Real-world Target Speaker Extraction with Schrödinger Bridge

Technical Report

Hao Liang*
Waseda University
lianghao@akane.waseda.jp

Siyu Chen*
Waseda University
chensiyu@ruri.waseda.jp

Xu Shen*
Waseda University
shenxu@ruri.waseda.jp

I. System Overview

The submitted system is based on USEF-TFGridNet and adopts a Schrödinger Bridge formulation for target speaker extraction as shown in Fig. 1. The bridge process uses the mixture speech as the endpoint at $t = 0$ and the target speaker speech as the endpoint at $t = 1$. During training, the model predicts the target speaker speech from intermediate bridge states. During inference, the model starts directly from the mixture endpoint and performs target speaker extraction with a single forward pass.

II. Backbone: USEF-TFGridNet

The backbone is USEF-TFGridNet [1], which combines the speaker-embedding-free target speaker extraction framework with the TF-GridNet separator. In USEF-TSE, the enrollment speech is not converted into a fixed speaker embedding by an external speaker embedding extractor. Instead, frame-level enrollment features are used through cross-attention to provide target-speaker information to the separator [1]. TF-GridNet performs target speech estimation in the complex time-frequency domain and models spectro-temporal dependencies with stacked TF-GridNet blocks [2].

III. Schrödinger Bridge Formulation

Schrödinger Bridge describes a stochastic transport path between two endpoint distributions [3]–[5]. Let y denote the mixture speech and s denote the target speaker speech. The bridge endpoints are defined as $x_0 = y$ and $x_1 = s$, where $t = 0$ corresponds to the mixture speech and $t = 1$ corresponds to the target speaker speech.

For a sampled time $t \in [0, 1]$, the intermediate bridge state is constructed as $x_t = w_y(t)y + w_s(t)s + \sigma(t)z$, where $z \sim \mathcal{N}_{\mathbb{C}}(0, I)$. Here, $w_y(t)$, $w_s(t)$, and $\sigma(t)$ are determined by the Schrödinger Bridge noise schedule. This construction satisfies $x_t \to y$ as $t \to 0$ and $x_t \to s$ as $t \to 1$.

IV. Training and One-Step Inference

During training, a time step $t \in [0, 1]$ is randomly sampled, and the bridge state x_t is constructed from the mixture-target pair (y, s). The model takes x_t, the enrollment speech r, and t as inputs, and predicts the target speaker speech. The training objective is the negative SI-SDR loss, $\mathcal{L} = -\text{SI-SDR}(\hat{s}, s)$.

During inference, the clean target speech is unavailable. Since the bridge endpoint at $t = 0$ is defined as $x_0 = y$, the input state is set to the mixture speech and t is fixed to 0. Target speaker extraction is performed with a single forward pass conditioned on the enrollment speech.

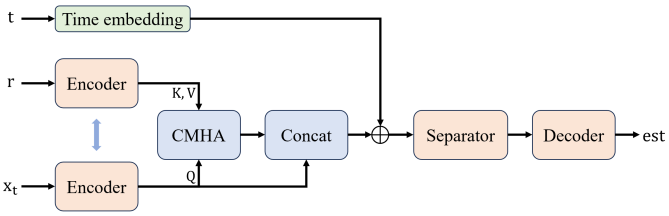

Fig. 1. Overview of the proposed structure.

V. Online Model

For the Online track, the causal model uses frame-wise RI-domain normalization instead of global waveform-level scale normalization, so the normalization statistics of each frame depend only on the current frame. The GroupNorm layer after the input convolution is replaced with channel layer normalization. In each TF-GridNet block, the inter_rnn is changed from bidirectional LSTM to unidirectional LSTM. Temporal self-attention is restricted by a limited-lookahead mask, where each query frame attends only to past frames and at most one future frame.

The configured algorithmic latency is 64 ms, consisting of 8 ms STFT latency, 8 ms causal convolution latency, and 48 ms look-ahead from six TF-GridNet blocks with one future frame per block. The latency verification result is 70 ms, which satisfies the 100 ms Online Track requirement.

VI. Training Dataset

To better match the REAL-TSE Challenge setting [11], training data are generated online with simulated multi-speaker meeting and dinner-party scenarios. Speech sources are sampled from LibriSpeech train-clean-100 and train-clean-360 [6], EARS [7], and VCTK [8]. Noise signals are sampled from the indoor scenes of RealMAN [9] and DEMAND [10].

*These authors contributed equally to this work.

Room acoustics are simulated on the fly using FRAM-RIR [12]. Each room is rectangular, with length 4.0–12.0 m, width 3.2–8.0 m, height 2.4–3.6 m, and RT60 0.05–0.60 s. A round, square, or rectangular table is randomly placed in the room. Speakers are arranged around the table, and the microphone is placed near the table center. The early RIR is defined by the first 50 ms of the room response.

Each sample contains a 6-second mixture, a target speaker signal, and a noisy enrollment utterance. The mixture contains 2–5 speakers. The target speaker is randomly selected from the mixture speakers, and enrollment is sampled from another utterance of the same speaker with duration 5–15 s. Two mixture-target branches are sampled with equal probability. In the first branch, both mixture sources and target use early RIRs. In the second branch, the mixture uses full reverberant RIRs and the target uses early RIRs. Enrollment is generated in an independently sampled room and can be either reverberant or early. Both mixture and enrollment are corrupted by additive indoor noise with SNR uniformly sampled from 5 to 30 dB, and their levels are normalized to -20 to -10 dBFS.

VII. Validation Strategy and Rule Compliance

The internal validation set is pre-generated by the online data generation pipeline and contains 1000 simulated samples. The model is trained continuously without a fixed epoch schedule. At regular intervals, the current top-10 checkpoints are selected according to the validation loss on the internal validation set and averaged to obtain a candidate averaged model. Candidate models are evaluated on the REAL-TSE development set, and the final submission model is selected according to the overall performance on TER, Speaker Similarity, Timing F1, and DNSMOS OVRL. The selected model is then used to generate the evaluation-set submission.

No additional post-processing is applied. The submitted waveforms are directly generated by the selected fixed model with one forward pass. For the evaluation set, the official evaluation models, scores, gradients, embeddings, or other feedback are not used to optimize, tune, select candidates, or post-process individual utterances.

VIII. Results

Table I reports the results on the REAL-TSE development set. The official non-causal baselines are included for comparison. Tables II and III show the evaluation results on the Online and Offline tracks, respectively. On the evaluation set, our team WasedaM ranks 6th in the Online track and 4th in the Offline track.

TABLE I
Results on the REAL-TSE development set.

Model	TER↓	F1↑	SIM↑	OVRL↑
BSRNN_EMB	0.693	0.841	0.501	1.660
BSRNN_TFMAP	0.703	0.838	0.521	1.500
ours	0.442	0.887	0.512	2.259

TABLE II
Evaluation results on the Online track.

Rank	Team	TER↓	F1↑	SIM↑	OVRL↑
1	CARTSE	0.699	0.848	0.504	3.436
2	DeepSound	0.713	0.849	0.524	2.151
3	SonicAGI	0.719	0.847	0.480	2.399
6	**WasedaM**	0.719	0.855	0.446	2.206

TABLE III
Evaluation results on the Offline track.

Rank	Team	TER↓	F1↑	SIM↑	OVRL↑
1	MERL-SA	0.593	0.866	0.989	4.060
2	YiJiaHe	0.639	0.871	0.565	3.460
3	CARTSE	0.651	0.857	0.544	3.364
4	**WasedaM**	0.675	0.858	0.480	2.324

References

[1] B. Zeng and M. Li, "USEF-TSE: Universal speaker embedding free target speaker extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 2110–2124, 2025.

[2] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[3] A. Jukic, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," in *Proc. Interspeech*, 2024, pp. 1175–1179.

[4] J. Yang, S. Wang, C. Wu, L. Guo, and F. Fan, "Schrödinger Bridge Mamba for one-step speech enhancement," arXiv preprint arXiv:2510.16834, 2026.

[5] J. Richter, D. de Oliveira, and T. Gerkmann, "Investigating training objectives for generative speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[7] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proc. Interspeech*, 2024, pp. 4873–4877.

[8] C. Veaux, J. Yamagishi, and S. King, "The Voice Bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[9] B. Yang, C. Quan, Y. Wang, P. Wang, Y. Yang, Y. Fang, N. Shao, H. Bu, X. Xu, and X. Li, "RealMAN: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

[10] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *Proc. Meetings on Acoustics*, vol. 19, no. 1, 2013, Art. no. 035081.

[11] S. Wang, Z. Qian, K. Zhang, J. Han, Z. Liu, H. Li, M. Delcroix, K. Yu, L. Xie, M. Li, and H. Li, "REAL-TSE Challenge: Real-world target speaker extraction from conversational recordings," Technical report, 2026.

[12] Y. Luo and R. Gu, "Fast random approximation of multi-channel room impulse response," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2024, pp. 449–453.