speaker Embedding Fusion," *Information Fusion*, vol. 112, 2024.

[22] S. Wang, K. Zhang, S. Lin, J. Li, X. Wang, M. Ge, J. Yu, Y. Qian, and H. Li, "WeSep: A Scalable and Flexible Toolkit Towards Generalizable Target Speaker Extraction," in *Proc. Interspeech*, 2024, pp. 4273–4277.

[23] 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.

[24] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-Net: Target Speaker Separation via Attention-Based Neural Network," in *Proc. Interspeech*, 2020, pp. 1411–1415.

[25] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural Target Speech Extraction: An Overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[26] T. Ochiai, M. Delcroix, T. Moriya, T. Ashihara, H. Sato, N. Tawara, T. Nakatani, and S. Araki, "Target Sound Information Extraction: Speech and audio Processing with Neural Networks Conditioned on Target Clues," *Acoustical Science and Technology*, vol. 46, no. 3, pp. 197–209, 2025.

[27] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[28] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. DImitriadis, J. Droppo, and Y. Gong, "Single-Channel Speech Extraction Using Speaker Inventory and Attention Network," in *Proc. ICASSP*, 2019, pp. 86–90.

[29] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target Speaker Extraction by Directly Exploiting Contextual Information in The Time-Frequency Domain," in *Proc. ICASSP*, 2024, pp. 10 476–10 480.

[30] B. Zeng, H. Suo, Y. Wan, and M. Li, "SEF-Net: Speaker Embedding Free Target Speaker Extraction Network," in *Proc. Interspeech*, 2023, pp. 3452–3456.

[31] B. Zeng and M. Li, "USEF-TSE: Universal Speaker Embedding Free Target Speaker Extraction," *IEEE Trans. Audio, Speech, Lang. Process.*, 2025.

[32] P. Shen, K. Chen, S. He, P. Chen, S. Yuan, H. Kong, X. Zhang, and Z.-Q. Wang, "Listen to Extract: Onset-Prompted Target Speaker Extraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 4832–4843, 2025.

[33] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and Reverberant Single-channel Speech Separation," in *Proc. ICASSP*, 2020, pp. 696–700.

[34] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending Speech Separation to Noisy Environments," *Proc. Interspeech*, 2019.

[35] 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," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[36] ——, "TF-GridNet: Making Time-Frequency Domain Models Great Again For Monaural Speaker Separation," in *Proc. ICASSP*, 2023, pp. 1–5.

[37] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - Half-Baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[38] C. Deng, S. Ma, Y. Sha, Y. Zhang, H. Zhang, H. Song, and F. Wang, "Robust Speaker Extraction Network Based on Iterative Refined Adaptation," in *Proc. Interspeech*, 2021, pp. 3530–3534.