The SonicAGI System for the REAL-TSE Challenge

Kai Li[1,2,†], Wendi Sang[1,†], Jintao Cheng[1], Xiaolin Hu[1,2,3,*]
[1]Department of Computer Science and Technology, Institute for AI, BNRist, Tsinghua University, Beijing, China
[2]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China
[3]Chinese Institute for Brain Research (CIBR), Beijing, China
[†]These authors contributed equally to this work. [*]Corresponding author.

***Abstract*—Real-world target speaker extraction (TSE) remains challenging because target speech, interference, and enrollment are recorded under mismatched acoustic conditions with reverberation, noise, and irregular conversational overlap. This paper describes the SonicAGI submission to the REAL-TSE Challenge (IEEE SLT 2026). We take a data-centric approach that combines fully simulated mixtures from clean speech with real meeting overlaps, and use a frozen offline enhancer to provide a denoised mirror of real targets for auxiliary supervision. For the online track, we introduce SwiftNet-Lookahead, which inserts a single bounded-lookahead module before a strictly causal iterative separator and keeps the total system latency at 96 ms. For the offline track, we use a frame-level enrollment cross-attention USEF-TFGridNet with a magnitude-domain fusion stage that trades off perceptual quality and speaker fidelity. In the official evaluation, SwiftNet-Lookahead ranks third in Track 1 and USEF-TFGridNet ranks fifth in Track 2, both exceeding the challenge baselines. These results suggest that real-data-oriented training and track-specific modeling are effective for conversational TSE.**

***Index Terms*—target speaker extraction, speech separation, low latency, causal modeling, data simulation, conversational speech**

I. Introduction

Target speaker extraction (TSE) extracts a designated speaker from a mixture using an enrollment utterance [1]. It is a key front-end for hearing aids, meeting transcription, and speech interaction. Although recent speech and query-based separation methods perform well on curated or synthetic mixtures [2]–[4], real conversational recordings contain reverberation, background noise, irregular overlap, and cross-scene enrollment. These factors expose a persistent mismatch between synthetic training and real evaluation conditions, making data realism a central bottleneck.

The REAL-TSE Challenge [5] evaluates this setting on real Mandarin and English meeting and dinner-party recordings, scored by transcription error, target-speaker timing, speaker similarity, and perceptual quality. Track 1 requires online processing below 100 ms algorithmic latency and verifies causality through input-perturbation tests, whereas Track 2 allows full-context offline processing. Development data are held-out partitions of AISHELL-4 [6], AliMeeting [7], AMI [8], DipCo [9], and CHiME-6 [10]; their development and test splits are excluded from training.

Our SonicAGI submission is built around three components. First, we construct training data with a two-route sampler: *Fully Simulated Mixing* provides controlled speaker overlap from clean speech, while *Real-recording Mixing* samples overlapping utterances from real meetings; a frozen offline enhancer further produces a cleaned mirror of real targets for auxiliary supervision. Second, for Track 1, we propose *SwiftNet-Lookahead*, which inserts a single chunked bidirectional LSTM lookahead module before a strictly causal SwiftNet separator, so that future context is available within the latency budget but does not accumulate across iterative updates. Third, for Track 2, we use a frame-level enrollment cross-attention USEF-TFGridNet and apply magnitude-domain fusion between the extractor output and its enhanced version, improving perceptual quality while preserving target-speaker cues. The resulting systems rank third in Track 1 and fifth in Track 2, both outperforming the challenge baselines.

II. Data Construction

All audio is 16 kHz mono, and training samples are generated on the fly. Each sample takes, with equal probability, one of two routes: a *Fully Simulated Mixing* route that builds the mixture from clean speech under simulated acoustic conditions, and a *Real-recording Mixing* route that remixes target and interferer segments cut from real meetings. The two routes share the enrollment-length sampling and the global-gain setting; enrollment lengths are drawn from $[3, 15]$ s (covering 94% of the development-set durations) and the training chunk is 3 s. Table I lists the main simulation parameters.

TABLE I
Key data-simulation parameters (shared by both routes unless otherwise noted).

Parameter	Value
Sample rate	16 kHz, mono
Fully Simulated / Real-recording Mixing prob.	0.5 / 0.5
Training chunk length	3.0 s
Enrollment length	$[3, 15]$ s
Number of speakers N	2–4
Signal-to-interference ratio	$[-5, 10]$ dB
Noise probability (synthetic)	0.6
Noise SNR (synthetic)	$[-5, 25]$ dB
Reverberation RT60 (synthetic)	$[0.1, 0.7]$ s
Global gain	$[-12, 0]$ dB
Cross-scene enrollment prob. (real)	0.5
Samples per epoch	20,000

A. Source corpora

The Fully Simulated Mixing route draws clean single-speaker speech from LRS3 [11] and VoxCeleb2 [12], transcoded from mp4 to PCM WAV. The Real-recording Mixing route draws overlapping speech from the official *training* splits of AISHELL-4 [6], AliMeeting [7], AMI [8] (single distant microphone), and CHiME-6 [10]. Following the challenge rules, we exclude the development and test sets of every corpus, and never use DipCo for training. WHAM! [13] noise serves as the additive background in the Fully Simulated Mixing route.

B. Fully Simulated Mixing route

A Fully Simulated Mixing sample first selects a target speaker and an anchor utterance, then draws $N \in [2, 4]$ speakers in total (of which $N - 1$ interferers come from the full speaker pool), and reverberates each source with an independent single-channel room impulse response generated by a fast random-image simulator [14] (RT60 in $[0.1, 0.7]$ s). Interferers are added at a signal-to-interference

ratio in $[-5, 10]$ dB; the mixture is peak-normalized jointly with the sources, an optional global gain in $[-12, 0]$ dB is applied, and finally WHAM! noise is added with probability 0.6 at $[-5, 25]$ dB. The enrollment uses *another* utterance of the target speaker, and is degraded *independently* of the mixture with its own room impulse response (probability 0.5) and additive noise (probability 0.5, $[0, 20]$ dB), so that the model cannot use a shared acoustic channel as a shortcut cue for identifying the target.

C. Real-recording Mixing route

A Real-recording Mixing sample is drawn entirely from one meeting: the target and one or more interferers are real utterances from the same recording, so the mixture keeps genuine acoustics and we add no artificial reverberation or noise [15]. Interferers are remixed at a signal-to-interference ratio in $[-5, 10]$ dB, followed by the same peak normalization and optional global gain. With probability 0.5 the enrollment prefers a *cross-scene* reference (the same speaker in another meeting), falling back to a same-scene one otherwise; this suppresses the model's shortcut dependence on the acoustic environment and matches the cross-meeting enrollment of the evaluation set.

Real meeting segments are never perfectly clean: the "target" still contains background noise and crosstalk. To use these segments without forcing the network to reproduce that noise, we precompute a cleaned mirror of the real corpora with a frozen offline enhancer (RE-USE [16]). The mixture is always built from the *noisy* (acoustically real) target, while the enrollment is loaded from the *cleaned* mirror. For the offline USEF-TFGridNet, the two views form a combined supervision target (noisy as the primary term, cleaned as a light auxiliary) that keeps the acoustic mapping faithful to the real target while improving perceptual quality; weights are given in Section IV. The online SwiftNet-Lookahead, for latency and stability, is supervised only with the noisy real target.

III. Systems

Both tracks use the data pipeline above, but differ in separator architecture and enrollment conditioning.

A. Online track: SwiftNet-Lookahead

For Track 1, *SwiftNet-Lookahead* introduces bounded future context into a causal TSE model. Its backbone is a weight-shared SwiftNet [17] separator with twelve iterative updates, each combining multi-scale causal convolutions, frequency-axis bidirectional modeling, time-axis unidirectional modeling, and causally masked self-attention. A frozen ResNet-34 speaker encoder [18] from WeSpeaker [19] provides the target representation, which is injected by feature-wise linear modulation (FiLM) [20] only at the first iteration; subsequent updates refine the same conditioned representation. The learnable STFT front end uses a 256-point window and a 128-sample hop, and adds no latency beyond the analysis-synthesis delay.

A naive lookahead inside each shared update would accumulate: n future frames per update would become $12n$ frames after twelve iterations. We instead insert a single *LookaheadModule* between the audio bottleneck and the causal separator. It partitions time into non-overlapping chunks of length S and applies a bidirectional LSTM independently within each chunk; the output is projected, layer-normalized, and added as a residual. Because no recurrent state crosses chunk boundaries, the maximum future context is

$$\text{lookahead}_{\max} = (S - 1)\, H, \tag{1}$$

where H is the STFT hop. The separator remains strictly causal, so no additional lookahead accumulates. We use a two-layer bidirectional LSTM with 128 hidden units per direction on the 256-channel bottleneck. With $S = 11$ and $H = 128$ samples (8 ms/frame), the maximum lookahead is 80 ms.

Under the challenge latency definition, the algorithmic latency is the 8 ms STFT delay plus the 80 ms lookahead, yielding 88 ms; including the 8 ms hop buffer gives 96 ms total system latency. The streaming schedule emits one hop after the corresponding bounded-lookahead chunk is available, and no recurrent state carries future information across chunks. A tail-perturbation probe reported effective lookaheads of 37.9, 75.0, and 22.6 ms at three positions, all below the structural bound in Eq. (1); these spot checks supplement, but do not replace, the architectural latency bound.

B. Offline track: USEF-TFGridNet

For Track 2, we use full-context modeling with a USEF-style TFGridNet extractor [21], [22]. Rather than compressing the enrollment into a fixed-dimensional speaker embedding, the model keeps frame-level enrollment features and conditions the mixture through per-frequency cross-attention, with mixture frames as queries and effective enrollment frames as keys and values. Since padded enrollment frames are excluded from the keys and values, conditioning is invariant to in-batch padding. The target-aware representation is concatenated with the mixture representation, processed by six GridNetV2 blocks, and decoded by a transposed convolution followed by an inverse STFT. The model uses a 256-point window, a 128-sample hop, a shared encoder with $C = 128$ channels, four attention heads, and GridNetV2 blocks with hidden dimension 256.

After extraction, we apply a validation-tuned magnitude-domain fusion to balance perceptual quality and speaker fidelity. Let $\mathbf{x}_r, \mathbf{x}_e \in \mathbb{R}^N$ denote the original extractor output and the output after a frozen RE-USE [16] enhancer for the same sample. The enhanced output suppresses noise but can attenuate the target, whereas the original output preserves target-speaker cues but retains residual noise. We therefore align RMS, fuse STFT magnitudes, and reuse the phase of $\mathbf{x}_r$.

Specifically, we first align the RMS of the enhanced output to that of the original output:

$$\tilde{\mathbf{x}}_e = g\, \mathbf{x}_e, \quad g = \text{clip}\left(\frac{\text{rms}(\mathbf{x}_r)}{\text{rms}(\mathbf{x}_e)},\, 0.1,\, 10\right). \tag{2}$$

Let $\mathbf{X}_r = \text{STFT}(\mathbf{x}_r)$ and $\mathbf{X}_e = \text{STFT}(\tilde{\mathbf{x}}_e)$, with magnitudes $\mathbf{M}_r = |\mathbf{X}_r|$ and $\mathbf{M}_e = |\mathbf{X}_e|$. We estimate frame-level speech activity from the enhanced magnitude:

$$e_t = \sqrt{\frac{1}{F}\sum_f \mathbf{M}_e(f,t)^2}, \quad a_t = \sigma\left(\frac{20\log_{10}\frac{e_t+\epsilon}{p_{20}+\epsilon} - \tau}{\Delta}\right), \tag{3}$$

where p_{20} is the 20th-percentile energy of $\{e_t\}$, $\tau = 10$ dB is the activity threshold above this floor, $\Delta = 6$ dB controls the sigmoid transition width, and ϵ is a small constant. We smooth a_t with a 5-frame moving average and compute the fused magnitude as

$$w_t = w - \rho a_t, \qquad \mathbf{M}_f(f,t) = w_t\, \mathbf{M}_e(f,t) + (1 - w_t)\mathbf{M}_r(f,t). \tag{4}$$

Here w is the background-frame enhancement weight and ρ reduces this weight on speech-active frames. We use validation-tuned values $w = 0.9$ and $\rho = 0.35$, so active frames retain more of the original TSE magnitude ($w - \rho = 0.55$) while background frames remain closer to the enhanced magnitude. The final waveform is reconstructed as

$$\hat{\mathbf{x}} = \text{iSTFT}(\mathbf{M}_f \exp(j\angle\mathbf{X}_r)). \tag{5}$$

IV. Training Setup

Both systems were trained with the same protocol: gradient clipping at norm 5.0 and a plateau learning-rate scheduler (decay factor 0.5) driven by the scale-invariant SNR on a validation set of held-out meetings with scenes disjoint from training. Checkpoint selection and early stopping both followed this validation metric, avoiding any tuning with leaderboard feedback.

The track-specific configurations differed in supervision and optimizer. SwiftNet-Lookahead minimized a plain SNR loss against the target speech, with its ResNet-34 speaker encoder frozen, and was trained on six GPUs with AdamW (initial learning rate 10^{-3}, weight decay 0.1). USEF-TFGridNet used a combined SNR loss weighting the noisy real target by 0.8 and the cleaned mirror by 0.2, and was optimized by Adam [23] with an initial learning rate of 5×10^{-4}. All external corpora and frozen checkpoints used by the submission are named in Sections II–III; no development or evaluation split is used for training, fine-tuning, or data augmentation.

V. Results

Evaluation protocol. The evaluation set comprises EVAL-1 (2000 "seen" mixture–enrollment pairs drawn without overlap from the development corpora) and EVAL-2 (3000 "unseen" real-world samples newly recorded across meeting rooms, cafés, homes, and vehicles with near-field, far-field, and mobile devices). Each pair was processed independently with all metadata withheld. The leaderboard scores four metrics: target-speaker error rate (TER; lower is better), presence-timing F1, speaker similarity (SIM, cosine), and DNSMOS [24] perceptual quality. The official rank is the average dense rank over these four metrics on the combined EVAL-1+EVAL-2 target; the public leaderboard[1] listed 27 Track 1 participants and 30 Track 2 participants at our final snapshot [5].

Online track. Table II reports Track 1. On the public leaderboard snapshot, SwiftNet-Lookahead ranks third overall among valid participants. The table compares it with the two official BSRNN baselines [25]; SwiftNet-Lookahead leads both on all four metrics. The ablation in Table III explains two design choices. For training data, the mixed setting attains the best TER, Timing F1, and SIM; synthetic-only data degrades the most (TER 0.857, DNSMOS 1.676), suggesting that real acoustics are important for generalization, while real-only data reaches a marginally higher DNSMOS but trails on intelligibility and similarity, being limited in scale and scene diversity. For the enrollment encoder, ResNet-34 gives the best TER and the best DNSMOS; TF-MAP/contextual embedding [26] matches it on Timing F1 and SIM but drops sharply to 1.923 DNSMOS, and ECAPA-TDNN [27] is comparable on TER yet weaker elsewhere. We therefore adopt mixed training data with the ResNet-34 conditioning representation.

TABLE II
Online track (Track 1).

System	TER↓	Timing F1↑	SIM↑	DNSMOS↑
SwiftNet-Lookahead (ours)	**0.719**	**0.847**	**0.480**	**2.399**
BSRNN-TFMAP	0.805	0.836	0.456	1.670
BSRNN-EMB	0.809	0.821	0.422	1.714

Offline track. Table IV reports Track 2. USEF-TFGridNet ranks fifth on the public snapshot and beats both baselines by a large margin (TER 0.680, DNSMOS 2.484). Its remaining speaker-similarity gap is consistent with the perception-distortion trade-off from fusing toward denoised magnitudes on background frames [28]. Table V shows that frame-level enrollment cross-attention gives the best TER and DNSMOS, whereas compressed embeddings have similar Timing F1/SIM but worse perceptual quality. Within the supervision-weight block, 0.8:0.2 is best on all metrics; a small cleaned-mirror weight balances fidelity to real acoustics against residual-noise suppression.

TABLE III
Ablation study of SwiftNet-Lookahead.

Configuration	TER↓	Timing F1↑	SIM↑	DNSMOS↑
Training-data composition				
Mixed synthetic + real (default)	**0.719**	**0.847**	**0.480**	2.399
Synthetic data only	0.857	0.839	0.454	1.676
Real data only	0.739	0.841	0.450	**2.477**
Enrollment audio encoder				
ResNet-34 (default)	**0.719**	0.847	0.480	**2.399**
ECAPA-TDNN	0.723	0.828	0.464	2.392
TF-MAP/contextual embedding	0.757	**0.848**	**0.483**	1.923

TABLE IV
Offline track (Track 2).

System	TER↓	Timing F1↑	SIM↑	DNSMOS↑
USEF-TFGridNet (ours)	**0.680**	**0.851**	**0.471**	**2.484**
BSRNN-EMB (baseline)	0.829	0.829	0.417	1.844
BSRNN-TFMAP (baseline)	0.838	0.829	0.443	1.612

TABLE V
Ablation study of USEF-TFGridNet.

Configuration	TER↓	Timing F1↑	SIM↑	DNSMOS↑
Enrollment audio encoder				
USEF (default)	**0.680**	0.851	0.471	**2.484**
ResNet-34	0.739	**0.853**	0.474	1.903
ECAPA-TDNN	0.728	0.850	**0.476**	2.248
TF-MAP/contextual embedding	0.781	0.845	0.470	2.138
Combined supervision weight (noisy real target:cleaned mirror target)				
0 : 1	0.756	0.844	0.454	2.271
0.5 : 0.5	0.739	0.841	0.450	2.477
0.8 : 0.2 (default)	**0.680**	**0.851**	**0.471**	**2.484**
1 : 0	0.684	0.840	0.452	2.249

VI. Conclusion

This paper presents the SonicAGI systems for the online and offline tracks of the REAL-TSE Challenge. The results suggest that real conversational TSE depends not only on separator design, but also on training-data realism and target construction. Our two-route data pipeline combines controllable synthetic mixtures with real meeting overlap, and uses a cleaned mirror target to reduce the effect of residual background noise in real recordings. SwiftNet-Lookahead confines future context to a single bounded injection, achieving third place in Track 1 with 96 ms total system latency. USEF-TFGridNet combines frame-level enrollment cross-attention with magnitude-domain fusion, ranking fifth in Track 2 while improving intelligibility and perceptual quality. The main remaining limitation is the trade-off between speaker similarity and perceptual quality; future work will explore similarity-aware training or fusion objectives and extend bounded lookahead to stronger cross-attention extractors.

[1] https://leaderboard.real-tse.com/, accessed 2026-06-30.

AI-Generated Content Disclosure

ChatGPT was used to assist with language editing, grammar refinement, and concision of the manuscript text. The authors reviewed and revised all AI-assisted edits.

References

[1] S. Wang, Z. Chen, K. A. Lee, Y. Qian, and H. Li, "Overview of speaker modeling and its applications: From the lens of deep speaker representation learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4971–4998, 2024.

[2] K. Li, G. Chen, W. Sang, Y. Luo, Z. Chen, S. Wang, S. He, Z.-Q. Wang, A. Li, Z. Wu *et al.*, "Advances in speech separation: Techniques, challenges, and future trends," *arXiv preprint arXiv:2508.10830*, 2025.

[3] K. Li, K. Gao, and X. Hu, "Efficient audio-visual speech separation with discrete lip semantics and multi-scale global-local attention," *arXiv preprint arXiv:2509.23610*, 2025.

[4] K. Li, J. Cheng, C. Zeng, Z. Yan, H. Wang, Z. Su, B. Zheng, and X. Hu, "A semantically consistent dataset for data-efficient query-based universal sound separation," *arXiv preprint arXiv:2601.22599*, 2026.

[5] REAL-TSE Challenge Organizers, "REAL-TSE Challenge: Real-world Target Speaker Extraction Challenge," https://real-tse.github.io/challenge/, 2026, a satellite challenge of IEEE Spoken Language Technology Workshop (SLT) 2026. Accessed: 2026-06-29.

[6] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 3665–3669.

[7] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6167–6171.

[8] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*. Springer Berlin Heidelberg, 2006, pp. 28–39.

[9] M. Van Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "DiPCo – dinner party corpus," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 434–438.

[10] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[11] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[12] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 1086–1090.

[13] 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 *Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 1368–1372.

[14] Y. Luo and J. Yu, "FRA-RIR: Fast random approximation of the image-source method," *arXiv preprint arXiv:2208.04101*, 2022.

[15] K. Li, W. Sang, C. Zeng, R. Yang, G. Chen, and X. Hu, "Sonicsim: A customizable simulation platform for speech processing in moving sound source scenarios," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 67 379–67 405.

[16] S.-W. Fu, R. Chao, X. Yang, S.-F. Huang, R. E. Zezario, R. Nasretdinov, A. Jukić, Y. Tsao, and Y.-C. F. Wang, "Rethinking training targets, architectures and data quality for universal speech enhancement," *arXiv preprint arXiv:2603.02641*, 2026.

[17] W. Sang, K. Li, R. Yang, J. Huang, and X. Hu, "A fast and lightweight model for causal audio-visual speech separation," in *European Conference on Artificial Intelligence (ECAI)*, 2025.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

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

[20] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

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

[22] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[24] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[25] 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.

[26] K. Zhang, J. Li, S. Wang, Y. Wei, Y. Wang, Y. Wang, and H. Li, "Multi-level speaker representation for target speaker extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 3830–3834.

[28] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6228–6237.