SLT 2026 REAL-TSE Challenge: Real-World Target Speaker Extraction from Conversational Recordings

Shuai Wang[1], Zihan Qian[1], Ke Zhang[2], Jiangyu Han[3], Zikai Liu[4], Xiaoyang Yu[1], Haoyu Li[1]
Marc Delcroix[5], Kai Yu[6], Lei Xie[4], Ming Li[2], and Haizhou Li[2]
[1]Nanjing University, China, [2]Chinese University of Hong Kong (Shenzhen), China
[3]Brno University of Technology, Czechia, [4]Northwestern Polytechnical University, China
[5]NTT, Inc., Japan, [6]Shanghai Jiao Tong University, China
realtse.challenge@gmail.com

arXiv:2607.15198v2 [eess.AS] 26 Oct 2026

***Abstract*—We introduce the REAL-TSE Challenge[1], an IEEE SLT 2026 satellite challenge on target speaker extraction (TSE) from real conversational recordings. Given a multi-speaker mixture and one or more enrollment utterances from a target speaker, participating systems must recover only the target speech. Unlike simulated read-speech benchmarks, REAL-TSE evaluates Mandarin and English recordings that contain natural overlap, reverberation, noise, channel mismatch, and conversational dynamics. The challenge defines two complementary tracks: an *Online* track for low-latency streaming extraction and an *Offline* track for full-context processing. Systems are evaluated with Token Error Rate (TER), Speaker Similarity (SpkSim), DNSMOS, and target-speaker activity F1. This overview covers the task, data, baselines, evaluation protocol, submitted systems, condition-wise findings, and lessons for future real-world TSE.**

I. Introduction

Target Speaker Extraction (TSE) aims to isolate a specified speaker from a multi-talker recording using auxiliary cues such as enrollment utterances [1]–[3]. Robust TSE is an enabling technology for teleconferencing, hearing assistance, smart assistants, and speaker-aware meeting transcription.

Despite substantial progress, most widely used TSE evaluations still rely on simulated mixtures such as LibriMix [4] and WSJ0-2Mix [5], where clean read-speech utterances are summed at predefined signal-to-noise ratios. These benchmarks are reproducible and analytically convenient, but they do not fully capture the conditions that make conversational extraction difficult: uncontrolled loudness variation, real reverberation, ambient noise, turn-taking, reactive overlap, disfluencies, and non-verbal vocalizations. Existing challenges have advanced far-field ASR and speech enhancement [6]–[10] in realistic acoustic environments, yet none directly benchmarks real-recorded TSE in both low-latency and full-context settings.

The **REAL-TSE Challenge** addresses this gap with four design choices. First, it evaluates real Mandarin and English conversational recordings, combining data derived from REAL-T [11] with an independently collected unseen subset. Second, it separates online and offline tracks so that streaming and batch systems can be assessed under appropriate constraints. Third, it uses a multi-dimensional scoring protocol that covers intelligibility, speaker consistency, perceptual quality, and target-speaker activity. Fourth, it releases baselines and validation tools built on WeSep [12][2], together with an official toolkit[3] for submission checking and metric computation. The challenge attracted registrations from more than 100 teams spanning over 15 countries and regions. Ultimately, 18 teams submitted 24 valid systems, with 12 submissions in each track. Their results are available on the Leaderboard.[4] Following the challenge, we further refined the dataset and made it publicly available.[5]

II. Related Datasets and Challenges

Simulated TSE sets such as WSJ0-2Mix [5] and LibriMix [4] provide clean references but omit natural conversational behavior. Meeting and far-field resources such as CHiME [6], [7], AMI [13], AISHELL-4 [10], and M2MeT [9] capture realistic acoustics, but primarily target ASR, diarization, or enhancement rather than enrollment-conditioned TSE.

REAL-T [11] is the closest prior real-conversation TSE resource. REAL-TSE extends it with a newly recorded bilingual unseen set, controlled device pairings, online/offline tracks, four-metric ranking, and post-challenge diagnostics for reproducible comparison under real conversational conditions.

III. Challenge Settings

Each REAL-TSE trial consists of a real multi-speaker mixture $\mathbf{x}$ and an enrollment utterance $\mathbf{e}$ from the target speaker. The system outputs an estimate $\hat{\mathbf{s}}$ of the target speech while suppressing non-target speakers and non-speech artifacts.

A. Track 1: Online Target Speaker Extraction

The online track targets latency-sensitive applications such as hearing devices and real-time communication. Systems must operate frame-wise or chunk-wise without full-utterance buffering, and their end-to-end algorithmic latency must not

[1]https://real-tse.github.io/challenge/

[2]https://github.com/REAL-TSE/wesep-real-tse

[3]https://github.com/REAL-TSE/REAL-TSE-Challenge

[4]https://real-tse.github.io/challenge/#rankings

[5]https://huggingface.co/datasets/REAL-TSE/REAL-T

exceed 100 ms. The latency budget includes future context introduced by STFT/iSTFT, overlap-add, look-ahead frames, smoothing, post-processing, and chunk accumulation, but excludes hardware computation and I/O time. Effective latency is verified with a perturbation-based response-delay test.

B. Track 2: Offline Target Speaker Extraction

The offline track targets batch scenarios such as meeting transcription and audio post-production. Systems may use the full utterance, global context, arbitrary model classes, and unrestricted inference cost. However, each mixture–enrollment pair must still be processed independently, without metadata or information from other evaluation samples.

C. Evaluation Metrics

Systems are ranked by dense ranking on four metrics, and the final rank is the average of the four metric ranks. As discussed in Sections VI and IX, this multi-objective design improves coverage of TSE quality but can also encourage metric gaming.

a) Intelligibility: Token Error Rate (TER): TER measures ASR transcription accuracy, using word error rate for English and character error rate for Mandarin. Lower TER indicates better preservation of the target speaker's linguistic content. The official ASR backbone is Zipformer [14].

b) Speaker Consistency: Speaker Similarity (SpkSim): SpkSim is the cosine similarity between speaker embeddings of $\hat{\mathbf{s}}$ and $\mathbf{e}$, computed with the WeSpeaker ResNet-34 encoder [15]. Higher scores indicate better target-speaker consistency. Tables and figures abbreviate this metric as SIM.

c) Perceptual Quality: DNSMOS: DNSMOS is a non-intrusive perceptual-quality estimator reflecting artifacts, distortion, and residual noise. The challenge initially considered the DNSMOS OVRL [16] sub-score for perceptual-quality ranking. However, post-submission reliability analysis and subjective listening tests showed that OVRL suffered from metric-specific over-optimization in this competitive setting. We therefore **adopt DNSMOS-P808 [17] as the official perceptual-quality metric**, which appeared not to have been the target of over-optimization, and report OVRL only for reference. Details are provided in Section VIII.

d) Target-Speaker Activity: F1 Score: F1 measures whether speech is emitted when, and only when, the target speaker is active. It combines temporal precision and recall over activity regions detected with FireRedVAD [18].

e) Evaluation References: EVAL-2 transcripts and target-activity references are human-annotated from the speaker-specific headset channels described in Section IV-C. The close-talk channels resolve the target content and timing during overlap, while FireRedVAD is applied only to determine activity in system outputs. This design avoids relying on the far-field mixture to decide which speaker produced a token or was active during overlap.

IV. Datasets

Most TSE benchmarks are simulated from clean read speech [19], which simplifies evaluation but introduces a mismatch with real reverberation, loudness variation, noise, and conversational turn-taking. REAL-TSE therefore grounds evaluation in real conversational recordings where overlap, speaker behavior, and channel effects arise naturally.

TABLE I
STATISTICS OF THE CHALLENGE DATASET. SAMPLES ARE MIXTURE-ENROLLMENT PAIRS; MIXTURES AND ENROLLMENTS ARE UNIQUE WAV COUNTS; TARGET SPEAKERS ARE DISTINCT SPEAKERS.

	DEV	**EVAL-1**	**EVAL-2**
Samples	1,991	1,930	3,000
Mixtures	528	576	1,186
Enrollments	242	177	1,913
Target speakers	64	48	77
Mixture audio (h)	2.60	2.90	5.67
Enrollment audio (h)	0.67	0.54	4.02

TABLE II
MEAN VALUES OF THE MAIN CHARACTERISTICS OF EACH DATA SPLIT. THE NUMBERS IN PARENTHESES SHOW THE MIN AND MAX VALUES.

Quantity	DEV	EVAL-1	EVAL-2
Mix dur. (s)	17.26 (6.24, 29.91)	18.33 (5.82, 29.99)	17.25 (5.48, 29.91)
Enroll dur. (s)	9.65 (5.00, 50.51)	11.00 (5.01, 59.82)	9.10 (3.00, 42.98)
Overlap ratio	0.49 (0.18, 0.99)	0.48 (0.18, 0.98)	0.53 (0.18, 0.95)
Target ratio	0.75 (0.27, 1.00)	0.75 (0.21, 1.00)	0.73 (0.31, 1.00)

A. Challenge Splits and Statistics

Table I summarizes the three challenge splits. Each sample is a released mixture–enrollment pair, and multiple samples may share mixture or enrollment files. In total, REAL-TSE contains 6,921 trials over 2,290 mixtures and 11.2 h of mixture audio in Mandarin and English.

Table II reports statistics for mixture duration, enrollment duration, overlap ratio, and target-speaker activity ratio.

B. Development Set

The development set contains 1,991 mixture–enrollment pairs from REAL-T [11], built from AISHELL-4 [10], AliMeeting [9], AMI [13], DipCo [20], and CHiME6 [7]. Each sample pairs a naturally overlapping mixture with a non-overlapping target-speaker enrollment segment of at least 5 seconds. The development set is intended only for tuning, comparison, and ablation, and must **not** be used for training or fine-tuning.

C. Evaluation Set

The evaluation set contains **4,930 mixture–enrollment pairs** in two subsets (Table I).

a) EVAL-1 (Seen, 1,930 pairs): EVAL-1[6] is drawn from the same source corpora as the development set but contains no overlapping samples, measuring in-domain performance.

b) EVAL-2 (Unseen, 3,000 pairs): EVAL-2 is newly collected and covers meetings, cafés, homes, and in-vehicle conversations. Each conversation was synchronously recorded with high-quality microphones (H1/H2), a phone, and one close-talk headset per speaker. Human annotators used the

[6]See Section XI for details on the evaluation corrections.

TABLE III
EVAL-2 MIXTURE×ENROLLMENT CHANNEL PAIRING (SAMPLE COUNTS). H1/H2: HIGH-QUALITY MICROPHONES; PHONE: MOBILE-PHONE RECORDING; HEADSET: SPEAKER-WORN CLOSE-TALK HEADSET; EXTERNAL: ADDITIONAL SINGLE-SPEAKER ENROLLMENT RECORDINGS.

Mix.\Enr.	Enrollment channel					Total
	H1	H2	phone	headset	external	
H1	**397**	127	129	141	218	1,012
H2	142	**415**	128	145	179	1,009
phone	122	135	**388**	131	203	979
Total	661	677	645	417	600	3,000

TABLE IV
BASELINE MODEL VARIANTS.

Model	Conditioning method	Extractor
`BSRNN_EMB`	ECAPA-TDNN emb.	Non-causal
`BSRNN_EMB_CAUSAL`	ECAPA-TDNN emb.	Causal
`BSRNN_TFMAP`	TF-Map + Context	Non-causal
`BSRNN_TFMAP_CAUSAL`	TF-Map + Context	Causal

clear per-speaker headset channels (which contain little cross-talk) to produce transcripts and target-activity labels, including during overlap. Target speakers also provided external single-speaker enrollment utterances. EVAL-2 tests generalization under unseen environments, devices, and channel mismatch.

Table III summarizes EVAL-2 channel pairings, including matched- and cross-channel cases used later for channel-mismatch analysis. For a fair blind evaluation, scenario labels, speaker identities, and device metadata are removed from the released samples, and each pair is processed independently.

D. Training Data Policy

REAL-TSE has no official training set. Participants may use open-source data and pretrained models, but they must document all data sources and checkpoints. Development/test splits of AliMeeting, AISHELL-4, AMI, DipCo, and CHiME6 are forbidden for pre-training, training, or augmentation, whereas their official training splits are allowed. This open-training policy encourages diverse data and modeling strategies while exposing the practical upper bound of current approaches.

V. BASELINE SYSTEMS

We provide four BSRNN-based baselines [21], [22] implemented with WESEP [12]. All baselines are trained on fully overlapped Libri2Mix-100 [4] at 16 kHz for 150 epochs. They compare two enrollment-conditioning methods under causal and non-causal extractors, as summarized in Table IV.

`BSRNN_EMB` uses a fixed speaker embedding from ECAPA-TDNN [15], [23], whereas `BSRNN_TFMAP` combines spectral TF-Map features with contextual embeddings [22].

The causal baselines keep the same conditioning modules but impose causality inside the BSRNN extractor. Specifically, they replace global normalization with feature-axis normalization and bidirectional temporal LSTMs with unidirectional LSTMs.

Table V reports the official baseline results. Trained only on synthetic English Libri2Mix, they are lower bounds: EVAL TER near or above 0.70 exposes the domain gap and motivates real-data adaptation. The gap also shows why these baselines provide a common reference rather than a competitive recipe.

TABLE V
BASELINE RESULTS ON DEV, EVAL-1, AND EVAL-2. LOWER TOKEN ERROR RATE (TER) IS BETTER; HIGHER SIM, DNSMOS P808/OVRL, AND F1 ARE BETTER. GRAY NUMBERS DENOTE THE PREVIOUSLY REPORTED DNSMOS OVRL SCORES.

Set	Model	TER	SIM	DNSMOS		F1
				P808	OVRL	
DEV	`BSRNN_EMB`	0.695	0.501	2.82	1.66	0.841
	`BSRNN_EMB_CAUSAL`	0.703	0.492	2.69	1.63	0.829
	`BSRNN_TFMAP`	0.703	0.521	2.72	1.50	0.838
	`BSRNN_TFMAP_CAUSAL`	0.651	0.535	2.76	1.55	0.844
EVAL-1	`BSRNN_EMB`	0.686	0.513	2.91	1.90	0.828
	`BSRNN_EMB_CAUSAL`	0.706	0.504	2.76	1.77	0.807
	`BSRNN_TFMAP`	0.712	0.540	2.80	1.73	0.823
	`BSRNN_TFMAP_CAUSAL`	0.667	0.558	2.84	1.77	0.836
EVAL-2	`BSRNN_EMB`	0.773	0.357	2.85	1.80	0.830
	`BSRNN_EMB_CAUSAL`	0.751	0.370	2.72	1.67	0.831
	`BSRNN_TFMAP`	0.766	0.382	2.73	1.52	0.833
	`BSRNN_TFMAP_CAUSAL`	0.742	0.391	2.75	1.59	0.835

VI. OVERVIEW OF SUBMITTED SYSTEMS

Across the two tracks, 18 distinct teams contributed 24 valid systems (12 per track). A valid submission complied with the challenge rules, produced valid evaluation outputs, and was accompanied by a system report. Rather than describing each system individually, we distill the recurring design choices that most consistently affected performance. Three of them stand out: (i) how real conversational data is exploited through simulation and adaptation; (ii) how the extractor is designed and conditioned under (or without) the latency budget; and (iii) how training objectives, post-processing, and test-time behavior interact with the automatic metrics.

Fig. 1 compares the top-five systems in each track with two baseline systems across the four official metrics. All metrics are normalized to a common scale, with TER reversed so that higher values consistently indicate better performance. Top systems substantially outperform both baselines across nearly all metrics. Nevertheless, **no leading system uniformly dominates across all evaluation dimensions**. Instead, the systems occupy distinct regions of the metric space: some perform better in speech intelligibility or target-speaker activity detection, whereas others more effectively preserve speaker identity or perceptual quality. These complementary strengths highlight the challenging multi-objective nature of real-world TSE, which must simultaneously preserve linguistic content, target activity, speaker identity, and perceptual naturalness.

A. Data Simulation and Real-Data Adaptation

Most teams synthesized mixtures from publicly available speech corpora such as LibriSpeech [24]/LibriMix [4], VoxCeleb [25], CN-Celeb [26], AISHELL [10], [27], VCTK [28], and EARS [29]; added noise from WHAM! [30], MUSAN [31], DEMAND [32], or DNS Challenge data [33]; and introduced reverberation with measured [34]/simulated RIRs [35], while varying the number of speakers, SNR/SIR, RT60, chunk length,

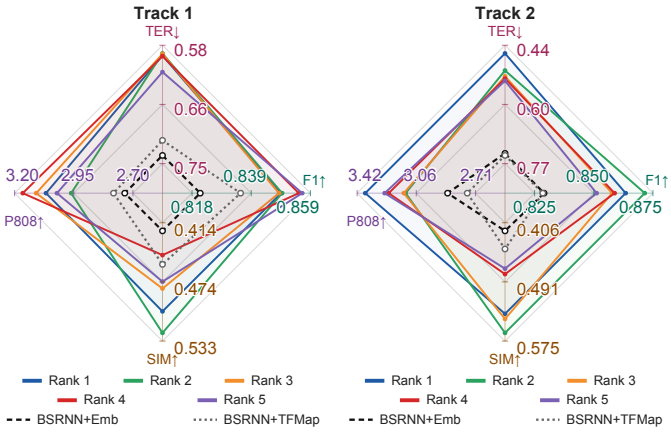


Fig. 1. Metric balance of the top-5 valid systems in each track.

and enrollment quality. Many teams also injected target-absent or noise-only segments, forcing systems not only to extract the target when present but also to suppress non-target speech and avoid false extraction.

The more consequential trend was **adaptation to real conversational data**. Since real-world mixtures rarely provide clean target references, competitive systems generated pseudo-labels using guided source separation, offline teacher models, model averaging, or self-training. The strongest approaches repeated this process over multiple rounds, gradually progressing from fully overlapped synthetic data to meeting-like mixtures and, ultimately, real far-field recordings. Across the reported results, real-data adaptation consistently outperformed synthetic-only training. Notably, the top three teams either incorporated real recordings into training or relied exclusively on real data.

B. Extractor Design and Speaker Conditioning

The latency constraint sharply narrowed the Track 1 design space: nearly all online systems were compact, causal discriminative extractors (mainly BSRNN and TF-GridNet [36] variants) adapted with causal normalization, unidirectional recurrence, and controlled look-ahead. Track 2 explored a much broader space, including SSL-based attractor initialization, diarization-aware front ends, and generative refinement based on diffusion, flow matching, or vocoder reconstruction. Both tracks explored where and how to inject enrollment information through global embeddings, frame-level features, TF-Map conditioning, prefix tokens, or speaker-aware state modulation.

Crucially, **the results suggest substantial headroom beyond the current submitted systems.** The top-performing entries were all based on BSRNN-style backbones similar to the official baseline, and some used nearly identical extractor architectures, yet achieved substantially better scores. This indicates that realistic data simulation, real-data adaptation, pseudo-label generation and filtering, loss design, metric-aware training, and post-processing can yield large gains even with an existing backbone. At the same time, these results should not be taken to imply that stronger backbones are unnecessary. Rather, they suggest that model architecture and training recipe have not yet been fully co-optimized for real-world TSE. With comparably strong data preparation and adaptation pipelines, more expressive backbones, such as TF-GridNet-style extractors, may therefore still offer considerable room for further improvement.

C. Objectives, Post-Processing, and Metric-Aware Behavior

Reconstruction losses (SI-SDR/SI-SNR, multi-resolution STFT) were routinely combined with auxiliary terms aligned to the official metrics, including speaker-similarity, DNSMOS-related, ASR/token-level, and VAD/activity losses, marking a clear shift from single-objective separation toward multi-objective optimization. Post-processing followed the same logic, particularly in Track 2: several teams appended enhancement or vocoder-based reconstruction modules, or produced raw and enhanced branches and selected between them using speaker-similarity criteria. Such processing typically improved DNSMOS but could reduce SpkSim or alter target activity, reflecting a trade-off rather than a uniform benefit.

This metric-aware behavior, however, exposed a fundamental weakness of the evaluation itself. Optimizing a neural metric can produce artifacts that raise the score without improving perceived quality. The system reports revealed borderline practices such as branch selection based on automatic scores and repeated use of the official validation scripts during development; in one extreme case, *adversarial waveform perturbations* were applied to inflate SpkSim and DNSMOS without genuinely improving extraction. That submission was ruled invalid and excluded from the official ranking and all valid-system analyses. These cases motivate stricter development rules that **prohibit use of final non-intrusive ranking metrics** (Section IX).

For Track 1, algorithmic latency mainly came from STFT/iSTFT delay, hop buffering, look-ahead, chunk accumulation, and other forms of non-causal context. Most submitted systems reported latency in the 20–100 ms range. The perturbation-based latency script was important for verifying these claims, since analytical estimates and measured delays did not always agree. In some cases, systems reported substantial look-ahead but showed relatively small measured delay; in others, latency was underestimated because buffering, normalization, or other implementation details were not fully accounted for. Future online TSE benchmarks should require both an analytical latency breakdown and standardized delay measurement.

VII. Condition-wise Analysis of EVAL-2 Results

Beyond system-report summaries, REAL-TSE supports aggregate analysis of how performance changes under real-world recording conditions. We focus on EVAL-2 because, unlike DEV and EVAL-1, it was newly recorded by the organizers to cover diverse scenarios, devices, and enrollment–mixture channel pairings. We analyze trends by device/channel pairing, target-speaker activity ratio, and scenario.

A. Device and Channel Mismatch Effects

We begin by examining channel effects between enrollment and mixture recordings. Figure 3 shows performance measured by the four metrics for different mixture/enrollment microphone

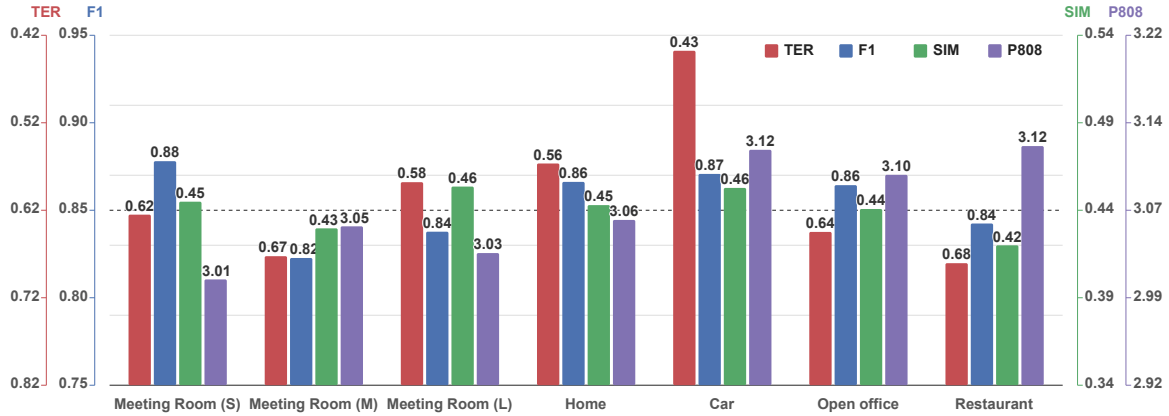


Fig. 2. Scenario-wise performance on the EVAL-2 set, computed by averaging all valid submissions separately for CN and EN and then taking a CN/EN macro average over shared scenarios.

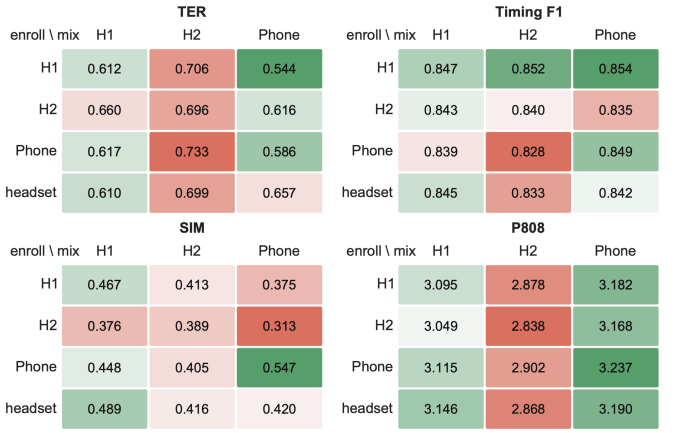


Fig. 3. Enrollment-device by mixture-device performance matrix on EVAL-2.

pairs, averaged over all valid submitted systems. As shown in Fig. 3, **the mixture recording device has a stronger effect than the enrollment device.** Mixtures recorded by the H2 microphone achieved consistently lower performance than H1 and Phone. This matches the recording setup, in which H2 is the farthest microphone and captures more reverberation and environmental interference. Robustness to far-field mixture channels thus remains a key open challenge for real-world TSE. Interestingly, mismatch has different effects across metrics, but systems are generally not over-sensitive except in speaker similarity, which uses enrollment as its reference. External enrollment is omitted from Fig. 3 because it has no corresponding mixture-device row.

B. Target Ratio and Utterance-level Activity

We further examine the relationship between utterance-level difficulty and the target ratio in Fig. 4, measured as the proportion of target-speaker activity in the mixture.

Overall, target ratio shows the clearest effect on target-activity F1. As the target speaker occupies a larger portion of the mixture, F1 increases, indicating more reliable detection and preservation of target-active regions. TER shows a weaker

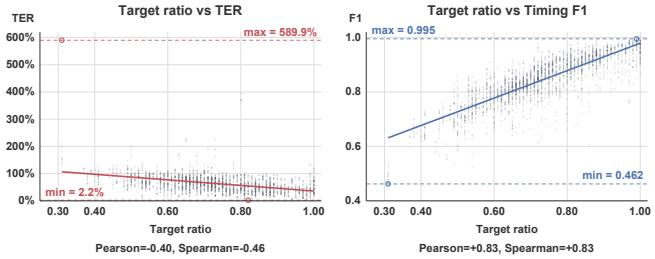


Fig. 4. Relationship between target ratio and F1/TER on EVAL-2.

decreasing trend, suggesting that target ratio alone does not fully explain utterance-level intelligibility errors.

C. Scenario and Conversational Difficulty

We next group EVAL-2 samples by recording scenario and compute average performance for each official metric. This tests whether labels for meeting rooms, restaurants, homes, and in-vehicle conversations reliably predict difficulty.

The observed trends do not consistently follow intuitive acoustic expectations. Meeting-room recordings might appear easier, and in-vehicle recordings seem manageable because the car interior is small and the recording geometry is relatively fixed. Conversely, restaurant-like scenarios might be expected to be more difficult because of stronger background noise and less controlled conversational behavior. However, Fig. 2 contradicts this intuitive ordering.

A likely reason is that EVAL-2 conversations were deliberately recorded in a natural and topic-driven manner. Speakers interacted freely rather than following a tightly controlled script, so each scenario contains substantial variability in turn-taking, overlap, speaker distance, loudness, device placement, and local noise events. Scenario-wise statistics are therefore a coarse diagnostic rather than a strict difficulty ranking.

VIII. Revisiting DNSMOS Reliability

Near the final stage of the challenge, we observed anomalous behavior in DNSMOS OVRL, the perceptual-quality score originally considered for ranking. Some submissions achieved

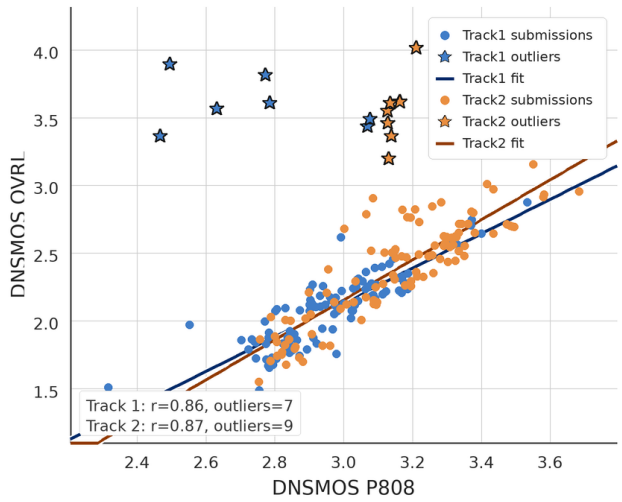


Fig. 5. Relationship between DNSMOS OVRL and DNSMOS P808 on participant submissions. The high-OVRL outliers do not show corresponding gains in P808, indicating metric-specific over-optimization rather than a consistent improvement in perceptual quality.

TABLE VI
CORRELATION BETWEEN HUMAN MOS AND DNSMOS VARIANTS FOR TOP-FIVE FINAL SUBMISSIONS.

		DNSMOS-OVRL		DNSMOS-P808		DNSMOS-PRO [39]	
Scope	n	LCC	SRCC	LCC	SRCC	LCC	SRCC
Track 1	500	+0.003	+0.040	+0.453	+0.380	+0.156	+0.115
Track 2	600	+0.182	+0.186	+0.466	+0.463	+0.229	+0.234
Overall	1100	+0.165	+0.186	+0.489	+0.460	+0.233	+0.230

unusually high OVRL scores that did not align with listening impressions or with other DNSMOS variants. We therefore conducted a post-submission reliability analysis using the submissions from each track.[7]

Fig. 5 shows the main symptom: several systems obtain high OVRL scores without comparable gains in P808. Since both scores are intended to estimate perceptual quality, such outliers suggest metric-specific over-optimization rather than uniform improvement in extracted-speech quality. To quantify this effect, Table VI compares DNSMOS variants against human MOS using linear correlation (LCC [37]) and rank correlation (SRCC [38]). OVRL is weakly correlated with human MOS, especially in Track 1, whereas P808 shows consistently higher agreement across both tracks. Based on this evidence, we use DNSMOS-P808 as the official perceptual-quality metric and keep OVRL only as a reference score.

These observations highlight a limitation of public reference-free neural metrics: training, post-processing, or selection may over-optimize or exploit them, so higher automatic scores need not indicate better perceived quality.

[7]The full analysis and official metric-update notice are available at https://real-tse.github.io/challenge/#evaluation. A similar issue was also spotted for speaker similarity, but it appeared to have less impact on evaluation. Due to space constraints, we thus focus our discussion on DNSMOS.

IX. Overall Lessons

The submissions suggest three lessons for future real-world TSE evaluations. First, most teams focused less on proposing fundamentally new extractor architectures and more on making existing backbones work under realistic conditions. Strong systems commonly combined BSRNN- or TF-GridNet-style extractors with careful data simulation, real-data adaptation, pseudo-label filtering, multi-objective training, post-processing, and latency control. This does not imply that architecture is unimportant; rather, backbone design and training recipes need to be co-optimized, and stronger models may still have substantial headroom with equally strong data preparation.

Second, REAL-TSE confirms that TSE quality is inherently multi-dimensional. No system consistently dominates TER, SpkSim, DNSMOS P808, and target-activity F1 at the same time. Different applications may therefore prefer different operating points: meeting transcription may emphasize intelligibility, hearing-assistance and communication scenarios may prioritize latency and perceptual quality, and speaker-aware applications may require stronger identity preservation and activity accuracy. Future leaderboards should report both aggregate rankings and use-case-specific views of performance.

Third, future benchmarks should keep a ranking signal hidden or human-audited, prohibit final-metric optimization or output selection, disclose metric-aware post-processing, and listen to cross-metric outliers. EVAL-2 also shows that fine-grained acoustic and interaction factors explain difficulty better than scenario labels.

X. Conclusions

REAL-TSE bridges controlled laboratory benchmarks and real conversational TSE through authentic bilingual recordings, online/offline tracks, and multi-dimensional scoring. The results highlight the joint importance of data construction, speaker conditioning, multi-objective optimization, post-processing, and transparent latency reporting. By releasing the refined dataset, baselines, and evaluation tools, we hope to advance robust, generalizable, and deployable real-world TSE systems.

XI. Post-challenge corrections

During data-release validation and camera-ready preparation, we identified two issues: (i) an audio-channel mapping error in AMI meeting IB4002, previously documented but still present in the source release used,[8] which caused incorrect speaker labels; and (ii) a text normalization error in our TER evaluation script. We excluded the 70 affected IB4002 samples, leaving 1,930 EVAL-1 samples (4,930 evaluation samples in total), and corrected the normalization. The results reported in this paper incorporate these corrections. Overall rankings remained largely unchanged. Original and corrected rankings are available at https://real-tse.github.io/challenge/#rankings.

[8]https://groups.inf.ed.ac.uk/ami/corpus/dataproblems.shtml

ACKNOWLEDGMENT

The REAL-TSE Challenge was sponsored by Huawei.

This work was supported by the National Natural Science Foundation of China (Grants 62401377 and 62271432), the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (Grant 2024CSJGG1100), and the Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant ZDSYS20230626091302006).

REFERENCES

[1] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[2] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[3] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[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] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[6] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[7] 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.

[8] R. Sutherland, J. Clarke, H. Elghazaly, T. Kuebert, M. Lugger, S. Petrausch, J. A. Ortiz, B. Xu, S. Goetze, and J. Barker, "Descriptor: Enhancing conversations for the hearing impaired in the 9th computational hearing in multisource environments challenge (CHiME9 ECHI)," *IEEE Data Descriptions*, vol. 3, pp. 73–81, 2026.

[9] 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 *Proc. ICASSP*. IEEE, 2022, pp. 6167–6171.

[10] 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 *Proc. Interspeech*, 2021, pp. 3665–3669.

[11] 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.

[12] S. Wang, K. Zhang, S. Lin, J. Li, X. Wang, M. Ge, J. Yu, Y. Qian, and H. Li, "WeSep: A Scalable and Flexible Toolkit Towards Generalizable Target Speaker Extraction," in *Proc. Interspeech*, 2024, pp. 4273–4277.

[13] 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 (MLMI 2005)*, ser. Lecture Notes in Computer Science, S. Renals and S. Bengio, Eds., vol. 3869. Berlin, Heidelberg: Springer, 2006, pp. 28–39.

[14] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *Proc. ICLR*, vol. 2024, 2024, pp. 44 440–44 455.

[15] 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*. IEEE, 2023, pp. 1–5.

[16] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2022, pp. 886–890.

[17] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*. IEEE, 2021, pp. 6493–6497.

[18] K. Xu, Y. Jia, K. Huang, J. Chen, W. Li, K. Liu, F.-L. Xie, X. Tang, and Y. Hu, "Fireredasr2s: A state-of-the-art industrial-grade all-in-one automatic speech recognition system," *arXiv preprint arXiv:2603.10420*, 2026.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] 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 *Proc. Interspeech*, 2020, pp. 434–436.

[21] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1215–1225, 2023.

[22] 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, pp. 1–5.

[23] 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.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[26] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging chinese speaker recognition dataset," in *Proc. ICASSP*, 2020, pp. 7604–7608.

[27] 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 *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[28] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2019, sound dataset. [Online]. Available: https://doi.org/10.7488/ds/2645

[29] 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.

[30] 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.

[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015. [Online]. Available: https://arxiv.org/abs/1510.08484

[32] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments," 2013, version 1.0. [Online]. Available: https://doi.org/10.5281/zenodo.1227121

[33] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[34] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[35] Y. Luo and R. Gu, "Fast random approximation of multi-channel room impulse response," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 449–454.

[36] 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.

[37] K. Pearson, "Notes on the history of correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 1920.

[38] C. Spearman, "The proof and measurement of association between two things." 1961.

[39] F. Cumlin, X. Liang, V. Ungureanu, C. K. Reddy, C. Schüldt, and S. Chatterjee, "DNSMOS Pro: A reduced-size DNN for probabilistic MOS of speech." in *Proc. Interspeech*, 2024, pp. 4818–4822.