Causal Target Speaker Extraction with Real-Distortion Interference: The ADT_audio System for the REAL-TSE Challenge Track-1

Feng Xu
ADT_audio
Ant Digital Technologies
Hangzhou, China
shiqi.xf@antgroup.com

Weixian Li
ADT_audio
Ant Digital Technologies
Hangzhou, China
liweixian.lwx@antgroup.com

Zhihai Zhu
ADT_audio
Ant Digital Technologies
Hangzhou, China
buqi.zzh@antgroup.com

***Abstract*—This report describes the ADT_audio submission to the REAL-TSE Challenge Track-1. The task requires online target speaker extraction under a 100 ms algorithmic latency limit and ranks systems by the mean dense-rank of four partly competing metrics: transcription error rate, timing activity F1, speaker similarity, and DNSMOS overall. Starting from the official causal BSRNN baseline, we keep the architecture strictly causal (K=0, measured maximum future dependency 26.0 ms) and make the main changes in data composition, augmentation, and training objectives. The submitted system, `BSRNN_TFMAP_CAUSAL_realaug_vitw_ft`, uses 32k VitW (Voices-in-the-Wild-2M) real distorted recordings as interference-only speakers, while clean LibriLight and AISHELL-1 utterances remain the target pool. This keeps the target reference clean for SI-SDR and speaker losses while exposing the separator to real far-field and device-distorted interference. The loss combines SI-SDR with MRSTFT, phase, speaker, complex-spectrum, and consistency terms. On the 5,000-utterance EVAL set, the single submitted model obtains 0.738 TER, 0.853 Timing-F1, 0.483 SIM, and 1.938 DNSMOS-OVRL, giving a dense-rank score of 6.50 and fourth place overall. Timing-F1 ranks second among all teams.**

***Index Terms*—target speaker extraction, causal speech separation, real-world speech, data augmentation, domain mismatch**

I. Introduction

The REAL-TSE Challenge Track-1 [1] requires recovering a target speaker from a multi-speaker mixture using a short enrollment utterance, under a 100 ms algorithmic latency budget. Systems are evaluated by transcription error rate (TER), timing activity F1 (Timing-F1), speaker similarity (SIM), and DNSMOS overall (DNSMOS-OVRL) [2]. The final score is the mean dense rank over the four metrics, lower being better.

Several constraints shaped our design. First, the online track leaves little room for non-causal processing. Second, the metrics are not independent: aggressive suppression of non-target energy may improve perceived cleanliness, but it can also remove target speech and hurt TER or SIM [3]. Third, the evaluation audio is recorded in real conditions and carries far-field acoustics, codec artifacts, bandwidth limits, and device-specific coloration that clean speech plus ordinary additive noise does not fully cover. Finally, the challenge rules require the evaluation set to be processed by a fixed submitted system, without per-utterance metric-driven optimization or candidate selection.

We made three official submissions. The first, `BSRNN_TFMAP_CAUSAL_realaug_ft` (June 16), strengthened the synthetic augmentation recipe and reached rank 6 overall with EVAL DNSMOS 2.051. The second, `BSRNN_TFMAP_CAUSAL_la_lookahead_ft` (June 24), added K=2 look-ahead frames; it improved DEV DNSMOS but transferred only a small DNSMOS gain to EVAL and regressed on TER and SIM. The final submission, `BSRNN_TFMAP_CAUSAL_realaug_vitw_ft` (June 26), replaced the intensified synthetic-only route with real distorted VitW interference and a structured multi-objective loss. It gave the most balanced EVAL result among our three submissions and is the system described below.

II. Architecture

The model is TSE_BSRNN_SPK, structurally identical to the official `tfmap_context_causal_100` baseline. The separator is a band-split RNN [4] operating on the complex STFT with a 512-sample window and 128-sample hop, corresponding to 32 ms and 8 ms at 16 kHz. The separator uses 128-dimensional features and 6 repeated blocks. The speaker branch is ECAPA-TDNN [5] (ECAPA_TDNN_GLOB_c512), mapping 80-dimensional filter-bank features from the enrollment utterance to a 192-dimensional speaker embedding. Speaker information is injected into the separator through tfmap multiplicative fusion together with a 512-dimensional context embedding. The waveform is reconstructed by inverse STFT.

All convolutions are strictly causal (K=0, no look-ahead). The official `latency_verification` script measures

Participant: Feng Xu; Team: ADT_audio. REAL-TSE Challenge Track-1 (Online).

maximum future dependency at 26.0 ms, with 23.0 ms minimum and 24.4 ms mean, below the 100 ms limit. We tested K=2 look-ahead during development; it raised DEV DNSMOS to about 2.024, but EVAL reached only 2.074 and TER/SIM regressed. Its measured latency was about 90 ms, leaving little margin. We therefore kept the submitted model at K=0.

III. Data Strategy

The data strategy is the main difference from the official baseline. Rather than changing the separator, we changed which distortions the separator sees during training and how they enter the mixture.

A. Motivation

Our first submission used stronger synthetic augmentation on clean speech: MUSAN noise [6], FRAM-RIR reverberation, and enrollment degradation. It reached rank 6 with EVAL DNSMOS 2.051, but further intensification of the same recipe improved DEV DNSMOS while lowering EVAL DNSMOS. We read this as a sign that the model was learning the specific noises and reverberation patterns in the augmentation set, rather than the broader capture-chain variation in real recordings.

Real conversations include effects that additive noise and ordinary room impulse responses do not fully reproduce: far-field capture, device frequency response, bandwidth limitation, low-bitrate or telephone-grade coding, and other recording-chain artifacts. The VitW corpus [7] provides real far-field and device-distorted speech at scale, so we use it to bring these conditions into training.

B. Interference-Only VitW

VitW recordings do not have clean references. Using them as targets would make SI-SDR and speaker-similarity supervision ambiguous, because the target reference itself would contain unknown distortion and background. We therefore use VitW only as interference. In the sampler, speakers tagged with a `vitw_` prefix are excluded from the target slot and enter only the interference pool. This keeps target references clean while injecting real far-field and device characteristics into the mixtures.

C. Data Composition

The final training pool contains 152,000 single-speaker sources. LibriLight [8] contributes 72k English clean utterances and AISHELL-1 [9] contributes 48k Chinese clean utterances; both are candidate targets. VitW contributes 32k real distorted recordings as interference. The VitW subset covers four scenes: `far_field` (12k), `far_field_recording` (8k), `recording` (6k), and `distortion` (6k), with durations between 2 and 8 s. The echo scene is excluded because it introduced audible artifacts in pilot runs. The number of speakers per mixture follows 1/2/3–4 = 0.1/0.6/0.3, capped at 4.

IV. Augmentation Rules

Figure 1 shows the online generation pipeline. The augmentation recipe is organized as a set of composition rules rather than independent switches:

1) **Noise.** MUSAN noise is added with probability 0.4, reduced from the stronger synthetic recipe to avoid overfitting to the MUSAN distribution.
2) **Noise skip.** MUSAN is skipped when a VitW source is present, because VitW already carries real background noise and stacking synthetic noise would create an unrealistic mixture.
3) **Reverberation.** FRAM-RIR is applied with probability 0.5, rt60 sampled from $[0.1, 0.6]$ s, and room size from 3 to 8 m.
4) **Reverberation skip.** Synthetic RIR is applied to clean sources but skipped for VitW interference, since VitW recordings already include real room and device effects.
5) **TIR.** Target-to-interference ratio is sampled from $[-10, 5]$ dB, lower than the baseline $[-5, 10]$ dB range, to cover low-target-fraction mixtures.
6) **Gain.** A global gain from $[-12, 0]$ dB is applied after mixing.
7) **Enrollment degradation.** With probability 0.3, one or two device-style degradations are applied to the enrollment: 8 kHz bandwidth limiting and resampling back, peaking-EQ coloration, μ-law coding, or light real-RIR reverberation. This models cross-device enrollment mismatch.
8) **Segment length.** All segments are trimmed or padded to 3 s, or 48,000 samples at 16 kHz.

V. Loss and Training

A. Multi-Objective Loss

The training loss is a six-term weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{SI-SDR}} + 0.2\,\mathcal{L}_{\text{MRSTFT}} + 0.1\,\mathcal{L}_{\text{ph}} + 0.1\,\mathcal{L}_{\text{spk}} + 0.05\,\mathcal{L}_{\text{com}} + 0.05\,\mathcal{L}_{\text{con}}. \tag{1}$$

Let s be the clean target waveform and $\hat{s}$ be the estimate. The SI-SDR term follows

$$s_{\text{tar}} = \frac{\langle \hat{s}, s \rangle}{\|s\|_2^2} s, \quad e = \hat{s} - s_{\text{tar}}, \tag{2}$$

$$\mathcal{L}_{\text{SI-SDR}} = -10 \log_{10} \frac{\|s_{\text{tar}}\|_2^2}{\|e\|_2^2 + \epsilon}. \tag{3}$$

For spectral terms, let $X_r = \text{STFT}_r(s)$ and $\hat{X}_r = \text{STFT}_r(\hat{s})$ at resolution r. The MRSTFT term is

$$\mathcal{L}_{\text{MRSTFT}} = \frac{1}{R} \sum_{r=1}^{R} \Big(\Big\| |X_r| - |\hat{X}_r| \Big\|_1 + \Big\| \log(|X_r| + \epsilon) - \log(|\hat{X}_r| + \epsilon) \Big\|_1 \Big). \tag{4}$$

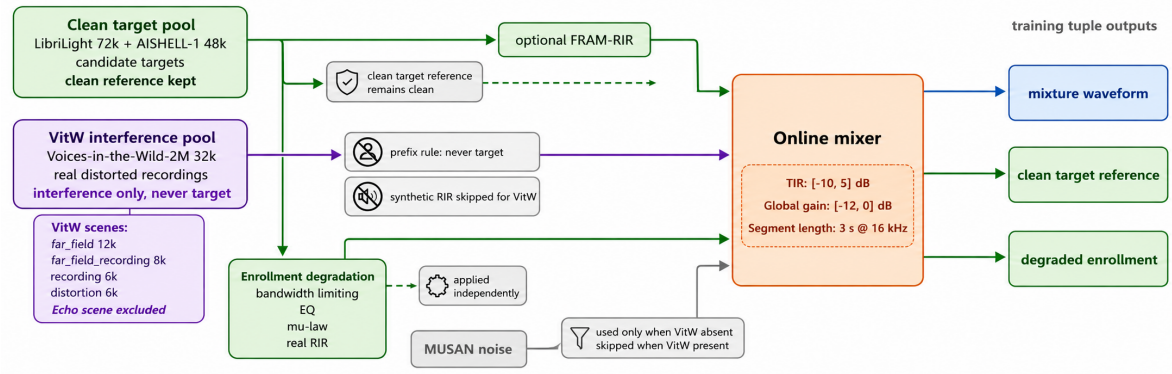

Fig. 1. Online data generation and augmentation. Clean LibriLight/AISHELL-1 utterances form the target pool. VitW recordings are tagged as interference-only and never enter the target slot. MUSAN and synthetic RIR are applied with skip rules to avoid stacking synthetic noise or reverberation on top of already distorted VitW recordings. Enrollment-side degradation is applied independently.

The phase, complex-spectrum, speaker, and consistency terms are

$$\mathcal{L}_{\text{ph}} = \left\| 1 - \cos(\angle X - \angle \hat{X}) \right\|_1, \tag{5}$$

$$\mathcal{L}_{\text{com}} = \left\| X - \hat{X} \right\|_1, \tag{6}$$

$$\mathcal{L}_{\text{spk}} = 1 - \frac{E(\hat{s})^\top E(s)}{\|E(\hat{s})\|_2 \|E(s)\|_2}, \tag{7}$$

$$\mathcal{L}_{\text{con}} = \left\| \hat{X} - \text{STFT}(\text{iSTFT}(\hat{X})) \right\|_1. \tag{8}$$

Here $E(\cdot)$ is the ECAPA speaker embedding extractor, and $X, \hat{X}$ denote the default-resolution complex spectra. All norms are averaged over the corresponding time-frequency bins and batch items. The auxiliary weights are intentionally small so that they guide training without overpowering SI-SDR [3].

B. Training Procedure

We start from the official baseline, fine-tune for 11 epochs on clean speech with synthetic augmentation to obtain an intermediate checkpoint, and then continue for 4 more epochs on the VitW-augmented mixture. Optimization uses Adam with an exponentially decaying learning rate from 2.0×10^{-5} to 1.0×10^{-5} and weight decay 10^{-4}. Training runs on 8 GPUs with distributed data parallel, batch size 8 per GPU, automatic mixed precision, gradient clipping at 5.0, and fixed seed 42.

The submitted model is a single seed, single checkpoint system: the final checkpoint of the last epoch is used. There is no weight averaging, model averaging, or ensembling. We tested checkpoint averaging over consecutive epochs and found it harmful: SIM dropped from 0.566 to below 0.38 and TER rose from 0.546 to above 0.77. We therefore did not treat neighboring checkpoints as safely interchangeable for this architecture.

VI. Inference and Post-Processing

At inference time, each utterance is processed once by the fixed submitted checkpoint. The only waveform reconstruction step is the model's own inverse STFT. We do not apply an additional denoiser, dereverberation model, enhancement model, loudness tuning, candidate reranking, or metric-based waveform refinement after the model output. In particular, there is no per-utterance post-processing selected with respect to DNSMOS, SIM, TER, Timing-F1, or any official evaluation model.

VII. Experiments

A. Setup

The challenge provides a DEV split of 1,991 utterances with references for local validation and an EVAL split of 5,000 utterances for official ranking. We use DEV for model development and diagnosis, and report the official leaderboard scores for EVAL. Latency is checked with the official `latency_verification` script.

B. Official EVAL Results

Table I shows the final EVAL results. The system places fourth overall with a dense-rank score of 6.50.

TABLE I
FINAL EVAL RESULTS FOR THE SUBMITTED SYSTEM.

Metric	Value	dense-rank
TER (↓)	0.738	6
Timing-F1 (↑)	0.853	2
SIM (↑)	0.483	5
DNSMOS-OVRL (↑)	1.938	13
Score	**6.50**	**4**

Timing-F1 is the submitted system's best-ranked metric: 0.853 ranks second among all teams and is within 0.002 of the best Timing-F1 score. The strictly causal 26 ms front end therefore did not prevent accurate target-activity timing. TER and SIM are competitive but not leading. DNSMOS-OVRL is the weakest metric and accounts for most of the remaining gap to the top systems.

C. Evolution Across Submissions

The three submissions trace the design path. The first submission, `BSRNN_TFMAP_CAUSAL_realaug_ft`, showed that stronger augmentation could help the overall system but also revealed limited EVAL headroom for a fixed synthetic recipe. The second, `BSRNN_TFMAP_CAUSAL_la_lookahead_ft`, tested K=2 look-ahead as an architectural change. DEV DNSMOS rose from 1.575 to 2.024 at epoch

11, but EVAL improved only slightly to 2.074 and TER/SIM regressed. The final submission, `BSRNN_TFMAP_CAUSAL_realaug_vitw_ft`, switched to real distorted VitW interference and gave the most balanced result across the four metrics.

D. DEV-EVAL Behavior

Table II summarizes the DNSMOS-OVRL pattern across representative development variants. The main observation is that DEV perceptual gains were not reliable predictors of EVAL perceptual gains.

TABLE II
DNSMOS-OVRL ACROSS REPRESENTATIVE VARIANTS. DEV GAINS DID NOT RELIABLY TRANSFER TO EVAL.

Variant	DEV OVRL	EVAL OVRL
`bsrnn_realaug_ft`	1.575	2.051
`bsrnn_la_lookahead_ft`	2.024	2.074
Perceptual loss + real RIR + CN data	1.892	1.955
`bsrnn_realaug_vitw_ft`	1.839	1.938

The background-noise component of DNSMOS showed the clearest DEV–EVAL split. On DEV, augmentation variants tended to raise BAK; on EVAL, the same family of changes lowered it, especially on unseen conversations. Within the VitW route, we also varied overlap ratio (timeline-controlled $[0.3, 1.0]$ instead of fixed full overlap) and relaxed TIR to $[-5, 5]$ dB. These changes did not improve EVAL DNSMOS. This suggests that the missing variation is not only mixture geometry, but the broader recording chain: bandwidth, codec, device response, and cross-device enrollment.

E. Alternatives Not Used

We explored several alternatives that were not included in the submitted system. A causal GTCRN backbone [10] with FiLM speaker conditioning was trained from scratch. Its training loss decreased, but DEV metrics did not reach the fine-tuned BSRNN: the best checkpoint reached DNSMOS 1.676, while SIM was about 0.524 and TER stayed near 0.62. A from-scratch causal backbone likely needs substantially more training data and time to match a strong fine-tuned BSRNN.

We also tested a speaker-consistency loss without the other auxiliary losses. It improved DEV SIM but hurt DNSMOS, reflecting the same suppression-versus-completeness trade-off seen in the official metrics: preserving more target identity can also leave more residual interference.

VIII. COMPLIANCE

The EVAL waveforms were produced by the fixed submitted model `BSRNN_TFMAP_CAUSAL_realaug_vitw_ft` in a single pass. The submitted system is an Online-track system: it uses K=0 causal processing, no look-ahead frames, and a verified maximum future dependency of 26.0 ms, below the 100 ms limit. We did not use per-utterance post-processing, metric-driven waveform refinement, candidate selection, or optimization based on official evaluation scores, gradients, embeddings, or model feedback. Official evaluation models and scripts were used only for allowed local validation and analysis during development, not as training objectives on evaluation utterances and not for selecting among candidates on the evaluation set.

IX. CONCLUSION

This report described the ADT_audio system for REAL-TSE Challenge Track-1. The submitted system keeps the official BSRNN architecture strictly causal, uses VitW recordings as interference-only real distorted data, and trains with a six-term objective centered on SI-SDR. It ranks fourth overall with a dense-rank score of 6.50, with Timing-F1 as the best-ranked metric. The main weakness is DNSMOS, where fixed augmentation and fixed real-distortion coverage did not transfer cleanly from DEV to EVAL. Future work should broaden capture-chain coverage–bandwidth limits, codec types, device profiles, and real rooms–rather than simply intensifying one augmentation corner.

ACKNOWLEDGMENT

The authors thank the REAL-TSE Challenge organizers for providing the dataset, baseline, rules, and evaluation infrastructure. This work used publicly available AISHELL-1, LibriLight, MUSAN, and VitW corpora.

REFERENCES

[1] REAL-TSE Challenge Organizing Team, "REAL-TSE challenge: Real-world target speaker extraction," IEEE SLT 2026 Challenge organizer materials, 2026, participant system-description instructions.

[2] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv preprint arXiv:2110.01763*, 2021, https://arxiv.org/abs/2110.01763.

[3] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[4] Y. Luo and J. Yu, "Music source separation with band-split RNN," *arXiv preprint arXiv:2209.15174*, 2022, https://arxiv.org/abs/2209.15174.

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

[6] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015, https://arxiv.org/abs/1510.08484.

[7] Z. Xie, K. Pang, H. Zhang, D. Ye, X. Hu, S. Yan, and C. Miao, "Mega-ASR: Towards in-the-wild^2 speech recognition via scaling up real-world acoustic simulation," 2026, introduces the Voices-in-the-Wild-2M dataset. Dataset page: https://huggingface.co/datasets/zhifeixie/Voices-in-the-Wild-2M.

[8] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, https://arxiv.org/abs/1912.07875.

[9] 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 *Oriental COCOSDA*, 2017, https://arxiv.org/abs/1709.05522.

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