AIVOX System Description for the REAL-TSE 2026 Online Track

Kaixing Ding, Xiaowei Fu, Aoyu Liu
Quectel Wireless Solutions Co., Ltd.
{kasen.ding, vic.fu, lay.liu}@quectel.com

***Abstract*—This report describes the AIVOX submission to the REAL-TSE 2026 online target speaker extraction track. The system is derived from the official causal TF-Map + Context BSRNN baseline and includes one architectural change: a zero-initialized causal temporal convolutional pre-filter placed before the temporal LSTM in each BSNet block. The baseline speaker-conditioning pathway, complex-mask estimation procedure, and online causality are otherwise retained. The model was trained with eligible open-source speech and noise data, dynamic two-speaker mixture generation, simulated reverberation, and an SI-SDR objective. Following the challenge definition, the algorithmic latency is 24 ms for the 512-sample STFT window and 128-sample hop at 16 kHz; including one 8 ms STFT-hop buffering step, the submitted total system latency is 32 ms. For the EVAL-1+EVAL-2 submission, the reported metrics were TER 0.751, Timing F1 0.847, speaker similarity 0.493, and DNSMOS OVRL 1.811.**

I. Submission Summary

REAL-TSE evaluates target speaker extraction in real conversational recordings [1], [3]. Given a multi-speaker mixture and an enrollment utterance from the target speaker, the task is to recover the target-speaker waveform. The online track further requires low-latency processing while handling overlap, reverberation, background noise, and conversational variability.

The submission was made under participant name kasen and team name AIVOX, matching the names shown on the challenge leaderboard. Table I summarizes the main technical configuration of the submitted system. The EVAL-1 and EVAL-2 waveforms were generated by a single fixed model, organized according to the official submission format, and submitted to the official evaluation server. Evaluation utterances were not used for training, fine-tuning, metric-driven waveform refinement, or utterance-level candidate selection.

TABLE I
Main configuration of the submitted system.

Item	Description
Track	Online target speaker extraction
Method	Causal Temporal Pre-Filter BSRNN
Base system	Official causal TF-Map + Context BSRNN
Added module	Causal temporal pre-filter in each BSNet block
Objective	SI-SDR loss
Sampling rate	16 kHz
Latency	24 ms algorithmic; 32 ms total submitted
Outputs	EVAL-1 and EVAL-2 waveforms

II. System Architecture

A. Overall Pipeline

The submitted system is a single-channel target speaker extraction model based on BSRNN-style subband processing [4]. As shown in Figure 1, the mixture waveform is first converted into a complex spectrogram. The system then forms target-speaker cues from the enrollment and mixture spectrograms, a causal BSRNN separator estimates a target complex mask, and the target waveform is reconstructed with inverse STFT.

The conditioning block in Figure 1 follows the released TF-Map + Context design [1], [5]. TF-Map is computed from the enrollment and mixture magnitude spectra, producing a one-channel cue that is concatenated with the real and imaginary mixture spectra before band splitting. In parallel, the Context branch extracts frame-level speaker representations from the enrollment waveform with an ECAPA-TDNN-style speaker encoder [10], [11]. These representations are aligned with the mixture subband features through cross-attention and fused by multiplicative speaker conditioning. The resulting target-conditioned subband representation is then passed to the causal BSRNN separator.

B. Causal BSRNN Backbone

The separator consists of six BSNet blocks with a feature dimension of 128 and a single target-speaker output. Within each block, temporal dependencies are modeled independently for each frequency band before information is exchanged across bands. To satisfy the online-track requirement, the temporal recurrent path is unidirectional and uses only current and past frames along the mixture time axis. The STFT uses a 512-sample window and a 128-sample hop at 16 kHz. The mask head estimates the real and imaginary components of the complex mask for each subband, and the mask is applied to the original mixture spectrum before iSTFT reconstruction.

C. Causal Temporal Convolutional Pre-filter

The only architectural modification to the official causal TF-Map + Context baseline is the causal temporal pre-filter shown in Figure 1b. The module is inserted before the temporal LSTM in every BSNet block. For an input tensor

$$x \in \mathbb{R}^{(B\cdot K)\times D\times T}, \tag{1}$$

where B denotes the batch size, K the number of frequency bands, $D = 128$ the feature dimension, and T the number of

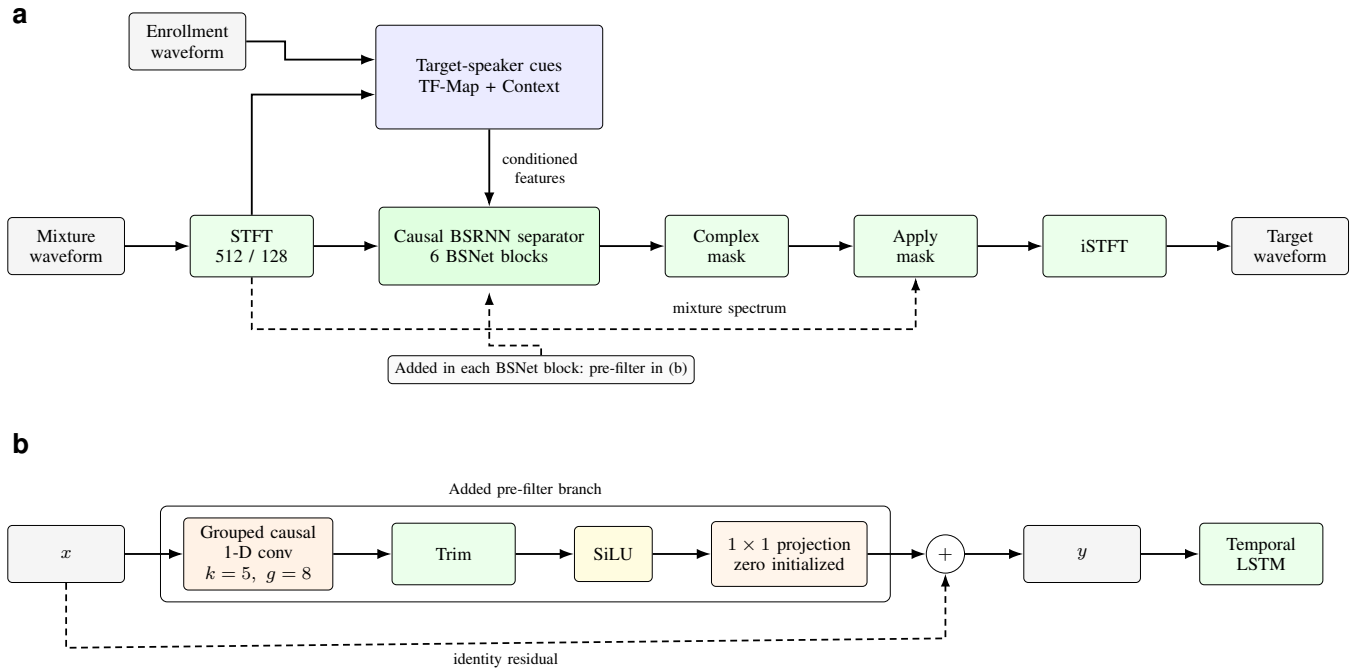

Fig. 1. Architecture of the submitted online target speaker extraction system. **a** Overall online TSE pipeline. The speaker-conditioning block represents the TF-Map cue and Context fusion used in the official TF-Map + Context baseline. TF-Map is computed from the mixture and enrollment magnitude spectra and concatenated with the real and imaginary mixture spectra before band splitting, whereas Context uses frame-level enrollment representations to condition subband features before separation. The dashed skip connection carries the original mixture spectrum for complex-mask application. **b** Causal temporal pre-filter added to each BSNet block. The pre-filter is inserted before the temporal LSTM and is initialized as an identity mapping through a zero-initialized 1×1 projection.

frames. The K subbands are folded into the batch dimension for per-band temporal modeling. The pre-filter computes

$$y = x + W_{1\times1}\left(\mathrm{SiLU}\left(\mathrm{Trim}_T\left(\mathrm{GConv1D}_{k=5,g=8}(x)\right)\right)\right). \quad (2)$$

The grouped 1-D convolution uses kernel size 5 and 8 groups. It is padded by $k-1$ frames and then trimmed back to the original length T, which is equivalent to left-causal padding and guarantees that frame t depends only on current and past frames. The subsequent 1×1 projection is initialized with zero weights and zero bias. Consequently, the added branch contributes zero at initialization, and the full module initially behaves as an identity mapping. This initialization allows the pre-filter to be inserted into a pretrained separator without changing its initial input–output function.

The pre-filter adds 26,880 parameters per BSNet block, or 161,280 parameters across all six blocks. This corresponds to approximately 0.6% of a 27M-parameter causal BSRNN separator. The module is intended to provide local temporal preprocessing before the LSTM, especially for short-time spectral transitions and target-speaker onsets, while leaving the speaker-conditioning path, mask estimation path, and online causality unchanged.

III. Training and Model Selection

A. Training Resources

The submitted system was initialized from the released causal TF-Map + Context BSRNN checkpoint and trained using only the data sources listed below. For dynamic mixture generation, the speech sources were LibriSpeech training splits [6], VoxCeleb1 dev [7], and VoxCeleb2 dev [8]. MUSAN [9] was used for additive noise augmentation, and reverberation was generated by online room simulation during training.

TABLE II
Data and resources used by the submitted system.

Resource	Usage in the submitted system
Released checkpoint	TF-Map + Context causal BSRNN initialization
Speech data	LibriSpeech train, VoxCeleb1 dev, VoxCeleb2 dev
Noise data	MUSAN augmentation, sampled with probability 0.35
Reverberation	Online simulated room augmentation, sampled with probability 0.35
Training validation	Held-out split from the same speech sources for SI-SDR monitoring
REAL-TSE DEV	Validation and model selection only
EVAL-1/EVAL-2	Final inference only

The submitted system does not use AliMeeting, AISHELL-4, AMI, DipCo, or CHiME6 for training, pretraining, fine-tuning, or augmentation.

B. Training Configuration

Both training stages used dynamic online mixture generation. Each training segment contained 48,000 samples,

corresponding to 3 seconds at 16 kHz, and each mixture contained two speakers. The SNR range was $[-5, 10]$ dB, and a global gain in the range $[-12, 0]$ dB was applied after peak normalization. Enrollment waveforms were used directly, without enrollment VAD, enrollment-side noise augmentation, or enrollment-side reverberation augmentation.

The final system was trained in two stages. In the first stage, the model was initialized from the released causal TF-Map + Context checkpoint and optimized with Adam, weight decay 10^{-4}, and a learning-rate schedule decaying from 10^{-4} to 5×10^{-6}. In the second stage, training resumed from the first-stage checkpoint with the same optimizer and weight decay, using a learning-rate schedule from 5×10^{-5} to 5×10^{-6}. Each stage used 102,400 dynamically sampled mixtures with batch size 4. Automatic mixed precision was not used.

C. Validation and Model Selection

A held-out split from the same open-source speech sources was used to monitor SI-SDR loss during training. REAL-TSE DEV was used only for validation and model selection with the released development evaluation pipeline. No EVAL-1 or EVAL-2 metric feedback was used to optimize individual utterances, select per-utterance candidates, or refine submitted waveforms.

IV. Inference and Latency

A. Inference and Post-processing

At inference time, the mixture and enrollment audio are loaded, resampled to 16 kHz when necessary, and processed by the fixed trained model. The estimated waveform is peak-normalized to 0.9 before being saved as a 16 kHz, mono, 16-bit PCM WAV file as required by the submission format. No ASR-, DNSMOS-, speaker-similarity-, or timing-metric-driven post-processing was applied to individual evaluation utterances.

B. Online Latency

The temporal recurrent path is unidirectional, and the added temporal pre-filter is causal. Thus, after STFT analysis, the separator introduces no future-frame lookahead along the mixture time axis.

The STFT configuration follows the official online BSRNN baseline, with a 512-sample window and a 128-sample hop at 16 kHz. This corresponds to a 32 ms window and an 8 ms hop. Under the challenge definition, the STFT/iSTFT algorithmic latency is therefore $32\ \mathrm{ms} - 8\ \mathrm{ms} = 24\ \mathrm{ms}$. Because the added pre-filter uses causal padding and trimming, it introduces no additional lookahead beyond the STFT/iSTFT delay. The algorithmic latency is therefore reported as 24 ms, below the 100 ms limit of the online track. The submitted latency metadata used a total system latency of 32 ms, consisting of the 24 ms algorithmic component plus one 8 ms STFT-hop buffering step.

The challenge latency verification script reported a future dependency of 24.4 ms on average, with a range of 23.3–25.7 ms across the checked utterances. This is consistent with the theoretical 24 ms algorithmic latency and indicates that neither the causal pre-filter nor the unidirectional temporal LSTM adds future-frame lookahead.

V. Evaluation Results

Table III reports the EVAL-1+EVAL-2 metrics returned by the official evaluation server for the submitted waveform package. The EVAL-1 and EVAL-2 references are not released, so these values were obtained through the official black-box evaluation process and were not re-computed locally. They match the metric values displayed on the leaderboard for the submitted entry, but no ranking information is included here because the leaderboard may be subject to organizer review.

TABLE III
Reported EVAL-1+EVAL-2 metric values for Team AIVOX.

Metric	Value
TER ↓	0.751
Timing F1 ↑	0.847
SIM ↑	0.493
DNSMOS OVRL ↑	1.811

These values are reported as official server feedback for the final submitted waveforms. No ASR-, DNSMOS-, speaker-similarity-, or timing-metric-driven feedback from EVAL-1 or EVAL-2 was used for training, model selection, per-utterance candidate selection, or waveform refinement.

VI. Rule Compliance

The submitted system follows the REAL-TSE participation rules in the following respects.

- Training used only eligible open-source speech and noise data, together with the released baseline checkpoint.
- All training data sources, augmentation sources, and checkpoints are documented in this report.
- REAL-TSE DEV was used only for validation and model selection, not for model training or fine-tuning.
- EVAL-1 and EVAL-2 were used only to generate the final submitted waveforms.
- Official evaluation scores, gradients, embeddings, and other utterance-level feedback were not used to optimize, select, or refine individual evaluation utterances.
- No AliMeeting, AISHELL-4, AMI, DipCo, or CHiME6 audio was used for training, pretraining, fine-tuning, or augmentation.
- The separator adds no future-frame lookahead after STFT analysis; algorithmic latency is 24 ms, and the submitted total latency metadata is 32 ms including one 8 ms buffering hop.

VII. Summary

The AIVOX online-track system is a lightweight extension of the official causal TF-Map + Context BSRNN baseline. It retains the original target-speaker conditioning and complex-mask reconstruction pipeline while adding a causal temporal convolutional pre-filter before the temporal LSTM in each

BSNet block. The added branch is zero-initialized, causal, and parameter-efficient. For EVAL-1+EVAL-2, we report only the official server metrics returned for the submitted waveform package, without claiming a local evaluation-set baseline comparison. Its 24 ms algorithmic latency, 24.4 ms latency-verification result, and 32 ms submitted total-latency metadata are all within the 100 ms online-track budget.

REFERENCES

[1] REAL-TSE Challenge, "Real-world Target Speaker Extraction Challenge," 2026. [Online]. Available: https://real-tse.github.io/challenge/

[2] REAL-TSE, "REAL-TSE Challenge repository and official baseline results," 2026. [Online]. Available: https://github.com/REAL-TSE/REAL-TSE-Challenge

[3] S. Li, S. Wang, J. Han, K. Zhang, W. Wang, and H. Li, "REAL-T: Real Conversational Mixtures for Target Speaker Extraction," in *Proc. Interspeech*, 2025, pp. 1923–1927, doi: 10.21437/Interspeech.2025-2662.

[4] Y. Luo and J. Yu, "Music Source Separation with Band-Split RNN," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[5] K. Zhang, J. Li, S. Wang, Y. Wei, Y. Wang, Y. Wang, and H. Li, "Multi-Level Speaker Representation for Target Speaker Extraction," in *Proc. ICASSP*, 2025. [Online]. Available: https://arxiv.org/abs/2410.16059

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. ICASSP*, 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.

[7] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620, doi: 10.21437/Interspeech.2017-950.

[8] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090, doi: 10.21437/Interspeech.2018-1929.

[9] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," arXiv:1510.08484, 2015. [Online]. Available: https://www.openslr.org/17/

[10] 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.

[11] 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.