SHNU Target Speaker Extraction System for SLT 2026 REAL-TSE Challenge

Yiming Yang[1], *Yu Wang*[1], *Yanhua Long*[**]

[1] Shanghai Normal University, Shanghai, China

y2379286479@outlook.com, yanhua@shnu.edu.cn

Abstract

This paper describes SHNU target speaker extraction system (SHNU-TSE, team name-"SHNU-TSE"), submitted to Track 1 and Track 2 of the SLT 2026 REAL-TSE Challenge. For Track 1, we propose an online low-latency target speaker extraction system based on a causal time-frequency mask estimation network. The system incorporates target speaker information through cross-attention and adopts a three-stage training strategy to improve robustness under realistic acoustic conditions. For Track 2, we adopt a speaker-embedding-free approach that captures time-frequency features via multi-head attention (MHA) and BLSTM. Moreover, we utilize approximately 600 hours of training data and a composite loss function, dynamically adjusting the weight of each loss component through an adaptive weighting mechanism. In Track 1, SHNU-TSE achieved an overall score of 9.75, ranking 10th , with a TER of 0.797 (14th), a Timing F1 of 0.842 (8th), a SIM of 0.462 (10th), and a DNSMOS OVRL of 2.253 (7th). In Track 2, SHNU-TSE achieved an overall score of 7.25, tying for 5th place, with a TER of 0.731, Timing F1 of 0.840, SIM of 0.507 (rank 5), and DNSMOS OVRL of 2.716 (rank 4).

Index Terms: target speaker extraction, real conversational environments

1. Introduction

Target speaker extraction (TSE)[1] aims to isolate the specific speaker from a multi-talker or noisy acoustic environment by leveraging a set of auxiliary clues, such as a reference enrollment utterance from the target speaker.

The REAL-TSE Challenge, organized as a satellite challenge of IEEE SLT 2026, focuses on TSE in real conversational environments. In contrast to conventional benchmarks built on simulated or read-speech data, REAL-TSE emphasizes naturally occurring overlap, authentic reverberation, ambient noise, and conversational dynamics in Mandarin and English, providing a more practical testbed for real-world TSE research. The competition consists of two tracks, namely Track 1 (Online TSE) and Track 2 (Offline TSE). For Track 1, TSE in latency-sensitive applications such as assistive hearing, real-time communication, and interactive voice systems, where the system must produce temporally responsive outputs under overlapping speech and background noise, enabling continuous and natural listening. Models must ensure input changes are reflected within a bounded delay ($\leq$ 100 ms), balancing extraction quality and responsiveness. For Track 2, TSE in offline scenarios such as speech transcription, meeting analysis, and audio post-processing, where the full utterance is available prior to inference. This track targets the upper performance bounds of TSE systems, focusing on speech quality, interference suppression, and robustness without latency constraints.

For both Track 1 and Track 2, SHNU-TSE follows an encoder-extractor-decoder architecture. In the encoder, the enrollment and mixture interact directly; the extractor employs MHA and BLSTM to capture information, while the decoder reconstructs the speech. Multi-objective joint optimization is used during training.

2. Data

For Track 1, the training process consists of three stages. In the first stage, the model is jointly trained on the mini versions of the WHAMR![2] and Libri2Mix [3]`2spk+noise/train-360` datasets using 2-second audio segments. In the second stage, the model is further fine-tuned using longer utterances (7–14 seconds) selected from the same WHAMR! and Libri2Mix datasets, enabling the model to better capture long-range temporal dependencies. Finally, the model is fine-tuned on self-enhanced AMI meeting recordings, where the enhanced speech is generated by the Stage 2 model, to reduce the domain gap between synthetic and real meeting recordings. During training, each mixture utterance is associated with five enrollment utterances from the target speaker. One enrollment utterance is randomly selected at the beginning of each training epoch to improve robustness to enrollment variability. The official REAL-T development set is used for model selection and hyperparameter tuning throughout all training stages.

For Track 2, the training data of SHNU-TSE comprises publicly available speech datasets, noise datasets, and simulated reverberation data. The clean speech component uses AISHELL-1[4], AISHELL-3[5], and LibriSpeech[6]. Specifically, the LibriSpeech data were used to construct mixed samples featuring three types of speaker pairings male-male, male-female, and female-female based on speaker gender labels; the AISHELL-1 and AISHELL-3 datasets were not subjected to this pairing classification and served primarily to supplement the training data for the extraction of Chinese target speakers. The total duration of the training set is approximately 600 hours. The noise data is primarily from the WHAM![7] and DNS2020 Challenge[8] noise datasets. Reverberant data is generated using gpuRIR, with the reverberation time (T60) set between 0.3 and 0.6 seconds. During training, target speech, interference speech, and noise are randomly sampled and dynamically mixed with room impulse responses to construct training samples featuring varying signal-to-noise ratios, interference intensities, and reverberation conditions.

[**] indicates the corresponding author.

3. Model Configuration

On Track 1, SHNU-TSE follows a causal time-frequency mask estimation architecture for online low-latency target speaker extraction. The input mixture waveform is first transformed into the time-frequency domain using a causal Short-Time Fourier Transform (STFT) frontend. The resulting time-frequency representation is processed by a stack of causal transformer blocks to estimate a complex ratio mask (CRM), which is applied to the mixture spectrum to extract the target speaker. Target speaker information is incorporated through a cross-attention module, where enrollment features extracted by a pre-trained speaker encoder are fused into each transformer block to guide mask estimation. Finally, the enhanced spectrum is converted back to the time domain through a causal inverse STFT. The model is trained using a three-stage training strategy consisting of synthetic pre-training, long-audio fine-tuning, and domain adaptation to improve robustness under realistic acoustic conditions. The algorithmic latency consists of an 80 ms look-ahead and an additional 8 ms corresponding to half of the STFT analysis window, resulting in a total theoretical latency of approximately 88 ms. Using the official `latency_verification.py` script provided by the challenge organizers, the maximum measured effective latency is 90.7 ms, which closely matches the theoretical latency.

On Track 2, SHNU-TSE follows an encoder-extractor-decoder architecture: the encoder transforms the mixture and enrollment signals into the time-frequency domain via STFT, applies dynamic range compression (DRC), and fuses them through an early interaction block before concatenation and convolution to produce a fused feature; the extractor processes this feature through stacked basic blocks to estimate a mask; the decoder applies the mask and reconstructs the target waveform via inverse compression and inverse STFT. Within the extractor, the basic blocks employ MHA and BLSTM to capture dependencies. This model is trained using a comprehensive multi-task joint optimization strategy that encompasses four key dimensions: signal reconstruction, speaker consistency, semantic preservation, and temporal alignment. This approach ensures high speech quality, accurate data acquisition, and robust speaker identity characteristics.

Table 1: *Evaluation results of SHNU-TSE for Track 1.*

Models	**TER**	**Timing F1**	**SIM**	**OVRL**
BSRNN_EMB_CAUSAL	0.809	0.821	0.422	1.714
SHNU-TSE-Stage1	**0.796**	0.835	0.373	2.172
SHNU-TSE-Stage2	0.799	0.841	0.443	**2.293**
SHNU-TSE-Stage3	0.797	**0.842**	**0.462**	2.253

Table 2: *EVAL results for SHNU-TSE in Track 2.*

Models	**TER**	**Timing F1**	**SIM**	**OVRL**
BSRNN-TFMAP	0.838	0.829	0.443	1.612
SHNU-TSE-base	0.840	**0.842**	0.415	2.554
SHNU-TSE-1	**0.708**	0.834	0.476	2.626
SHNU-TSE-2	0.731	0.840	**0.507**	**2.716**

4. Results

Table 1 summarizes the evaluation results on the official evaluation set. SHNU-TSE-Stage1 represents the model trained on the combined WHAMR! and Libri2Mix datasets using 2-second audio segments. Although SHNU-TSE-Stage1 achieves competitive performance over the official baseline (*BSRNN_EMB_CAUSAL*), it is trained using only short utterances, which limits its ability to model long-context speech. Therefore, the model is further fine-tuned on selected 7–14-second utterances from the same training datasets to obtain SHNU-TSE-Stage2, leading to substantial improvements in the SIM score, Timing F1, and DNSMOS OVRL. Finally, the model is fine-tuned on self-enhanced AMI meeting recordings, where the enhanced speech is generated by the Stage 2 model to further improve robustness and speaker discrimination.

On Table 2, SHNU-TSE-base represents the results obtained by training our model on the Libri2Mix and WHAMR! dataset for Track 2. We found the SIM score for this result to be too low. Consequently, we reconfigured the training set using data from AISHELL-1, AISHELL-3, and LibriSpeech, while selecting noise from the WHAM! and DNS2020 Challenge datasets and simulating reverberation using gpuRIR. SHNU-TSE-1 is precisely this result. We further enhance SHNU-TSE-1 using TF-GridNet[9]. Then, we perform time-domain fusion of this result with SHNU-TSE-1 to obtain SHNU-TSE-2.

5. 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] 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.

[3] 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.

[4] 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.

[5] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[7] G. Wichern, J. Antognini, M. Flynn *et al.*, "WHAM!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[8] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[9] Z.-Q. Wang, S. Cornell, S. Choi *et al.*, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.