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/ICASSP55912.2026.11461577

MC-LEXT: MULTI-CHANNEL TARGET SPEAKER EXTRACTION WITH ONSET-PROMPTED SPEAKER CONDITIONING MECHANISM

Tongtao Ling[1], *Shulin He*[1], *Pengjie Shen*[1,2], *and Zhong-Qiu Wang*[1]

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

lingtt2025@mail.sustech.edu.cn, wang.zhongqiu41@gmail.com

ABSTRACT

Multi-channel target speaker extraction (MC-TSE) aims to extract a target speaker's voice from multi-speaker signals captured by multiple microphones. Existing methods often rely on auxiliary clues such as direction-of-arrival (DOA) or speaker embeddings. However, DOA-based approaches depend on explicit direction estimation and are sensitive to microphone array geometry, while methods based on speaker embeddings model speaker identity in an implicit manner and may degrade in noisy-reverberant conditions. To address these limitations, we propose *multi-channel listen to extract* (MC-LExt), a simple but highly-effective framework for MC-TSE. Our key idea is to prepend a short enrollment utterance of the target speaker to each channel of the multi-channel mixture, providing an onset-prompted conditioning signal that can guide TSE. This design allows the deep neural network (DNN) to learn spatial and speaker identity cues jointly in a fully end-to-end manner. Evaluation results on noisy-reverberant benchmarks, including WHAMR! and MC-Libri2Mix, show the effectiveness of MC-TSE.

Index Terms— Multi-channel target speaker extraction

1. INTRODUCTION

Target speaker extraction (TSE) aims to separate the speech of a targeted speaker from a mixture speech containing multiple speakers given an auxiliary clue. It has broad applications in real-world scenarios such as virtual assistants, hearing aids, and smartphones [1]. While recent studies [2, 3] have achieved notable progress in monaural TSE under anechoic scenarios, many existing methods struggle in noisy-reverberant, multi-microphone settings [4], where complex acoustic interference and spatial ambiguity can cause significant performance degradation.

Existing multi-channel TSE (MC-TSE) approaches generally fall into two categories: methods that rely on pre-extracted speaker embeddings to condition the network, and methods that exploit direction-of-arrival (DOA) cues to utilize spatial information [5]. For speaker-embedding-based methods, a monaural enrollment utterance is typically recorded and a speaker model is utilized to obtain a fixed-length speaker embedding, which serves as a global identity cue to guide the extraction network [6, 7]. In addition, the speaker embedding can be integrated into the intermediate layers of the extraction network, enabling more fine-grained guidance [8–12]. However, speaker embedding only provides a fixed-length vector of the target speaker and lacks frame-level details that are helpful for precise TSE. Differently, DOA-based methods aim to model the acoustic scene and localize the target speaker within the mixture. By leveraging the spatial cues, the extraction network can separate overlapping speakers based on their spatial positions in noisy-reverberant environments [13–15]. Nevertheless, these approaches usually require explicit DOA estimation or carefully designed beamforming techniques [16–19], which can be sensitive to noise, reverberation, and microphone array geometry. In addition, they become less effective when the speakers are spatially close to each other.

On the other hand, both of the multi-channel methods often rely on complex pre-processing pipelines and may not generalize well across diverse acoustic environments, motivating the need for a simpler and more robust paradigm. Recently, the *listen to extract* (LExt) technique [20] introduces an extremely-simple while highly-powerful method for monaural TSE. It first pre-pends an enrollment utterance to each training mixture, and then trains a DNN on the resulting mixtures to predict the target speech. It shows that the prepended enrollment utterance can serve as a very effective prompt for the TSE network to identify the target speaker and extract the target speech. In this context, this paper aims to design a prompting-based multi-channel TSE system that avoids explicit spatial modeling while obtaining strong TSE performance.

To this end, we extend LExt to the multi-channel setting, proposing *multi-channel LExt* (MC-LExt). MC-LExt prepends a short enrollment utterance to each mixture channel, constructing a speaker conditioning signal that triggers the extraction process. Considering that the prepending mechanism results in a longer signal to process, to reduce the computation cost, we design a lightweight downsampler to reduce the computation spent on the enrollment speech. Furthermore, we propose to integrate conventional speaker embedding based conditioning mechanisms into MC-LExt to ensure that sufficient speaker information is retained to guide the model to perform TSE. Experiments on the WHAMR! [21] and MC-Libri2Mix [22] datasets demonstrate that MC-LExt achieves strong TSE performance in noisy-reverberant conditions. The contributions of this paper can be summarized as follows:

- We propose MC-LExt, a simple but highly-effective framework for MC-TSE. It eliminates the need for explicit spatial parameter estimation, enabling fully end-to-end training for TSE.
- We introduce a computationally-efficient down-sampling module to reduce the computation spent for the enrollment utterance.
- We integrate speaker embedding based conditioning into MC-LExt to provide sufficient speaker identity information for TSE.

2. REVIEW OF MONAURAL LEXT

LExt [20] proposes a simple and effective approach for monaural TSE. It prepends an enrollment utterance to each training mix-

This research was supported by National Key Research and Development Program of China (Grant No. 2025YFF0518003). *Corresponding author: Zhong-Qiu Wang*.

ture and trains DNNs to extract the target speaker based on the prepended mixtures. The rationale is that prepending the enrollment utterance can create an earliest speech onset for the target speaker and such an onset can prompt DNNs to extract the target speaker based on the prepended mixture. This onset-prompting strategy enables the model to identify the target speaker without relying on explicit speaker embeddings, getting rid of complicated DNN modules fusing speaker and mixture embeddings. Despite its success in monaural settings, LExt has not been examined in multi-channel scenarios, where spatial cues may interact with the onset prompt in non-trivial ways. Moreover, in LExt, longer enrollment utterance prepended to the mixture increases the length of the signal to process and hence requires more computation. This could limit the practical deployment of LExt. In this work, we extend LExt to multi-channel TSE by designing efficient prompting strategies that reduce the computation spent on enrollment speech, while incorporating speaker embedding based conditioning into MC-LExt, aiming to balance the efficiency and extraction performance.

3. MC-LEXT

3.1. Overview of MC-LExt

Fig. 1 illustrates the proposed MC-LExt. Given a multi-channel mixture $y \in \mathbb{R}^{C\times N}$ and the corresponding clean target $s \in \mathbb{R}^{N}$, the goal is to reconstruct s from the mixture y. Following monaural LExt [20], MC-LExt prepends an enrollment utterance $e \in \mathbb{R}^{E}$ to each channel of the mixture, forming the input to the DNN for MC-TSE. In the time domain, this procedure can be denoted as

$$\widetilde{y} = [\,\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_C\,]^\top \in \mathbb{R}^{C\times(E+N)}, \text{ with } \widetilde{y}_c = [\,e; y_c\,], \quad (1)$$

where E and N are respectively the number of time-domain samples of the enrollment and mixture, and C is the number of microphone channels. Based on the prepended mixture, $\widetilde{y}$, MC-LExt trains a DNN to predict the target speech s via supervised learning. The prepending strategy introduces a unified onset-like prompt to all microphones simultaneously, enabling the DNN to condition its TSE process on the temporal and spectral characteristics afforded by the enrollment utterance.

MC-LExt is trained via multi-microphone complex spectral mapping [23–25], where the real and imaginary (RI) components of input mixture signals are stacked as input feature to predict the RI components of the target signal. In detail, as shown in Fig. 1, after utterance-level concatenation, we transform the fused signal $\widetilde{y}$ into T-F domain via short-time Fourier transform (STFT). We further compute the RI components and append the magnitude spectrum as an additional feature map, resulting in a stacked representation:

$$\mathbf{X}_{\text{Spec}} = \text{STFT}(\widetilde{y}) \in \mathbb{C}^{C\times T\times F}, \quad (2)$$

$$\mathbf{X}_{\text{RI+Mag}} = \big[\mathcal{R}(\mathbf{X}_{\text{Spec}}), \mathcal{I}(\mathbf{X}_{\text{Spec}}), |\mathbf{X}_{\text{Spec}}|\big] \in \mathbb{R}^{(2C+1)\times T\times F}, \quad (3)$$

where $\mathcal{R}(\cdot)$, $\mathcal{I}(\cdot)$ and $|\cdot|$ respectively extract the real component, imaginary component and magnitude of a complex spectrogram, F denotes the number of frequency bins, and $T = T_{\text{enroll}} + T_{\text{mix}}$ denotes the number of time frames, with T_{enroll} and T_{mix} representing the number of frames of the enrollment segment and the original mixture. We then apply a Conv2D layer over the concatenated input $\mathbf{X}_{\text{RI+Mag}}$ to project it into a higher-dimensional embedding:

$$\mathbf{H} = \text{Conv2D}(\mathbf{X}_{\text{RI+Mag}}) \in \mathbb{R}^{D\times T\times F}, \quad (4)$$

where D denotes the embedding dimension of each T-F unit. This high-dimensional representation is then processed by a stack of

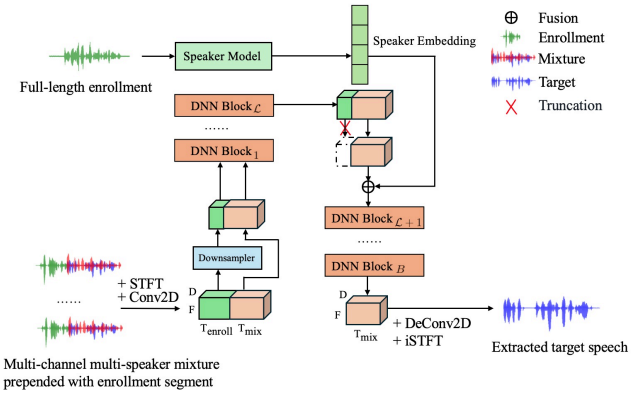

Fig. 1: Overview of MC-LExt. Best viewed in color.

DNN blocks to extract the target speech. To reduce computation, we propose to ① downsample the embedding of the enrollment segment along time and ② only pass it through a subset of the DNN blocks, whereas the mixture segment is fully processed by all the DNN blocks. Additionally, a fixed-length speaker embedding is computed from the full-length enrollment and we propose to ③ use it to condition the MC-LExt model, helping the MC-LExt model which already leverages onset-based conditioning to better extract the target speaker. Finally, the DNN blocks produce a high-dimensional representation $\widetilde{\mathbf{Z}}_B^{\text{mix}} \in \mathbb{R}^{D\times T_{\text{mix}}\times F}$, and a 2D transposed convolution (DeConv2D) is applied to project it back to the complex spectrogram space:

$$\mathbf{S} = \text{DeConv2D}(\widetilde{\mathbf{Z}}_B^{\text{mix}}) \in \mathbb{R}^{2\times T_{\text{mix}}\times F}, \quad (5)$$

where the first dimension corresponds to the RI components of the estimated spectrogram, and B is the number of DNN blocks. Then, iSTFT is applied to estimate the target speech in the time domain:

$$\hat{s} = \text{iSTFT}(\mathbf{S}) \in \mathbb{R}^{N}. \quad (6)$$

In default, the DNN is trained using the scale-invariant signal-to-distortion ratio (SI-SDR) loss [26].

In the rest of this section, we describe the proposed techniques marked using ①, ② and ③ in the previous paragraph.

3.2. Downsampling Enrollment Utterance

The computation of MC-LExt grows linearly with the length of the prepended enrollment utterance. To reduce the computation while preserving essential spectro-temporal information in the enrollment utterance, we apply a `downsampler` module consisting of a stack of Conv2D blocks to reduce the time steps of the embedding of the enrollment utterance. Each convolution block consists of a GroupNorm, ReLU activation, and a Conv2D layer with a kernel size of 3×3 and a stride of 2×1. This effectively halves the temporal resolution at each block while maintaining the frequency dimension, i.e.,

$$\text{Downsampler} = \big[\text{Conv2D, ReLU, GroupNorm}\big]_{\times G}, \quad (7)$$

where $G = \log_2(\lfloor \frac{E}{E'} \rfloor)$ and E' is the length of compressed enrollment speech. We then use the downsampler to compress the enrollment segment:

$$\mathbf{U} = \big[\text{Downsampler}(\mathbf{H}_{\text{enroll}}); \mathbf{H}_{\text{mix}}\big] \in \mathbb{R}^{D\times T'\times F}, \quad (8)$$

where $\mathbf{H}_{\text{enroll}}$ and $\mathbf{H}_{\text{mix}}$ respectively denote the embeddings of enrollment segment and mixture segment. After that, $\mathbf{U}$ is fed into a stack of DNN blocks (e.g., TF-GridNet blocks [3]) for training.

3.3. Selective Blocks for Enrollment Segment

To further reduce computation, we propose to use a subset of DNN blocks to process the entire prepended signal (consisting of the enrollment segment and mixture segment), while the remaining blocks operate solely on the mixture segment. Specifically, let the input be $\mathbf{Z}_0 = \mathbf{U}$ in Eq. (8), the output of the first $\mathcal{L}$-th blocks are computed as

$$\mathbf{Z}_i = \text{Block}_i(\mathbf{Z}_{i-1}), \quad 1 \leq i \leq \mathcal{L}. \tag{9}$$

where $\mathcal{L}$ is a tunable hyper-parameter. After the first $\mathcal{L}$ blocks, to reduce computation we discard the enrollment segment, and retain only the last T_{mix} frames corresponding to the mixture segment, i.e., $\mathbf{Z}_{\mathcal{L}}^{\text{mix}} = \mathbf{Z}_{\mathcal{L}}[:, T' - T_{\text{mix}} :, :]$. The subsequent blocks process the intermediate representations in the following way:

$$\mathbf{Z}_j^{\text{mix}} = \text{Block}_j(\mathbf{Z}_{j-1}^{\text{mix}}), \quad \mathcal{L}+1 \leq j \leq B. \tag{10}$$

3.4. Speaker Embedding Fusion

The intermediate mixture representation $\mathbf{Z}_j^{\text{mix}}$ can be further fused by incorporating speaker identity information. In real-world scenarios, enrollment utterances are often captured via microphone under less constrained conditions and can be very long (e.g., 30 seconds or more). In MC-LExt, directly prepending such lengthy enrollments for TSE is computationally expensive and inefficient. Therefore, we maintain the speaker embedding conditioning mechanism and use a speaker model (e.g., ECAPA-TDNN [27]) to obtain fixed-length embeddings from full-length enrollment utterances. It can be fused with $\mathbf{Z}_j^{\text{mix}}$ through various fusion methods, such as concatenation [9], addition [11], multiplication [10] and FiLM [12]). The fusion process is defined as follows:

$$\widetilde{\mathbf{Z}}_j^{\text{mix}} = \texttt{Fusion}(\mathbf{Z}_j^{\text{mix}}, v) \in \mathbb{R}^{D \times T_{\text{mix}} \times F}, \tag{11}$$

where $v \in \mathbb{R}^K$ is a fixed-length speaker embedding. When using this method, we replace $\mathbf{Z}_j^{\text{mix}}$ in Eq. (10) with $\widetilde{\mathbf{Z}}_j^{\text{mix}}$.

3.5. Loss Functions for Negative Pairs

MC-LExt is trained in an end-to-end manner to optimize the quality of the reconstructed target speech using the SI-SDR loss. To improve model robustness, we further consider the situation where the enrollment speaker is not present in the mixture, in which case the model should output silence. We adopt a contrastive training objective: for positive pairs, where the enrollment speaker is present in the mixture, the model is trained to reconstruct target speaker speech; for negative pairs, where the enrollment speaker is absent in the mixture, the model is instead encouraged to output silence. The standard SI-SDR loss is undefined for a silent target signal, as the scaling factor $\alpha = \langle \hat{s}, s \rangle / \langle s, s \rangle$ becomes undefined due to a division by 0 when the target signal s is all zeros. Using the standard SI-SDR loss would thus lead to numerical instability and training failure. Following previous work [28], we use log-MSE loss to handle negative pairs. Given the mixture signal y, target signal s, and estimated output $\hat{s}$, the log-MSE loss is defined as

$$\text{LOG-MSE}(s, \hat{s}, y) = \begin{cases} 10\log_{10}(||s - \hat{s}||^2 + \tau||s||^2), & s \neq \mathbf{0} \\ 10\log_{10}(||\hat{s}||^2 + \tau||y||^2), & s = \mathbf{0} \end{cases}$$

where $\tau = 10^{-\text{SNR}_{\text{max}}/10}$, with SNR_{max} set to 30 dB.

4. EXPERIMENTAL SETUP

We conduct experiments on two datasets: 1) **WHAMR!** [21]: Each mixture comprises two concurrent speech sources from the original WSJ0-2mix dataset, combined with a noise signal sampled from WHAM! [29] and convolved with room impulse responses (RIRs) simulated using the Pyroomacoustics toolkit [30]. The dataset contains $20,000$ training, $5,000$ validation, and $3,000$ test utterances of two-speaker mixed speech. 2) **MC-Libri2Mix** [22]: a multi-channel extension of the original Libri2Mix dataset [31], containing 2-speaker mixtures recorded with 4 microphones. It comprises $63,528$ training, $1,172$ validation, and $3,000$ test utterances of two-speaker mixed speech. For both datasets, we use 8 kHz sampling rate and the *min* version. For WHAMR!, we adopt the 2-channel version, and, for MC-Libri2Mix, the 4-channel version. Each speaker enrollment utterance is randomly selected from the original WSJ0 [32] and LibriSpeech [33] corpus. For evaluation, the enrollment utterances are chosen following the same setting as in prior works [20, 22].

For STFT, we use 16 ms window size, 8 ms hop size, and the square-root Hanning analysis window. Each training sample consists of a 4-second segment of enrollment utterance and a 4-second segment of mixture. When a downsampler is applied, an 8-second enrollment segment is used and is downsampled by 50%. We use TF-GridNet [3] with two configurations as the DNN architectures. For **TFGridNetV1**, we use $B = 4$ TF-GridNet blocks, each configured with the following hyper-parameters: $D = 128, I = 1, J = 1, H = 200, E = 16$, and $L = 4$, following the notations of TF-GridNet [3]. For **TFGridNetV2**, we set them to $B = 6, D = 128, I = 1, J = 1, H = 256, E = 16$, and $L = 4$. The smaller V1 model is used to verify the effectiveness of each proposed component, while the larger V2 model is used for comparison with state-of-the-art systems. We use ECAPA-TDNN [27] to extract 192-dimensional speaker embeddings. MC-LExt is trained for up to 100 epochs with ADAM [34]. The learning rate starts from 5×10^{-4} and is halved after 2 stagnant validation epochs. Performance is measured using SI-SDR improvement (SI-SDRi) and SDR improvement (SDRi).

5. EVALUATION RESULTS

Table 1 compares MC-LExt with monaural LExt based on the WHAMR! dataset. Compared to monaural LExt, MC-LExt achieves clearly better SDRi and SI-SDRi across all configurations, demonstrating the benefit of leveraging multi-channel information. Compared to a **Vanilla TSE** baseline which conditions the DNN only on a speaker embedding, MC-LExt improves SI-SDRi from 17.4 to 18.6 dB, demonstrating its stronger speaker-aware modeling capability. Enabling the downsampler (DS) reduces the enrollment embedding length, slightly improving performance (from 18.6 to 18.8 dB SI-SDRi) without harming model performance. Meanwhile, enabling only the speaker embedding (SE) yields 18.9 dB SI-SDRi, showing that conditioning the network on compact speaker identity features can strengthen TSE performance even without downsampling. Combining DS and SE achieves the best results (19.1 dB SI-SDRi). We also experiment with the Log-MSE loss, which resulted in inferior performance (18.8 dB SI-SDRi), likely because Log-MSE focuses on signal energy rather than waveform structure, causing the model to be overly conservative on negative pairs and degrading reconstruction quality on positive pairs. In Table 2, we supply each mixture in WHAMR! test set with a speech signal uttered by a speaker different from the ones in the mixture, and report how silent the output is via an *energy suppression ratio* metric defined as $10 \times \log_{10}(\|y\|_2^2 / \|\hat{s}\|_2^2)$. From the results, we can see that

Table 1: Results on WHAMR!. "DS" means whether using downsampler and "SE" means whether speaker embedding.

System	DNN arch.	C	DS	SE	Loss	SDRi (dB)	SI-SDRi (dB)
LExt [20]	TFGridNetV1	1	✗	✗	SI-SDR	15.5	17.1
Vanilla TSE	TFGridNetV1	2	✗	✓	SI-SDR	15.9	17.4
MC-LExt	TFGridNetV1	2	✗	✗	SI-SDR	17.1	18.6
MC-LExt	TFGridNetV1	2	✓	✗	SI-SDR	17.2	18.8
MC-LExt	TFGridNetV1	2	✗	✓	SI-SDR	17.4	18.9
MC-LExt	TFGridNetV1	2	✓	✓	SI-SDR	**17.6**	**19.1**
MC-LExt	TFGridNetV1	2	✓	✓	LOG-MSE	17.3	18.8

Notes: Using 8-second enrollment utterance when downsampler is available.

Table 2: Energy suppression ratio on negative pairs for Vanilla TSE and MC-LExt based on WHAMR! test set.

System	Energy suppression ratio (dB)
Vanilla TSE	45.7
MC-LExt	**61.0**

Table 3: Results of various speaker embedding fusion types (without downsampler and with SI-SDR loss).

System	Fusion type	DNN arch.	SDRi (dB)	SI-SDRi (dB)
MC-LExt	Concatenation	TFGridNetV1	17.4	18.9
MC-LExt	Addition	TFGridNetV1	17.2	18.6
MC-LExt	Multiplication	TFGridNetV1	**17.6**	**19.1**
MC-LExt	FiLM	TFGridNetV1	17.4	18.8

Table 4: Results of using various number of enrollment forward blocks on 2-channel WHAMR! dataset (without downsampler and speaker embedding, and with SI-SDR loss). GMAC/S is computed based on a 4-second mixture and 4-second enrollment speech.

System	DNN arch.	GMAC/s	SDRi (dB)	SI-SDRi (dB)
MC-LExt ($\mathcal{L}$=4)	TFGridNetV1	316.7	17.1	18.6
MC-LExt ($\mathcal{L}$=3)	TFGridNetV1	277.3	**17.7**	**19.2**
MC-LExt ($\mathcal{L}$=2)	TFGridNetV1	237.9	17.4	19.0
MC-LExt ($\mathcal{L}$=1)	TFGridNetV1	198.5	17.4	18.9

MC-LExt can better produce silent outputs than Vanilla TSE. In addition, for negative pairs, Fig. 2 illustrates a case study showing that the vanilla TSE system trained via Log-MSE fails to remain silent, while MC-LExt successfully outputs silence.

Table 3 investigates different fusion strategies for integrating speaker embeddings into MC-LExt. Multiplication achieves the highest performance, yielding 17.6 dB SDRi and 19.1 dB SI-SDRi. FiLM-based conditioning performs competitively, slightly outperforming concatenation and addition. These results indicate that the choice of fusion mechanism can further boost the discriminative power of the speaker embedding in the extraction process.

Table 4 examines the trade-off between computational cost and performance by varying the number of forward blocks ($\mathcal{L}$) processing the enrollment segment. Reducing the number of enrollment forward blocks from 4 to 1 lowers the GMAC/S from 316.73 to 198.48 ($\approx$37% reduction). Moreover, $\mathcal{L} = 1$ and $\mathcal{L} = 2$ produce very similar results (SI-SDRi difference < 0.1 dB), and $\mathcal{L} = 3$ achieves the highest SI-SDRi (19.2 dB) and SDRi (17.7 dB). However, the performance trend is non-monotonic increasing and $\mathcal{L} = 4$ yields the lowest scores despite having the largest computational cost. These results suggest that over-processing the enrollment segment may introduce redundancy or noise that slightly degrades extraction accuracy. We can conclude that forwarding fewer blocks for the enrollment is an effective strategy to reduce computational load without sacrificing extraction accuracy.

Table 5 and 6 respectively compare the results of MC-LExt and state-of-the-art TSE systems on MC-Libri2Mix and WHAMR!. On WHAMR!, MC-LExt significantly outperforms 1- and 2-channel baselines, including LExt [20] and the U-Conv-based TSE model [35]. This highlights the effectiveness of our onset-prompting strategy combined with the downsampling and speaker embedding mechanisms for MC-TSE. On the 4-channel MC-Libri2Mix dataset, MC-LExt also delivers substantial improvements. Compared to a conventional spatial filtering based method (Mask-based MVDR) [36] and recent DOA-based methods like L-SpEx [22], MC-LExt also delivers substantial improvements (16.3 dB SDRi and 14.7 dB SI-SDRi). Compared with the strong L-SpEx baseline (7.4 dB SI-SDRi), MC-LExt more than doubles the improvement, demonstrating strong generalization to clean, multi-channel, and highly-overlapped conditions. These results confirm that MC-LExt not only excels in noisy-reverberant conditions but also maintains robustness and scalability across diverse datasets.

Table 5: Results on 4-channel MC-Libri2Mix dataset. * denotes results reproduced by us.

System	SDRi (dB)	SI-SDRi (dB)
Mask-based MVDR Beamforming [36]	7.6	6.2
Pretrained Speaker Localizer [22]	7.0	5.7
L-SpEx w/o E2E Train [22]	9.0	7.2
L-SpEx [22]	9.2	7.4
HSRL-TSE* [37]	8.6	8.4
MC-LExt (TFGridNetV1) & SI-SDR	16.1	14.5
MC-LExt (TFGridNetV2) & SI-SDR	**18.6**	**16.3**
MC-LExt (TFGridNetV1) & LOG-MSE	15.1	13.7
MC-LExt (TFGridNetV2) & LOG-MSE	16.5	15.6

Table 6: Comparison of MC-LExt with other systems based on 2-channel WHAMR! dataset. * denotes results reproduced by us.

System	#CH	SDRi (dB)	SI-SDRi (dB)
SpEx+ [38]	1	10.0	10.9
X-TF-GridNet [39]	1	14.2	15.3
X-CrossNet [40]	1	14.1	14.6
DCF-NEt [41]	1	14.5	15.8
USEF-TSE [42]	1	14.9	16.1
CIENet-C2F-mDPTNet [43]	1	16.0	17.5
LExt (TFGridNetV2) [20]	1	16.7	18.3
Multi-TasNet [44]	2	-	12.1
Conv-TasNet-CD* [15]	2	11.3	12.2
U-Conv-based Extraction [35]	2	-	13.4
HSRL-TSE* [37]	2	9.0	9.1
MC-LExt (TFGridNetV2) & SI-SDR	2	**18.5**	**20.0**
MC-LExt (TFGridNetV2) & LOG-MSE	2	18.2	19.7

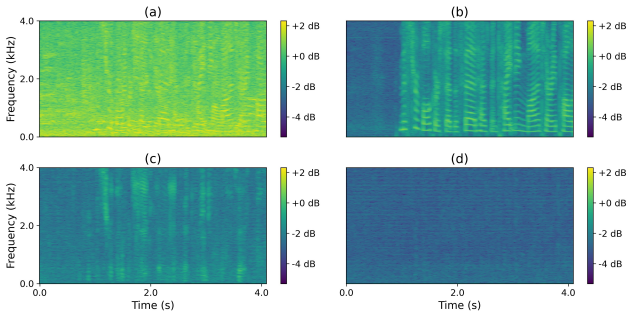

Fig. 2: Illustration, based on a negative pair (the enrollment speaker is absent in the mixture), of log spectrograms of (a) mixture; (b) target speech; (c) estimated speech by Vanilla TSE; and (d) estimated speech by MC-LExt.

6. CONCLUSIONS

We have proposed MC-LExt, an onset-prompted framework for MC-TSE in noisy-reverberant environments, which compresses long enrollment speech while retaining the speaker embedding conditioning mechanism to integrate spatial and speaker identity cues. Experiments on the WHAMR! and MC-Libri2Mix datasets show that MC-LExt consistently surpasses existing TSE models by a clear margin.

7. REFERENCES

[1] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita *et al.*, "Neural Target Speech Extraction: An Overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[2] C. Quan and X. Li, "SpatialNet: Extensively Learning Spatial Information for Multichannel Joint Speech Separation, Denoising and Dereverberation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 32, pp. 1310–1323, 2024.

[3] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee *et al.*, "TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation," in *Proc. ICASSP*, 2023, pp. 1–5.

[4] C. Zorilă, M. Li, and R. Doddipatla, "An Investigation into the Multi-Channel Time Domain Speaker Extraction Network," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 793–800.

[5] S. Zhang, J. Zhang, Y. Wang, and H. Yan, "DOA or Speaker Embedding: Which is Better for Multi-Microphone Target Speaker Extraction," *IEEE Signal Processing Letters*, vol. 32, pp. 3350–3354, 2025.

[6] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai *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.

[7] K. Zhang, J. Li, S. Wang *et al.*, "Multi-Level Speaker Representation for Target Speaker Extraction," in *Proc. ICASSP*, 2025, pp. 1–5.

[8] S. He, H. Zhang, W. Rao, K. Zhang *et al.*, "Hierarchical Speaker Representation for Target Speaker Extraction," in *Proc. ICASSP*, 2024, pp. 10 361–10 365.

[9] J. Wang, J. Chen, D. Su, L. Chen *et al.*, "Deep Extractor Network for Target Speaker Recovery from Single Channel Speech Mixtures," in *Proc. Interspeech*, 2018, pp. 307–311.

[10] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita *et al.*, "Improving Speaker Discrimination of Target Speech Extraction with Time-Domain Speakerbeam," in *Proc. ICASSP*, 2020, pp. 691–695.

[11] S. Wang, K. Zhang, S. Lin, J. Li *et al.*, "WeSep: A Scalable and Flexible Toolkit Towards Generalizable Target Speaker Extraction," in *Proc. Interspeech*, 2024, pp. 4273–4277.

[12] E. Perez, F. Strub, H. De Vries, V. Dumoulin *et al.*, "FiLM: Visual Reasoning with a General Conditioning Layer," in *Proc. AAAI*, vol. 32, no. 1, 2018.

[13] D. Choi and J.-W. Choi, "Multichannel-to-Multichannel Target Sound Extraction Using Direction and Timestamp Clues," in *Proc. ICASSP*, 2025, pp. 1–5.

[14] G. Li, S. Liang, S. Nie, W. Liu *et al.*, "Direction-Aware Speaker Beam for Multi-Channel Speaker Extraction," in *Proc. Interspeech*, 2019, pp. 2713–2717.

[15] J. Han, X. Zhou, Y. Long, and Y. Li, "Multi-Channel Target Speech Extraction with Channel Decorrelation and Target Speaker Adaptation," in *Proc. ICASSP*, 2021, pp. 6094–6098.

[16] M. Souden, J. Benesty, and S. Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2009.

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

[18] R. Gu, L. Chen, S.-X. Zhang, J. Zheng *et al.*, "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," in *Proc. Interspeech*, 2019, pp. 4290–4294.

[19] M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Beamformer-Guided Target Speaker Extraction," in *Proc. ICASSP*, 2023, pp. 1–5.

[20] 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 and Lang. Process.*, vol. 33, pp. 4832–4843, 2025.

[21] 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.

[22] M. Ge, C. Xu, L. Wang, E. S. Chng *et al.*, "L-SpEx: Localized Target Speaker Extraction," in *Proc. ICASSP*, 2022, pp. 7287–7291.

[23] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single-and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[24] ——, "Multi-Microphone Complex Spectral Mapping for Utterance-Wise and Continuous Speech Separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[25] K. Tan, Z.-Q. Wang, and D. Wang, "Neural Spectrospatial Filtering," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 30, pp. 605–621, 2022.

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

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[28] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel *et al.*, "What's All the Fuss About Free Universal Sound Separation Data?" in *Proc. ICASSP*, 2021, pp. 186–190.

[29] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu *et al.*, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[30] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[31] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge *et al.*, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," *arXiv:2005.11262*, 2020.

[32] 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.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[34] D. P. Kingma and J. Ba, "ADAM: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.

[35] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "Time-Domain Speech Extraction with Spatial Information and Multi Speaker Conditioning Mechanism," in *Proc. ICASSP*, 2021, pp. 6084–6088.

[36] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel *et al.*, "Improved MVDR Beamforming using Single-Channel Mask Prediction Networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[37] S. He, W. Xue, Y. Yang, H. Zhang *et al.*, "Enhancing Target Speaker Extraction with Hierarchical Speaker Representation Learning," *Neural Networks*, vol. 188, p. 107388, 2025.

[38] M. Ge, C. Xu, L. Wang, E. S. Chng *et al.*, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[39] 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, p. 102550, 2024.

[40] V. A. Kalkhorani and D. Wang, "TF-CrossNet: Leveraging Global, Cross-Band, Narrow-Band, and Positional Encoding for Single-and Multi-Channel Speaker Separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 32, pp. 4999–5009, 2024.

[41] K. Xue, R. Fan, S. Yu, C. Sun *et al.*, "DualStream Contextual Fusion Network: Efficient Target Speaker Extraction by Leveraging Mixture and Enrollment Interactions," *arXiv:2502.08191*, 2025.

[42] B. Zeng and M. Li, "USEF-TSE: Universal Speaker Embedding Free Target Speaker Extraction," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 33, pp. 2110–2124, 2025.

[43] X. Yang, C. Bao, and X. Chen, "Coarse-to-Fine Target Speaker Extraction Based on Contextual Information Exploitation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 32, pp. 3795–3810, 2024.

[44] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On End-to-End Multi-channel Time Domain Speech Separation in Reverberant Environments," in *Proc. ICASSP*, 2020, pp. 6389–6393.