XTSE-Net: A Unified Offline and Online Target Speaker Extraction System for the REAL-TSE Challenge

Yihao Meng*, Xin Hai*, Yuyang Fan*, Zhuan Yi*, Wei Liu*,
Gongping Huang*, Andong Li†‡
* School of Electronic Information, Wuhan University, Wuhan, China
† Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
‡ University of Chinese Academy of Sciences, Beijing, China

***Abstract*—We present XTSE-Net, a unified system for target speaker extraction in the SLT challenge, supporting both offline and online scenarios under a consistent architecture. The offline system adopts a non-causal X-TF-GridNet to fully exploit contextual information, while the online system is designed with a causal variant to satisfy strict low-latency constraints. To further improve speech quality, pretrained speech enhancement models, including SGMSE and causal GTCRN, are incorporated as post-processing modules. The online system achieves an average algorithmic latency of 72.5 ms under the official latency evaluation setting. Experimental results on the official evaluation set demonstrate that the proposed system achieves competitive performance across both tracks, ranking 10th in the offline setting and 7th in the online setting.**

***Index Terms*—SLT challenge, target speaker extraction, offline, online.**

I. Introduction

Target speaker extraction (TSE) aims to recover the speech of a target speaker from a multi-speaker mixture given an enrollment utterance. Recent advances in deep neural networks have significantly improved TSE performance through more effective speaker representation learning and extraction architectures [1]–[3]. Meanwhile, powerful time-frequency separation backbones, such as Band-Split RNN (BSRNN) [4], have further promoted the development of high-quality speech extraction systems. Nevertheless, most existing TSE systems are trained and evaluated on simulated mixtures generated from clean speech corpora, such as LibriMix [5], limiting their generalization to realistic conversational scenarios.

To bridge this gap, *REAL-T: Real Conversational Mixtures for Target Speaker Extraction* [6] introduces a large-scale corpus of naturally overlapped conversations containing spontaneous speech, background noise, and room reverberation, providing a realistic benchmark for target speaker extraction. Built upon this corpus, the REAL-TSE Challenge[1] evaluates systems from multiple perspectives, including transcription accuracy, speaker similarity, and perceptual speech quality, posing higher requirements on both interference suppression and speech fidelity.

In this paper, we present **XTSE-Net**, our submission to both the offline and online tracks of the REAL-TSE Challenge. The proposed systems are built upon X-TF-GridNet [7] for target speaker extraction. Specifically, the offline system adopts the original non-causal X-TF-GridNet, while the online system follows the causal design of DSINet [8] to construct a causal variant of X-TF-GridNet. To further improve the quality of the extracted speech, we leverage pre-trained speech enhancement models as post-processing modules, where SGMSE [9] is employed for the offline track and causal GTCRN [10] is adopted for the online track. Finally, the enhanced speech is fused with the extraction output at the waveform level to better preserve target speaker characteristics while improving perceptual speech quality. Experimental results demonstrate that XTSE-Net consistently outperforms the official baselines on both challenge tracks.

[1]https://real-tse.github.io/challenge/

II. Proposed Systems

In this section, we present the proposed XTSE-Net for both the offline and online tracks of the REAL-TSE Challenge. The unified architectures of the two systems are illustrated in Fig. 1.

A. Offline XTSE-Net

In the offline setting, given a mixture waveform and an enrollment utterance of the target speaker, both signals are first transformed into complex STFT representations and fed into an X-TF-GridNet-based target speaker extraction network. Following the original X-TF-GridNet framework, the mixture and enrollment features are first projected into a latent embedding space by convolutional encoders. The enrollment features are then processed by the built-in speaker encoder to extract speaker representations, which are progressively injected into each GridNet block through speaker-guided fusion modules. Finally, the enhanced latent representation is decoded to estimate the complex spectrum of the target speech, and inverse STFT is applied to reconstruct the extracted waveform.

Although X-TF-GridNet effectively suppresses interfering speakers, residual interference and speech distortion may still remain under realistic acoustic conditions. To further improve speech quality, we adopt the officially released pre-trained SGMSE model as a post-processing module. The official checkpoint, trained for speech enhancement on the WSJ0-2Mix and WHAMR! datasets, is directly adopted without additional fine-tuning, and all model parameters remain frozen during the training of XTSE-Net. The waveform estimated by X-TF-GridNet is then fed into the frozen SGMSE model for

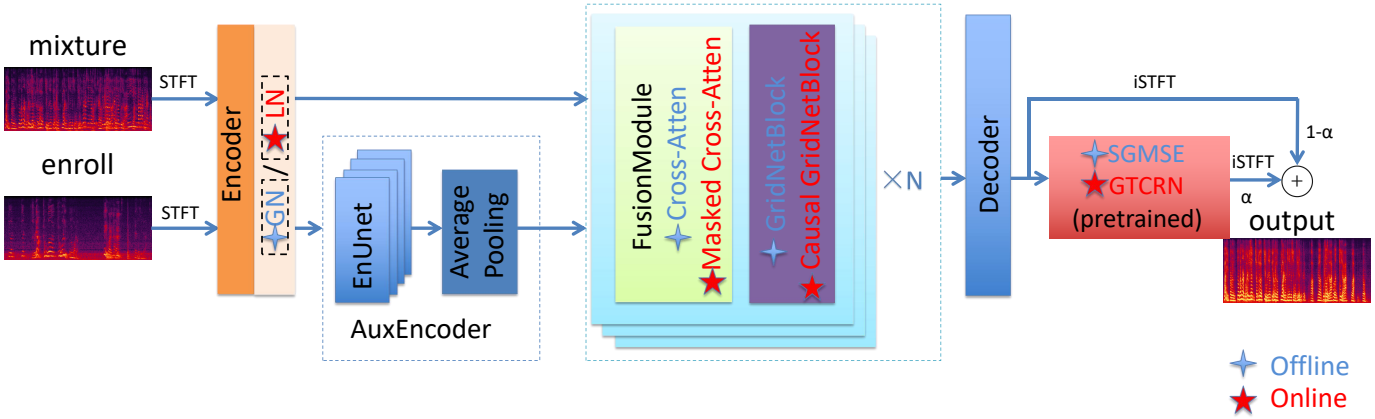

Fig. 1. Overview of the proposed XTSE-Net architecture for both offline and online tracks. Blue stars indicate offline-specific components, while red stars indicate online-specific components.

further refinement. This design enables XTSE-Net to leverage the strong denoising capability of large-scale pre-trained speech enhancement models while keeping the extraction framework unchanged.

B. Online XTSE-Net

Similar to the offline system, the online model also adopts a two-stage framework consisting of target speaker extraction followed by speech enhancement. Given a mixture waveform and an enrollment utterance, both signals are first transformed into complex STFT representations and processed by a causal X-TF-GridNet-based extraction network. The extracted waveform is subsequently refined by a lightweight causal GTCRN speech enhancement model.

To satisfy the real-time requirement, we redesign X-TF-GridNet into a fully causal architecture while preserving its original framework as much as possible. Specifically, the bidirectional temporal LSTM in each GridNet block is replaced with a unidirectional LSTM so that temporal modeling only relies on historical frames. In addition, a causal attention mask is introduced into the cross-frame self-attention module to prevent future information leakage. Furthermore, the mixture encoder and decoder are implemented using causal convolution and causal transposed convolution, respectively, while the speaker encoder remains unchanged since the enrollment utterance is fully available before inference.

Compared with the offline system that employs SGMSE, the online system adopts the officially released pre-trained GTCRN model as the post-processing module. The official checkpoint trained on the DNS Challenge 3 dataset is directly adopted without further fine-tuning, and all model parameters remain frozen during the training of XTSE-Net. This lightweight enhancement stage effectively suppresses residual interference and speech distortion while satisfying the low-latency requirement.

C. Training Objective and Waveform Fusion

Both the offline and online extraction networks are trained using the same objective as X-TF-GridNet, which combines the SI-SNR loss and an auxiliary speaker classification loss:

$$\mathcal{L} = \mathcal{L}_{\text{SI-SNR}} + \lambda \mathcal{L}_{\text{CE}}, \tag{1}$$

where $\mathcal{L}_{\text{SI-SNR}}$ denotes the scale-invariant signal-to-noise ratio loss, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for speaker classification, and λ is a weighting factor.

For both systems, the enhanced speech produced by the post-processing module is finally fused with the extraction output through a waveform-level fusion strategy:

$$\hat{s} = \alpha s_{\text{ext}} + (1-\alpha) s_{\text{enh}}, \tag{2}$$

where s_{ext} denotes the output of the extraction network, s_{enh} denotes the output of the speech enhancement module (SGMSE for the offline system and GTCRN for the online system), and α is the fusion coefficient. This fusion strategy effectively balances target speaker preservation and perceptual speech quality, consistently outperforming either individual output.

III. Experiments

A. Implementation

The training dataset was synthesized online using Chinese (AISHELL-1, AISHELL-3, THCHS-30) [11]–[13] and English (LibriTTS, VCTK, EARS, VoxCeleb1) [14]–[17] speech corpora, with all signals resampled to 16 kHz. Background noise was sourced from the DNS Challenge and WHAM! datasets [18], [19]. To better emulate real-world target speaker extraction (TSE) scenarios, data generation incorporated both simulated and measured room impulse responses (RIRs). Specifically, simulated RIRs were generated via Pyroomacoustics using the image-source method (ISM), with RT_{60} values randomly sampled from 0.15 s to 1.10 s. Measured RIRs were drawn from diverse public databases, including AIR, BUT ReverbDB, DECHORATE, ACE, and OpenSLR28/RVB2014.

To facilitate training convergence, we introduced a curriculum learning strategy that progressively scaled data augmentation over a 100-epoch schedule, with each epoch comprising 50k dynamically generated samples. During the first 90 epochs, the probabilities of injecting background noise and simulated reverberation were jointly increased from 0.0 to 0.5 with a step size of 0.1 across successive epoch intervals. The final 10 epochs served as a real-world adaptation stage: while the noise probability remained at 0.5, the 0.5 reverberation probability was equally partitioned between simulated and measured RIRs. Additionally, enrollment utterances were augmented with noise and reverberation, each applied with a probability of 0.2. Throughout training, the signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR) were randomly sampled from $[-5, 10]$ dB and $[-5, 20]$ dB, respectively.

B. Model Configurations

For the offline system, the target speaker extractor is implemented using the non-causal X-TF-GridNet, while the online system adopts the proposed causal X-TF-GridNet described in Section II. Unless otherwise specified, the two extraction networks share identical hyperparameter settings. Both models employ 3 GridNet blocks with an embedding dimension of 64, a temporal LSTM hidden size of 256, and a 4-head attention module. The STFT is computed using a 512-point FFT with a window length of 512 samples and a hop size of 128 samples. According to the challenge definition, the causal X-TF-GridNet introduces a theoretical algorithmic latency of 24 ms. Combined with the official GTCRN post-processing module, whose STFT configuration introduces an additional 16 ms algorithmic latency, the total theoretical algorithmic latency of the proposed online system is 40 ms.

Both extraction networks are trained for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. The learning rate is exponentially decayed to 2.5×10^{-5} during training, and gradient clipping with a maximum norm of 5.0 is applied. The batch size is set to 16 for the offline system and 20 for the online system.

During inference, the official pre-trained SGMSE and GTCRN models are adopted as frozen post-processing modules for the offline and online systems, respectively. The waveform fusion coefficient α is fixed to 0.8 for the offline system and 0.5 for the online system, as determined on the development set.

C. Results

TABLE I
FINAL SYSTEM PERFORMANCE OF THE OFFLINE SYSTEM ON THE OFFICIAL EVALUATION SET.

Method	TER↓	Timing F1↑	SIM↑	DNSMOS↑
BSRNN_TFMAP	0.838	0.829	0.443	1.612
BSRNN_EMB	0.829	0.829	0.417	1.844
XTSE-Net (Offline)	0.759	**0.850**	0.457	**2.006**
XTSE-Net w/o SGMSE	**0.757**	0.849	**0.467**	1.818

TABLE II
FINAL SYSTEM PERFORMANCE OF THE ONLINE SYSTEM ON THE OFFICIAL EVALUATION SET.

Method	TER↓	Timing F1↑	SIM↑	DNSMOS↑
BSRNN_TFMAP_CAUSAL	0.805	0.836	0.456	1.670
BSRNN_EMB_CAUSAL	0.809	0.821	0.422	1.714
XTSE-Net (Online)	0.742	**0.852**	0.465	**2.019**
XTSE-Net w/o GTCRN	**0.733**	**0.852**	**0.477**	1.742

Table I and Table II summarize the performance of the proposed system compared with baseline methods on the official evaluation set, where XTSE-Net consistently outperforms the baselines[2] across both offline and online settings. In the offline system, the variant with SGMSE achieves better DNSMOS and SIM than the version without post-processing, while other metrics remain comparable; in the online system, adding GTCRN shows a similar trend with improved DNSMOS and SIM and negligible changes in TER and Timing F1. Overall, the results show that the proposed system achieves consistent gains over baselines and that the post-processing modules mainly affect perceptual quality while maintaining stable extraction performance.

Latency Analysis. The algorithmic latency of the online system is measured using the official perturbation-based latency evaluation protocol defined in the challenge. The proposed system achieves an average latency of 72.5 ms, with a minimum of 64.9 ms and a maximum of 82.0 ms. The results indicate that the system operates within a consistently bounded delay range under the evaluation setting, satisfying the real-time constraint of the challenge, as shown in Figure 2.

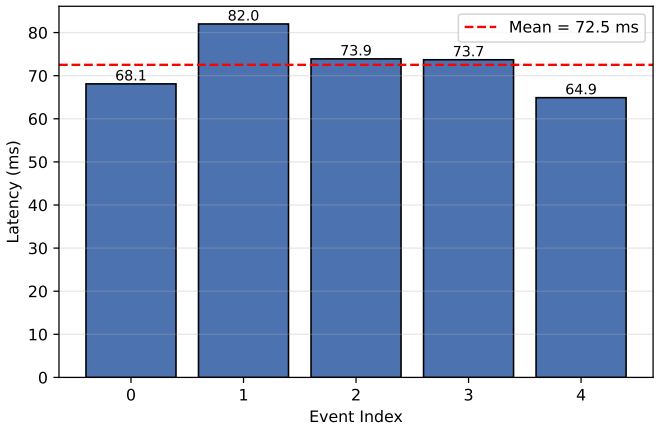

Fig. 2. Per-event algorithmic latency of the proposed streaming system. The red dashed line indicates the mean latency.

IV. CONCLUSION

This paper proposes a unified offline and online target speaker extraction framework based on X-TF-GridNet. Experimental results demonstrate strong performance under both accuracy and low-latency constraints, validating its practicality for real-time deployment.

[2]https://github.com/REAL-TSE/REAL-TSE-Challenge

V. REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[2] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] K. Zhang, J. Li, S. Wang, Y. Wei, Y. Wang, Y. Wang, and H. Li, "Multi-level speaker representation for target speaker extraction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[4] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[5] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[6] S. Li, S. Wang, J. Han, K. Zhang, W. Wang, and H. Li, "Real-t: Real conversational mixtures for target speaker extraction," in *Proc. Interspeech 2025*, 2025, pp. 1923–1927.

[7] 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.

[8] F. Hao, A. Li, X. Li, and C. Zheng, "Dsinet: Towards real-time target speaker extraction with dynamic speaker information fusion," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[9] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," in *Proc. Interspeech*, 2022, pp. 2928–2932.

[10] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, and J. Lu, "Gtcrn: A speech enhancement model requiring ultralow computational resources," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 971–975.

[11] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[12] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin tts corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.

[13] D. Wang and X. Zhang, "THCHS-30: A free chinese speech corpus," 2015.

[14] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[15] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[16] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proc. Interspeech 2024*, 2024, pp. 4873–4877.

[17] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[18] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2023 deep noise suppression challenge," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024.

[19] 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," in *Proc. Interspeech 2019*, 2019, pp. 1368–1372.