The AGH-JHU System Description for REAL-TSE Challenge 2026

Magdalena Rybicka*†, Tianyu Cao†, Jesús Villalba†, Najim Dehak†
*Institute of Electronics, AGH University of Krakow, Krakow, Poland
†Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
Email: mrybicka@agh.edu.pl

***Abstract*—In the following we describe systems used to build a submission for REAL-TSE Challenge by AGH-JHU team. All our systems were based on the SoloSpeech target-speaker extraction system with different modifications. Our submission was part of the Track 2: Offline Target Speaker Extraction.**

***Index Terms*—target-speaker extraction, generative model, diffusion model, SoloSpeech**

I. Introduction

This document describes systems submitted by AGH-JHU team to the REAL-TSE Challenge, Track 2: Offline Target Speaker Extraction. The core of the proposed system is SoloSpeech [1] target-speaker extraction system that is based on the cascaded generative pipeline. We develop five TSE systems that differ in their training data and the additional processing techniques employed. As preprocessing, we investigate speech enhancement of the enrollment recordings and the incorporation of Target-Speaker Voice Activity Detection (TS-VAD) applied to the input mixture. As postprocessing, to mitigate speaker confusion where the model extracts an interfering speaker instead of the target speaker, we propose a residual speech extraction mechanism that performs a second target-speaker extraction on the residual mixture. Finally, we introduce a simple speaker similarity score-based system fusion strategy. For each system output, we compute the speaker embedding similarity between the extracted speech and the enrollment utterance. The output with the highest similarity score is selected as the final prediction.

II. Target-Speaker Extraction

The target-speaker extraction (TSE) pipeline is based on SoloSpeech [1]. SoloSpeech adopts a cascaded architecture consisting of a compressor, a target-speaker extractor, and a corrector. The compressor is a time-frequency variational autoencoder (VAE) that maps the input waveform into a latent representation. The target-speaker extractor is based on a diffusion model whose task is to estimate the latent representation of the target speaker conditioned on the enrollment utterance. The VAE decoder then reconstructs the waveform from the estimated latent representation. Following the original SoloSpeech pipeline [1], four candidate outputs are generated for each input and reranked according to their speaker similarity to the enrollment utterance. To assess the similarity, an ECAPA-TDNN[1] speaker verification model [2] is used. The candidate with the highest similarity score is selected and subsequently refined by the corrector, which is also implemented as a time-frequency diffusion model. The corrector further enhances the reconstructed speech and improves its perceptual quality. All architectural settings and hyperparameters follow the pretrained models released by the authors[2].

III. Additional processing techniques

A. Speech enhancement of the enrollment recording

As one of the investigated preprocessing techniques, we apply speech enhancement to the enrollment recordings. Since the recordings provided in the challenge contain background noise, speech enhancement is used to improve their quality before they serve as reference signals for target-speaker extraction. The speech enhancement pipeline is based on DiT-Flow [3]. It is a latent-space generative enhancement system composed of an audio compressor, a flow-matching enhancement module, and an audio decompressor. The same compressor and decompressor were used as the ones for Target-Speaker Extraction. The enhancement module is based on flow matching with a Diffusion Transformer backbone, whose task is to estimate the clean speech latent from the distorted latent by learning a time-dependent velocity field and solving the corresponding ODE during inference. The decompressor reconstructs the enhanced waveform from the estimated clean latent. The model was trained on StillSonicSet [3] at 16 kHz, a newly constructed synthetic dataset with acoustically realistic conditions by incorporating complex room geometries, with reverberation, noise, and codec-compression distortions. The architecture setup can be found in [3].

B. Target-Speaker Voice Activity Detection

To guide the TSE pipeline when the target speaker is active, one of our systems incorporates Target-Speaker Voice Activity Detection (TS-VAD) applied to the mixture recording. We adopt a standard embedding-based approach. First, a reference speaker embedding is extracted from the enrollment utterance. The input mixture is then segmented into 1.5 s windows with a 0.25 s hop size, and a speaker embedding is extracted from each segment. The cosine similarity between the segment embedding and the enrollment embedding is computed to

[1]https://huggingface.co/yangwang825/ecapa-tdnn-vox2

[2]https://huggingface.co/OpenSound/SoloSpeech-models

estimate target speaker activity. To determine the start and end of speech regions, we apply hysteresis thresholding with empirically selected activation and deactivation thresholds of 0.10 and 0.05, respectively, determined on the development set. Rather than completely suppressing the non-target regions, we attenuate them by 20 dB, which mitigates potential TS-VAD errors while preserving sufficient speech information for the downstream TSE model. Speaker embeddings are extracted using the ECAPA-TDNN model [2].

C. Target-Speaker Extraction from Residual Speech

One of the failures of target-speaker extraction is speaker confusion, where the model extracts the speech of an interfering speaker instead of the target speaker. To mitigate this problem, we propose target-speaker extraction from residual speech mechanism.

After obtaining an initial estimate of the target speaker, we subtract it from the input mixture to obtain a residual signal,

$$\mathbf{r} = \mathbf{m} - g\mathbf{e}, \tag{1}$$

where $\mathbf{m}$ denotes the input mixture, $\mathbf{e}$ is the estimated target-speaker signal, and g is a gain factor that compensates for possible energy mismatch between the estimated signal and the mixture,

$$g = \frac{\mathbf{m}^\top \mathbf{e}}{|\mathbf{e}|^2}. \tag{2}$$

The residual signal is then used as the initialization for the diffusion process in the second TSE pass, instead of random initialization. Since the interfering speaker has been partially removed, the second extraction may recover the target speaker when the initial estimate corresponds to the wrong speaker. To determine which extraction corresponds to the target speaker, we compute the cosine similarity between the speaker embedding extracted from each estimated waveform and the enrollment embedding. The estimate with the higher similarity score is selected as the final output. Speaker embeddings are extracted using the ECAPA-TDNN [2] model.

D. System Fusion

The submitted systems exhibit complementary strengths, with each performing better on different types of examples. Therefore, instead of relying on a single model, we perform output selection based on speaker similarity. For each system output, we extract a speaker embedding and compare it with the enrollment embedding using cosine similarity. Both embeddings are extracted with the ECAPA-TDNN model [2]. The output achieving the highest similarity score is selected as the final prediction.

IV. Datasets

For training, we used the Libri2Mix dataset [4] together with the AISHELL-1 corpus [5]. From Libri2Mix, we used the train-360 subset with added WHAM! noise [6], sampled at 16 kHz. The resulting training, validation, and test sets contain approximately 212 h, 11 h, and 11 h of audio, corresponding to 50800, 3000, and 3000 mixtures, respectively.

Since Libri2Mix contains only English speech, whereas the challenge includes both English and Mandarin, we additionally incorporated AISHELL-1. As AISHELL-1 is a corpus designed for single-speaker speech recognition, we generated a simulated two-speaker mixture dataset, referred to as AISHELL2Mix. Two utterances from different speakers were randomly selected and mixed, while preserving the original train, development, and test speaker partitions to avoid speaker overlap across splits. All mixtures were generated at a sampling rate of 16 kHz with additive WHAM! noise.

To create each mixture, the relative level between the two speakers was first adjusted by randomly sampling a signal-to-noise ratio (SNR) from the range [-5, 5] dB. Subsequently, the resulting speech mixture was mixed with WHAM! noise using a randomly selected SNR in the range [0, 5] dB. The final AISHELL2Mix dataset consists of 50000 training mixtures (52.31 h), 3000 development mixtures (3.18 h), and 1500 test mixtures (1.73 h).

TABLE I
Overview of the developed TSE systems.

System	Training data	Additional processing
A	Libri2Mix	Enrollment enhancement
B	Libri2Mix + AISHELL2Mix	–
C	Libri2Mix + AISHELL2Mix	Enrollment enhancement
D	Libri2Mix + AISHELL2Mix	TS-VAD on mixture
E	Libri2Mix + AISHELL2Mix	TSE from Residual Speech

V. Submitted systems and Results

Table I summarizes the configurations of the developed TSE systems. All systems were based on the SoloSpeech framework. The table presents the training data used for each system and the additional preprocessing or postprocessing applied. Inference was performed using 200 diffusion sampling steps, except System E where 40 steps were used.

Table II presents the performance of the individual systems and their combinations on the development set. Based on these results, we selected the best-performing configurations for evaluation, whose results are reported in Table III. Throughout the paper, systems are referred to by the labels A–E corresponding to the list above, while expressions such as A+B denote the score-based output selection (fusion) of the corresponding systems.

TABLE II
System results on the development set.

System	TER	Timing F1	SIM	DNSMOS OVRL
A	0.667	0.838	0.393	2.545
B	0.652	0.820	0.411	2.244
C	0.641	0.833	0.421	2.296
D	0.631	0.823	0.425	2.177
E	0.619	0.848	0.440	2.128
A + B + C + D	0.607	0.842	0.450	2.291
A + B + C + E	0.604	0.850	0.451	2.258
A + B + C + D + E	0.599	0.848	0.454	2.242

TABLE III
SYSTEM RESULTS ON THE EVALUATION SET.

System	TER	Timing F1	SIM	DNSMOS OVRL
A	0.799	0.835	0.380	2.701
A + B + C + D	0.743	0.837	0.434	2.513
A + B + C + E	0.759	0.842	0.437	2.444
A + B + C + D + E	0.742	0.840	0.440	2.436

Comparing Systems A and C, we observe that incorporating the AISHELL2Mix dataset into the training data improves the TER, F1, and SIM metrics. Comparing Systems B and C further shows that speech enhancement of the enrollment recordings provides additional performance gains. Likewise, comparing Systems B and D demonstrates that incorporating TS-VAD improves extraction performance. Among the individual systems, System E achieves the best overall performance. Finally, the proposed score-based system fusion further improves upon the best individual system, indicating that the evaluated systems exhibit complementary strengths. It is worth noting, however, that these improvements are not consistently reflected in the DNSMOS scores.

ACKNOWLEDGMENT

This research was supported by the National Science Centre, Poland under Grant 2021/42/E/ST7/00452 and 2023/49/B/ST7/04100 and by program ”Excellence initiative – research university” for the AGH University of Krakow. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational Grants PLG/2025/019035 and PLG/2025/018799.

REFERENCES

[1] H. Wang, J. Hai, D. Yang, C. Chen, K. Li, J. Peng, T. Thebaud, L. Moro-Velázquez, J. Villalba, and N. Dehak, “Solospeech: Enhancing intelligibility and quality in target speech extraction through a cascaded generative pipeline,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 2408–2420, 2026.

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

[3] T. Cao, H. Wang, A. Frummer, Y. Sieradzki, A. Arbel, L. M. Velazquez, J. Villalba, O. Gal, T. Thebaud, and N. Dehak, “Dit-flow: Speech enhancement robust to multiple distortions based on flow matching in latent space and diffusion transformers,” 2026. [Online]. Available: https://arxiv.org/abs/2603.21608

[4] 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.

[5] X. N. B. W. H. Z. Hui Bu, Jiayu Du, “AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline,” in *Oriental COCOSDA 2017*, 2017.

[6] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, “WHAM!: Extending Speech Separation to Noisy Environments,” in *Interspeech 2019*, 2019, pp. 1368–1372.