SEP2TSE: SEPARATION-FIRST CASCADED TARGET SPEAKER EXTRACTION FOR THE REAL-TSE CHALLENGE

Xiaoxia Yang, Yiqian Pan, Liangliang Wang, and Xi Cai

China Telecom Cloud Technology Co., Ltd., Beijing, China

{yangxx16, panyq9, wangll52, caixi}@chinatelecom.cn

ABSTRACT

The REAL-TSE Challenge requires extracting a target speaker from real meeting mixtures with enrollment cues under background noise, reverberation and speech overlapping. For this challenge, we propose **Sep2TSE** (*Sep*aration-to-*TSE*), a two-stage framework unlike typical cascaded TSE systems where both stages perform target enhancement. Stage 1 uses source separation and enrollment-guided soft restitch to raise SIR; Stage 2 applies enrollment-conditioned TSEWavLM with a neural vocoder for denoising and reconstruction. Front-loading interference suppression improves robustness under adverse acoustics. On the REAL-TSE offline leaderboard, Sep2TSE ranks 5th among all submissions, showing that a modular separation-first design is competitive for real meeting scenarios.

***Index Terms*—** Target speaker extraction, speech enhancement, REAL-TSE

1. INTRODUCTION

The cocktail party problem remains a fundamental obstacle for speech-related systems. Practical microphone signals are corrupted by overlapping speech, background noise, and reverberation. Unlike general speech enhancement or blind separation, target speaker extraction (TSE) uses enrollment cues to suppress interference and recover a designated speaker.

Data-driven TSE methods now dominate and fall into two lines: *discriminative* approaches that estimate spectral masks or discriminative features, and *generative* approaches that reconstruct waveforms for higher fidelity. Representative discriminative systems include TEA-PSE [1, 2, 3], which injects enrollment speaker embeddings into the TSE network; TF-GridNet [4], often adopted as the backbone of TSE networks; and USEF-TSE [5], which conditions separation on enrollment via cross-attention without external speaker embeddings.

Generative TSE methods reconstruct high-fidelity target speech. A representative discretization–vocoder framework [6] predicts discrete speech tokens from the mixture and enrollment via self-supervised discretization, and resynthesizes the target waveform with an enrollment-conditioned discrete vocoder.

The REAL-TSE Challenge moves enrollment-conditioned TSE from simulated benchmarks to real meeting recordings with natural overlap, reverberation, and background noise. The challenge provides **online** and **offline** tracks; we compete in the offline track, which permits full-utterance context. Official ranking averages dense ranks over WER (intelligibility), speaker similarity, DNSMOS (perceptual quality), and target-presence F1.

In this work, we present **Sep2TSE**, a two-stage pipeline that differs from conventional single-pass TSE. **Stage 1** performs source separation followed by enrollment-guided soft restitch to raise the signal-to-interference ratio (SIR). **Stage 2** carries out enrollment-conditioned target enhancement with TSEWavLM and a dual-stream vocoder. Our system ranked **5th** on the REAL-TSE offline leaderboard; evaluation-set results also outperform official BSRNN baselines [7].

2. PROPOSED METHOD

2.1. Architecture

Fig. 1(a) summarizes **Sep2TSE**. Target speaker extraction (TSE) and source separation share the same underlying goal—disentangling overlapping voices—yet they differ in how speaker identity is specified. TSE conditions on an enrollment utterance to extract a *designated* speaker, whereas separation partitions the mixture without explicit identity cues.

We therefore adopt a *separation-first* workaround. **Stage 1** applies a separation model to split the multi-talker mixture into candidate streams, then fuses them with enrollment-guided **soft restitch** to obtain a high-SIR preliminary target estimate. **Stage 2** performs the final target-speaker enhancement and denoising in a *generative* model. The following subsections detail each component.

2.2. Stage 1: Separation and Soft Restitch

Many deep separation architectures are available for the front end. We adopt **TIGER** [8] for its favorable accuracy–efficiency trade-off: compared with TF-GridNet, TIGER

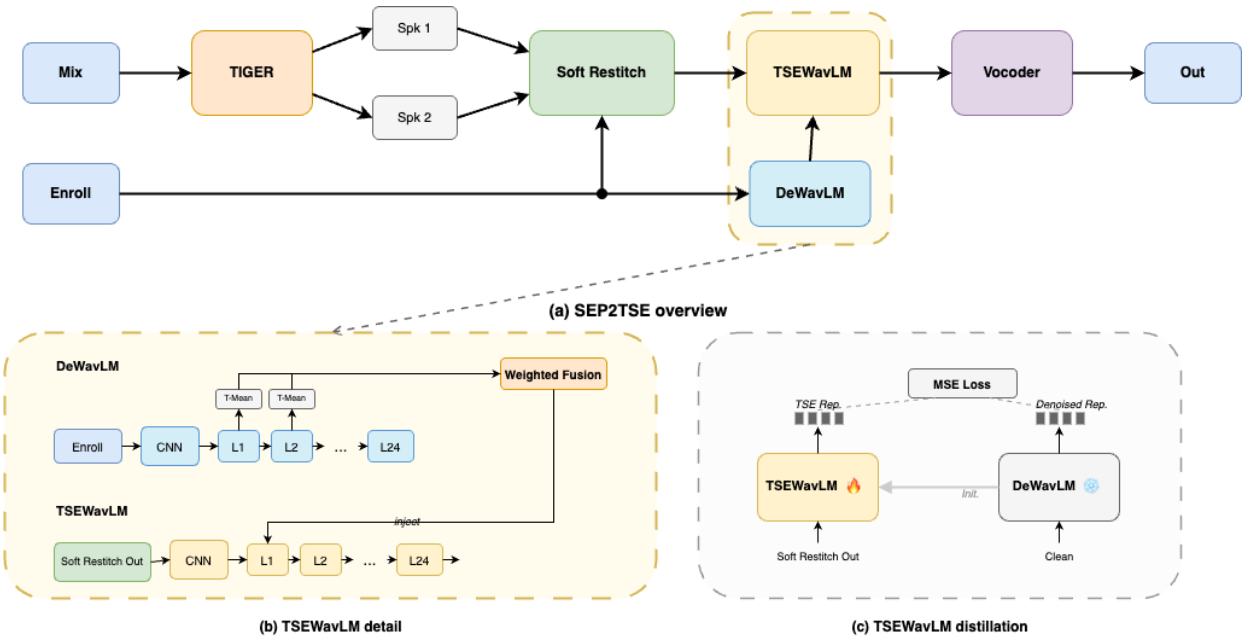

Fig. 1. Sep2TSE system overview (a), TSEWavLM encoder detail (b), and enrollment-conditioned distillation (c).

reduces parameters by 94.3% and multiply–accumulate operations (MACs) by 95.3%. TIGER also outperforms existing SOTA models in complex acoustic environments, which aligns with the reverberant and overlapping speech conditions.

In our implementation, long utterances are processed chunk-wise with overlap rather than in a single forward pass. Within a short segment, simultaneous talk from three or more speakers is relatively uncommon; a two-output separator therefore covers most real-world scenarios while keeping computation tractable. Accordingly, TIGER produces two candidate streams per chunk (Spk 1 / Spk 2 in Fig. 1(a)).

To obtain a temporally coherent target track, we fuse the two streams chunk-wise with enrollment-guided **soft restitch**. WeSpeaker ResNet34 [9] computes cosine similarities between the enrollment and Spk 1 / Spk 2, yielding s_1 and s_2; softmax weights $w_i = \exp(s_i/\tau)/\sum_{j\in\{1,2\}}\exp(s_j/\tau)$ blend chunk waveforms $\mathbf{x}_1$ and $\mathbf{x}_2$ into a pre-fusion signal $\tilde{\mathbf{x}} = w_1\mathbf{x}_1 + w_2\mathbf{x}_2$. With $s_{\max} = \max(s_1, s_2)$ and trust tiers $(\tau_h, \tau_m, \tau_l) = (0.7, 0.65, 0.6)$, the chunk output $\mathbf{y}$ is:

$$\mathbf{y} = \tilde{\mathbf{x}}, \qquad s_{\max} \geq \tau_h \quad \text{(trust-high)}, \tag{1}$$

$$\mathbf{y} = c_{\text{mid}}\,\tilde{\mathbf{x}}, \qquad \tau_m \leq s_{\max} < \tau_h \quad \text{(trust-mid)}, \tag{2}$$

$$\mathbf{y} = c_{\text{low}}\,\tilde{\mathbf{x}}, \qquad s_{\max} < \tau_m \quad \text{(trust-low)}, \tag{3}$$

where $c_{\text{mid}} = 0.85 + 0.15\,\frac{s_{\max}-\tau_m}{\tau_h-\tau_m}$ and $c_{\text{low}} = 0.45 + 0.40\,\frac{s_{\max}-\tau_l}{\tau_m-\tau_l}$ if $\tau_l \leq s_{\max} < \tau_m$, else $c_{\text{low}} = 0.15\,\frac{s_{\max}}{\tau_l}$. Tier-dependent temperatures $\tau \in \{\tau_h, \tau_m, \tau_s\}$ sharpen or soften the softmax blend. Overlapping chunks are overlap-added to form the *soft restitch out* fed to Stage 2.

2.3. Stage 2: TSEWavLM and Vocoder

Stage 2 further suppresses residual interference, dereverberates, and denoises the soft-restitched waveform to recover a clean target estimate. Our design follows the generative paradigm of PASE [10]: instead of estimating a time–frequency mask, a WavLM-based encoder produces enhanced latent representations that a neural vocoder reconstructs into waveform. PASE adapts WavLM into a denoising expert via representation distillation and guides a dual-stream vocoder with complementary phonetic and acoustic features. We extend this idea to target speaker extraction by conditioning TSEWavLM on enrollment cues extracted by a frozen DeWavLM branch.

2.3.1. Representation distillation

Fig. 1(b) shows how enrollment guides TSEWavLM. Because meeting enrollments are often noisy, a **frozen DeWavLM** encoder [10]—pretrained for denoising representation distillation—extracts shallow acoustic features (timbre and prosody) from early layers; **T-Mean** pooling and **weighted fusion** inject them into TSEWavLM to provide a denoised, identity-focused cue.

TSEWavLM is trained with the same distillation paradigm (Fig. 1(c)): a frozen teacher targets clean-speech features; the student, initialized from the teacher, predicts final-layer representations from soft-restitch outputs via an MSE loss.

2.3.2. Dual-stream vocoder reconstruction

Following PASE [10], a dual-stream vocoder reconstructs the waveform from TSEWavLM features: the final layer supplies **phonetic** content for intelligibility, while an early layer supplies **acoustic** cues (timbre and prosody) for speaker consistency. The acoustic stream is linearly projected and added to the phonetic stream before decoding with a PASE-style dual-stream vocoder.

3. EXPERIMENTS

3.1. Experimental setup

3.1.1. Datasets

Training data. The training data are mixtures generated on the fly from multilingual clean speech. Chinese utterances are drawn from AISHELL [11], AISHELL-3 [12], and the Interview, Speech, and Vlog partitions of CN-Celeb [13]; English utterances are drawn from LibriSpeech [14]. Background noise and room impulse responses (RIRs) are taken from the DNS Challenge corpus [15]. For each training sample, a target and an interfering speaker are mixed with controlled same-gender pairing (70% same-gender, 30% opposite-gender). The mixture signal-to-interference ratio (SIR) is uniformly sampled in $[-5, 15]$ dB and the signal-to-noise ratio (SNR) in $[-5, 10]$ dB. With probability 0.6, reverberation is applied independently to the target and interferer before mixing. Enrollment utterances are augmented with additive noise at SNR $\in [0, 15]$ dB and, with probability 0.6, additional reverberation.

3.1.2. Training procedure

Training focuses on Stage 2; the three steps are:

1. **Separation (no fine-tuning).** Stage 1 adopts the publicly released TIGER checkpoint [8] as-is; soft restitch is applied at inference without any additional training.

2. **TSEWavLM distillation.** The student is initialized from a pretrained DeWavLM teacher. On on-the-fly mixtures with augmented enrollments (Sec. 3.1.1), TSEWavLM is optimized with an MSE loss to match the teacher's final-layer representations on clean target speech, while enrollment cues from the frozen DeWavLM branch are injected into the student.

3. **Vocoder training.** With TSEWavLM frozen, a PASE-style dual-stream vocoder is trained to reconstruct clean waveforms from the fused phonetic and acoustic encoder features.

Table 1. Development-set results on REAL-T-dev.

System	TER↓	Timing F1↑	SIM↑	DNSMOS OVRL↑
BSRNN_EMB	0.693	0.841	0.501	1.66
BSRNN_TFMAP	0.703	0.838	0.521	1.50
Sep2TSE	**0.488**	**0.870**	0.510	**2.45**

Table 2. Official evaluation-set results on the REAL-T-EVAL

System	TER↓	Timing F1↑	SIM↑	DNSMOS OVRL↑
BSRNN_EMB	0.829	0.829	0.417	1.844
BSRNN_TFMAP	0.838	0.829	0.443	1.612
Sep2TSE	**0.670**	**0.847**	**0.471**	**2.522**

3.2. Experimental results

Tables 1 and 2 report development- and evaluation-set results, respectively. We compare **Sep2TSE** against two challenge-provided BSRNN baselines [7]. Metrics follow the challenge protocol: TER from Zipformer-ZH/EN ASR (lower is better), Timing F1, speaker similarity (SIM), and DNSMOS OVRL.

On REAL-T-dev (Table 1), Sep2TSE substantially reduces TER (0.488 vs. 0.693/0.703) and improves DNSMOS OVRL (2.45 vs. 1.66/1.50), while maintaining competitive speaker similarity (0.510 vs. 0.501/0.521) and the highest Timing F1 (0.870). On the official evaluation leaderboard (Table 2), the same trend holds: Sep2TSE achieves the best TER, Timing F1, SIM, and DNSMOS OVRL among the three systems, ranking **5th** overall (dense-rank average across the four metrics).

4. CONCLUSION

We presented **Sep2TSE**, a separation-first cascaded TSE system for the REAL-T challenge. Stage 1 raises SIR with a pretrained TIGER separator and enrollment-guided soft restitch; Stage 2 applies enrollment-conditioned TSEWavLM with DeWavLM distillation and a dual-stream vocoder to denoise and reconstruct the target. On the offline leaderboard our system ranked 5th overall; Sep2TSE outperforms the BSRNN baselines on both REAL-T-dev and the official evaluation set.

Several design choices remain open. The current soft restitch relies on hand-crafted similarity thresholds and tiered fusion rules; an alternative is to bypass this step and feed both separated streams directly into Stage 2, allowing a learnable backend to decide how much of each component to retain. Stage 2 itself is built on a PASE-style WavLM–vocoder pipeline, but other TSE architectures may offer complementary trade-offs under real meeting conditions. Joint optimization across separation, fusion, and extraction, as well as tighter alignment with challenge metrics such as WER and SpkSim, are promising directions for further improvement.

5. REFERENCES

[1] Y. Ju, W. Rao, X. Yan, Y. Fu, S. Lv, L. Cheng, L. Xie, Y. Wang, and S. Shang, "TEA-PSE: Tencent-Ethereal-Audio-Lab personalized speech enhancement system for ICASSP 2022 DNS challenge," in *Proc. ICASSP*, 2022.

[2] Y. Ju, S. Zhang, W. Rao, Y. Wang, T. Yu, L. Xie, and S. Shang, "TEA-PSE 2.0: Sub-band network for real-time personalized speech enhancement," in *Proc. SLT*, 2023.

[3] Y. Ju, J. Chen, S. Zhang, S. He, W. Rao, W. Zhu, Y. Wang, T. Yu, and S. Shang, "TEA-PSE 3.0: Tencent-Ethereal-Audio-Lab personalized speech enhancement system for ICASSP 2023 DNS challenge," in *Proc. ICASSP*, 2023.

[4] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, 2023.

[5] Z. Bang et al., "USEF-TSE: Universal speaker embedding free target speaker extraction," *arXiv preprint arXiv:2409.02615*, 2024.

[6] L. Yu, W. Zhang, C. Du, L. Zhang, Z. Liang, and Y. Qian, "Generation-based target speech extraction with speech discretization and vocoder," in *Proc. ICASSP*, 2024.

[7] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1893–1901, 2023.

[8] M. Xu, K. Li, G. Chen, and X. Hu, "TIGER: Time-frequency interleaved gain extraction and reconstruction for efficient speech separation," in *Proc. ICLR*, 2025.

[9] 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 *Proc. ICASSP*, 2023.

[10] X. Rong, Q. Hu, M. Yesilbursa, K. Wojcicki, and J. Lu, "PASE: Leveraging the phonological prior of WavLM for low-hallucination generative speech enhancement," 2025.

[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 *Proc. O-COCOSDA*, 2017.

[12] Y. Shi, H. Bu, X. Na, L. Li, X. Zhang, and J. Du, "AISHELL-3: A multi-speaker Mandarin TTS corpus," in *Proc. INTERSPEECH*, 2020.

[13] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Yu, "CN-Celeb: A challenging Chinese speaker recognition dataset," in *Proc. ICASSP*, 2020.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[15] H. Dubey et al., "ICASSP 2023 deep noise suppression challenge," *IEEE Open J. Signal Process.*, vol. 5, pp. 725–737, 2024.