REAL-TSE Challenge – Track 2

Official System Description

Team Name	CUDA_OUT_OF_MEMORY
Affiliation	National Taiwan Normal University, Speech and Machine Intelligence Laboratory (SMIL)
Members	Hua Chiun-Yu, Tsai Hung-Shao, Kuo Bing-Long, Tsai Cheng-Han, Tsai Ping-Chun, Mao Wei-Jun
Track	Track 2

1. System Overview

Our final submitted system performs target speech extraction (TSE) by directly applying SoloSpeech [1], a pretrained cascaded generative pipeline, to the mixture and enrollment audio pairs provided by the REAL-TSE Challenge, without any task-specific fine-tuning. SoloSpeech integrates compression, extraction, reconstruction, and correction stages, and is reported to generalize well to out-of-domain, real-world recordings. On the DEV set, this approach achieved a Target-source Error Rate (TER) of 0.682, substantially outperforming both the official baseline (BSRNN_EMB, TER = 0.770) and every alternative pipeline we constructed during development.

This report summarizes the development trajectory leading to this system, including approaches that were explored and ultimately not adopted, before describing the final configuration used for the EVAL1 and EVAL2 submissions.

2. Explored Approaches

Several extraction strategies were evaluated on the DEV set before settling on the final system. A summary is provided below; full implementation details are deferred to the accompanying paper submission.

2.1 Off-the-Shelf Pretrained TSE Models

Prior to adopting SoloSpeech, we benchmarked several publicly available pretrained TSE models directly on the DEV set, including xTF-GridNet [2], USEF-TSE [3], and UniSE [4]. As summarized in Table 1 below, all of these models underperformed the official baseline across TER, speaker similarity, and DNSMOS, with TER ranging from 0.890 to 1.144 compared to the baseline's 0.770. This indicated that most off-the-shelf TSE checkpoints, typically trained on simulated or non-meeting corpora, generalize poorly to the real, multi-speaker meeting recordings used in REAL-TSE.

System	TER ↓	SIM ↑	DNSMOS OVRL ↑	F1 ↑
Baseline (EMB)	0.693	0.501	1.66	0.841
Baseline (TFMAP)	0.703	0.521	1.50	0.838
xTF-GridNet	1.144	0.099	1.261	0.603

USEF-TSE	0.890	0.347	1.376	0.834
UniSE	0.988	0.383	2.982	0.837

Table 1: DEV-set comparison of off-the-shelf pretrained TSE models against the official baseline. F1 denotes the mean timing-segment F1 score.

2.2 Speaker-Diarization-Based Separation (pyannote)

We built a pipeline using pyannote's speech-separation model [5] to decompose each mixture into per-speaker sources. Speaker selection was then performed by computing a weighted combination of two scores for each candidate source: (1) cosine similarity between the candidate's ECAPA-TDNN speaker embedding and the enrollment embedding, and (2) the candidate speaker's estimated speaking duration within the mixture, as inferred from pyannote's diarization output. The final selection score was computed as:

score = 0.7 × sim_norm + 0.3 × dur_norm

where sim_norm and dur_norm are min-max normalized similarity and duration values respectively. The candidate with the highest score was selected as the extracted speech. The best variant reached TER = 0.753 on DEV, which outperformed the official baseline (TER = 0.770).

2.3 Fine-Tuning on In-Domain Data

We attempted two rounds of in-domain fine-tuning. First, UniSE [4] was fine-tuned on real mixture/target pairs, although TER slightly dropped from 0.988 (pretrained) to 0.940 (constructed pairs), it worsened to 0.993 when using a stricter “real-aligned” pairing scheme, with RATIO F1 collapsing from 0.837 to as low as 0.091—indicating the fine-tuned model frequently failed to produce timing-consistent output. Second, we fine-tuned the official BSRNN_EMB baseline on real mixture/target pairs constructed from the CHiME6, AMI, and AliMeeting training sets, using time-aligned far-field/close-talk microphone pairs. This consistently degraded performance relative to the unmodified baseline across all datasets, including DipCo, which was never part of fine-tuning data. Root-cause analysis identified the following issues: (i) AISHELL-4 lacks close-talk recordings, forcing mixture-equals-target pairs that taught the model an identity mapping; (ii) a subset of AliMeeting sessions exhibited substantial near/far microphone misalignment (up to several minutes), corrupting roughly 20% of fine-tuning pairs; (iii) the combination of these factors led to catastrophic forgetting of the model's pretrained separation capability. This direction was not pursued further within the challenge timeline.

2.4 SoloSpeech → Baseline Cascade

To test whether SoloSpeech's output could serve as a cleaner input for the baseline TSE model, we cascaded SoloSpeech's DEV-set predictions into both BSRNN_EMB and BSRNN_TFMAP. Counter-intuitively, this degraded TER to 0.886 and 0.878 respectively, worse than either model alone, and noticeably degraded DNSMOS as well. We attribute this to the baseline models being trained to perform extraction from multi-speaker mixtures; feeding them an already near-single-speaker input falls outside their training distribution and appears to introduce processing artifacts rather than further improvement. This cascade approach was discarded.

2.5 Pre/Post-Processing with Speech Enhancement

Denoising the mixture prior to separation (via DeepFilterNet) and applying speech enhancement to separated outputs were both evaluated on the pyannote pipeline. Neither produced a consistent improvement over the un-enhanced pipeline at full scale, despite encouraging behavior on small subsets; we attribute this discrepancy to high inter-dataset variance, particularly within CHiME6.

3. Final System: SoloSpeech

3.1 Model Description

SoloSpeech [1] is a cascaded generative pipeline for target speech extraction comprising four stages: (1) a compressor that encodes waveforms into a compact latent representation via a VAE-style autoencoder; (2) an extractor, a diffusion-based model conditioned on a speaker embedding (ECAPA-TDNN) derived from the enrollment utterance, which generates the latent representation of the target speaker's speech from the mixture latent; (3) a candidate re-ranking step, in which multiple extraction candidates are sampled and the one with highest speaker-embedding cosine similarity to the enrollment is retained; and (4) a corrector, a score-based generative model that refines the extracted waveform in the spectral domain to improve perceptual quality. We used the publicly released pretrained checkpoints without any additional fine-tuning on REAL-TSE data.

3.2 Inference Configuration

We used a DDIM sampler with num_infer_steps = 200 and num_candidates = 4 as the default configuration for the re-ranking step. For mixture/enrollment pairs that triggered CUDA out-of-memory errors during inference — typically arising from longer audio durations (e.g., enrollment utterances exceeding 40 seconds) — we fell back to num_candidates = 1 to complete extraction for the affected samples. This fallback was applied to a small subset of EVAL1 (DipCo) and EVAL2 utterances; the corrector stage and all other settings were kept unchanged.

4. Results

4.1 DEV Set Performance

Table 1 compares our final system against the official baseline and the best-performing alternative pipeline (pyannote-based separation) on the DEV set, aggregated across all five datasets (AISHELL-4, AMI, AliMeeting, CHiME6, DipCo).

System	TER ↓	SIM ↑	DNSMOS OVRL ↑
Official Baseline (BSRNN_EMB)	0.770	0.501	1.660
pyannote-based separation (best)	0.753	0.204	1.524
SoloSpeech (final submission)	**0.682**	**0.379**	**2.499**

Table 2: DEV-set comparison (1,991 utterances across 5 datasets). SIM denotes ECAPA-TDNN cosine similarity between extracted output and enrollment; DNSMOS OVRL is the overall perceptual quality score.

4.2 Official Leaderboard (EVAL)

Table 2 reports the official challenge leaderboard results for our submission (Team: CUDA_OUT_OF_MEMORY).

TER ↓	Timing F1 ↑	SIM ↑	DNSMOS OVRL ↑	Overall Score ↓
0.827 (rank 18)	0.819 (rank 16)	0.364 (rank 20)	2.643 (rank 5)	**14.75 (rank 14)**

Table 3: Official EVAL leaderboard results. Our system achieved particularly strong perceptual quality (DNSMOS OVRL, rank 5), reflecting SoloSpeech's generative correction stage, while TER and SIM ranked comparatively lower—likely reflecting the gap between DEV and EVAL data distributions and the num_candidates fallback applied to a subset of long-duration utterances.

5. Conclusion

We submit SoloSpeech, used in its pretrained, off-the-shelf form, as our final Track 2 system. Despite exploring several directions to better adapt extraction to the REAL-TSE domain—including in-domain fine-tuning, model cascading, and pre/post-processing with speech enhancement—none surpassed the performance of SoloSpeech's pretrained generalization, underscoring its robustness to out-of-domain, real-world meeting recordings. A more detailed analysis of all explored methods, including failure modes and ablations, will be presented in our accompanying paper submission.

References

[1] H. Wang, J. Hai, D. Yang, C. Chen, K. Li, J. Peng, T. Thebaud, L. Moro Velazquez, J. Villalba, and N. Dehak, “SoloSpeech: Enhancing Intelligibility and Quality in Target Speech Extraction through a Cascaded Generative Pipeline,” arXiv:2505.19314, 2025.

[2] “X-TF-GridNet: A Time-Frequency Domain Target Speaker Extraction Network with Adaptive Speaker Embedding Fusion,” 2024.

[3] “USEF-TSE: Universal Speaker Embedding Free Target Speaker Extraction,” 2024/2025.

[4] “UniSE: A Unified Framework for Decoder-only Autoregressive LM-based Speech Enhancement,” 2025.

[5] pyannote/speech-separation-ami-1.0, https://huggingface.co/pyannote/speech-separation-ami-1.0