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

***Abstract*—We introduce the REAL-TSE Challenge[1], a satellite challenge of IEEE SLT 2026 focused on Target Speaker Extraction (TSE) in real-world conversational environments. Given a multi-speaker mixture recorded in natural settings such as meetings and dinner-party interactions, together with one or more enrollment utterances from a target speaker, participants are asked to recover the target speaker's voice from the mixture. Unlike prior benchmarks that rely on simulated or read-speech data, REAL-TSE is grounded in real recordings spanning Mandarin and English, covering authentic reverberation, ambient noise, and naturally occurring conversational dynamics. The challenge features two tracks—an *Online* (low-latency, causal) track and an *Offline* (full-context) track—and evaluates systems along four complementary dimensions: intelligibility measured by Token Error Rate (TER), speaker consistency measured by Speaker Similarity (SpkSim), perceptual quality measured by Deep Noise Suppression Mean Opinion Score (DNSMOS), and target-speaker activity measured by the F1 score. We describe the challenge design, datasets, baseline systems, evaluation protocol, and the final DNSMOS metric adjustment, and highlight the scientific questions the challenge aims to address.**

I. Introduction

Target Speaker Extraction (TSE) aims to isolate the speech of a *specific* target speaker from a multi-talker mixture by leveraging auxiliary cues such as pre-recorded enrollment utterances [1]–[3]. In contrast to blind source separation, TSE enables speaker-aware interactive systems, making it directly applicable to teleconferencing, hearing aids, smart assistants, and meeting transcription pipelines.

Despite rapid advances in model architectures [4], [5], most prior work has been evaluated on *simulated* benchmarks such as LibriMix [6] and WSJ0-2Mix [7]. These simulated mixtures are typically created by artificially summing clean utterances at predefined signal-to-noise ratios, which can distort natural loudness relationships and often fail to capture authentic room reverberation and ambient noise. Moreover, most existing mixtures are dominated by *read* speech, and therefore do not reflect the spontaneous turn-taking, reactive overlaps, disfluencies, and non-verbal vocalizations that characterize real conversational scenarios.

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

Several prior speech challenges have addressed related problems. The CHiME series [8], [9] focuses on multi-speaker ASR in far-field settings. The recent CHiME-9 Task 2 (ECHI) [10] targets speech enhancement for wearable devices. AliMeeting [11] and AISHELL-4 [12] provide multi-channel Mandarin meeting data. However, none of these challenges directly benchmarks the *core TSE technology*—extracting a single target speaker from a real conversational mixture—across both online and offline processing modes and across multiple languages and diverse acoustic environments.

To close this gap, we introduce the **REAL-TSE Challenge**, whose main contributions are:

- **A real-world TSE benchmark dataset.** The challenge provides a development set derived from REAL-T [13], a real-recorded open-source corpus constructed from five speaker diarization corpora (AISHELL-4, AliMeeting, AMI [14], DipCo [15], CHiME6 [9]) covering Mandarin and English. The evaluation set further includes an in-domain subset (EVAL-1) and an independently collected unseen subset (EVAL-2), enabling assessment of both familiar-condition robustness and real-world generalization.
- **A dual-track task design.** The Online Track evaluates low-latency (≤ 100 ms) streaming TSE, while the Offline Track targets the upper performance bound of full-context systems without computational restrictions.
- **Multi-dimensional evaluation metrics.** Systems are scored along four axes: Token Error Rate (TER) for intelligibility, Speaker Similarity (SpkSim) for target-speaker consistency, Deep Noise Suppression Mean Opinion Score (DNSMOS) for perceptual quality, and target-speaker activity F1 for temporal precision and recall. Together, these metrics assess whether a TSE system extracts the right speaker, preserves intelligible speech, produces perceptually acceptable audio, and suppresses non-target regions.
- **Open-source baselines and validation tools.** We release four pretrained baseline models and reproducible recipes for both tracks, implemented using the previously released WeSep

framework [16][2]. We also provide an official validation toolkit for checking submission format and computing challenge metrics[3].

The rest of this paper is organized as follows. Section II describes the challenge tracks and metrics. Section III outlines the scientific questions the challenge addresses. Section IV presents the datasets. Section V describes the baseline system. Section VI explains the final DNSMOS metric adjustment. Section VII concludes and outlines future analysis.

II. Challenge Tracks and Evaluation

The REAL-TSE Challenge features two tracks, and participants may enter one or both. In each track, participants receive a set of *mixture–enrollment pairs*: a multi-speaker mixture $\mathbf{x}$ recorded in a real environment and an enrollment utterance $\mathbf{e}$ from the target speaker. The task is to produce an estimate $\hat{\mathbf{s}}$ of the target speaker's clean speech from $\mathbf{x}$.

A. Track 1: Online Target Speaker Extraction

This track evaluates TSE systems in latency-sensitive scenarios, such as assistive hearing devices, real-time communication, and interactive voice assistants. These applications require low-latency outputs under challenging acoustic conditions with overlapping speech and background noise.

System requirements. Participating systems must satisfy the following requirements:

- **Latency constraint.** The end-to-end algorithmic latency must not exceed 100 ms. In this challenge, the end-to-end algorithmic latency refers to the maximum amount of future input required by the whole inference pipeline to generate the corresponding output. This includes delays introduced by STFT, iSTFT, overlap-add, look-ahead frames, non-causal model context, smoothing, post-processing, and any required input accumulation or chunk-level buffering. Hardware-dependent computation time and I/O latency are not included.
- **Streaming output.** Models must support frame-wise or chunk-wise input and generate outputs continuously as new audio becomes available. They must operate in a streaming manner, without relying on full-utterance buffering or post-hoc batch processing.

Latency verification. In addition to self-reported algorithmic latency, the effective system latency will be verified by measuring the output response delay to a controlled input perturbation. This verification provides an objective check of real streaming behavior without requiring participants to implement a full streaming inference pipeline.

B. Track 2: Offline Target Speaker Extraction

This track evaluates TSE systems in offline scenarios, such as meeting transcription, audio post-production, and batch speech analytics. These applications have access to the entire recording before inference begins and therefore allow systems to exploit full-utterance context for higher extraction quality.

System requirements. Participating systems must satisfy the following requirements:

- **Full-context processing.** Systems may use the entire utterance and exploit global temporal dependencies across the recording. Unlike the Online Track, no causality or look-ahead constraint is imposed, allowing methods to perform full-context processing and offline optimization.
- **No architectural restrictions.** Model type, inference duration, and computational cost are unrestricted. Participants may use discriminative, generative, iterative, or hybrid approaches for offline processing.
- **Independent processing.** Each mixture–enrollment pair must be processed independently, without using metadata or information from other evaluation samples.

Encouraged directions. Long-context modeling, generation-based approaches, and iterative refinement methods are especially encouraged, as this track is intended to reveal the upper performance bound of full-context TSE systems.

C. Evaluation Metrics

Systems are evaluated along four complementary dimensions. *During the challenge and subsequent analysis stage, we found that neural non-intrusive metrics can become unreliable when heavily targeted for optimization. In particular, the originally announced DNSMOS OVRL metric was found to suffer from substantial over-optimization, making it unreliable for the final ranking; it was thus replaced by DNSMOS P808, as discussed in Section VI.* For each metric, all valid submissions are ranked independently using *dense ranking*, and the final challenge ranking is determined by **averaging the dense rankings across the four metrics**.

a) Intelligibility—Token Error Rate (TER): TER measures how accurately the extracted speech can be transcribed by an automatic speech recognition (ASR) system. It is computed as word error rate for English utterances and character error rate for Mandarin utterances, reflecting the standard tokenization used for the two languages. For TSE, this metric reflects whether the target speaker's linguistic content is preserved after extraction; a lower TER indicates better intelligibility. Official ASR backbone is zipformer [17], selected for stable transcription with reduced hallucinations.

b) Speaker Consistency—Speaker Similarity (SpkSim): SpkSim measures the cosine similarity between speaker embeddings extracted from the estimated signal $\hat{\mathbf{s}}$ and the enrollment utterance $\mathbf{e}$. In the context of TSE, this metric assesses whether the extracted signal preserves the target speaker's voice characteristics, rather than leaking speech from or switching to an interfering speaker; a higher similarity indicates better target-speaker consistency. The official scoring script uses a fixed speaker encoder based on the WeSpeaker ResNet-34 backbone [18], released with the validation toolkit, to ensure consistent embedding extraction across submissions.

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

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

c) Perceptual Quality—Deep Noise Suppression Mean Opinion Score (DNSMOS): DNSMOS [19] is a family of neural Mean Opinion Score estimators designed for non-intrusive perceptual speech quality assessment. In the context of TSE, it evaluates whether the extracted audio is natural and devoid of severe artifacts, distortion, or residual noise. The final leaderboard adopts DNSMOS P808 as the official perceptual quality metric, while previously reported DNSMOS OVRL scores are retained solely for reference.

d) Target-Speaker Activity—F1 Score: The F1 score measures whether the system outputs speech *only* in regions where the target speaker is active, combining temporal precision and recall. For TSE, precision penalizes non-target leakage or false speech in inactive regions, while recall penalizes missed target-speaker speech. All audio signals are normalized using a unified scaling procedure before this metric is computed; speech activity regions are detected using the FireRedVAD [20] model; a higher F1 is better.

III. Scientific Goals

The REAL-TSE Challenge aims to address fundamental questions in the field of real-world target speaker extraction:

- **Online vs. offline performance gap.** How large is the performance gap between causal low-latency systems (Track 1) and full-context systems (Track 2)? Can efficient streaming architectures approach the quality of offline counterparts?
- **Generalization to unseen acoustic environments.** Do systems trained on open-source corpora generalize to unseen acoustic scenarios (e.g., café or in-vehicle environments)? The EVAL-2 subset is designed to explore this question.
- **Impact of speaker conditioning.** How do different speaker representation strategies (e.g., fixed pretrained embeddings vs. jointly learned representations vs. multi-level conditioning) affect extraction quality in real conversational settings with short or overlapping enrollment segments?
- **Effect of training data domain.** Can models trained solely on simulated mixtures generalize to real-world conversational scenarios? Furthermore, which training strategies—such as data augmentation, domain adaptation, or advanced noise simulation—are most effective in bridging this domain gap?
- **Cross-lingual robustness.** As the evaluation spans both Mandarin and English, a key question arises: are current TSE architectures language-agnostic, or do they exhibit language-specific performance discrepancies?
- **Failure mode analysis.** What conversational phenomena (e.g., short target-speaker segments, high overlap ratio, rapid speaker turns, laughter, non-speech vocalizations) most severely degrade TSE performance?

IV. Datasets

A. Data Philosophy

Existing TSE benchmarks are predominantly constructed from simulated mixtures of clean, read speech. While such simulations facilitate controlled evaluation, they introduce systematic mismatches with real-world conditions: reverberation is typically approximated via the image-source method [21], loudness relationships are synthetically prescribed rather than naturally occurring, and turn-taking patterns derived from audiobooks fail to reflect spontaneous conversational dynamics. REAL-TSE addresses these limitations by grounding *all evaluation data* in real conversational recordings, where overlap, reverberation, noise, and speaker behavior emerge naturally.

TABLE I
Statistics of the three challenge splits. Samples are released mixture–enrollment pairs; mixtures and enrollments are unique WAV counts; target speakers are distinct target-speaker labels.

	DEV	**EVAL-1**	**EVAL-2**
Samples	1,991	2,000	3,000
Mixtures	528	595	1,186
Enrollments	242	183	1,919
Target speakers	64	49	77
Mixture audio (h)	2.60	2.98	5.67
Enrollment audio (h)	0.67	0.55	4.05

TABLE II
Per-sample statistics for the main split difficulty descriptors.

Split	**Quantity**	**mean**	**std**	**min**	**max**
DEV	Mix dur. (s)	17.26	6.23	6.24	29.91
	Enroll dur. (s)	9.65	5.90	5.00	50.51
	Overlap ratio	0.49	0.17	0.18	0.99
	Target ratio	0.75	0.16	0.27	1.00
EVAL-1	Mix dur. (s)	18.23	6.02	5.82	29.99
	Enroll dur. (s)	10.94	8.02	5.01	59.82
	Overlap ratio	0.48	0.16	0.18	0.98
	Target ratio	0.75	0.16	0.21	1.00
EVAL-2	Mix dur. (s)	17.25	6.03	5.48	29.91
	Enroll dur. (s)	9.10	8.65	3.00	42.98
	Overlap ratio	0.53	0.17	0.18	0.95
	Target ratio	0.73	0.14	0.31	1.00

B. Challenge Splits and Statistics

Table I summarizes the three challenge splits. Each released sample is one mixture–enrollment pair; multiple samples may share the same mixture or enrollment, so the unique-WAV counts are smaller than the sample counts. In total, REAL-TSE provides 6,991 extraction trials over 2,309 distinct mixtures and 11.3 h of mixture audio, spanning Mandarin and English.

Table II reports per-sample statistics for the main difficulty descriptors used to characterize the splits: mixture duration, overlap duration, overlap ratio, target-speaker active ratio, and enrollment duration. These statistics are included to make the released data transparent and to support comparable system-description papers; a comprehensive impact analysis on submitted systems is deferred to the post-challenge publication.

C. Development Set

The development set contains 1,991 mixture–enrollment pairs derived from REAL-T [13], a real-recorded open-source corpus designed for TSE under complex acoustic conditions. REAL-T is constructed from five speaker diarization corpora: **AISHELL-4**, **AliMeeting**, **AMI**, **DipCo**, and **CHiME6**, covering Mandarin and English in meeting and dinner-party scenarios.

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

Each sample in the development set comprises:

- a multi-speaker mixture featuring spontaneous, naturally overlapping speech in real-world scenarios.
- an enrollment utterance from the target speaker (a non-overlapping segment of at least 5 seconds).

The construction pipeline extracts overlapping segments as mixtures and selects non-overlapping segments as enrollment utterances through an automated procedure. This approach captures realistic overlap patterns, reverberation, ambient noise, and conversational turn-taking as they naturally manifest in unconstrained environments.

The development set is intended for hyper-parameter tuning, model comparison, and ablation studies. It must **not** be used for model training or fine-tuning.

D. Evaluation Set

The evaluation set consists of two subsets totaling **5,000 mixture-enrollment pairs**, as summarized in Table I:

a) EVAL-1 (Seen, 2,000 pairs): Derived from the same open-source corpora as the development set, sharing similar acoustic conditions and conversation styles. There is no overlap between EVAL-1 and the development set. EVAL-1 measures system performance under familiar, in-domain settings.

b) EVAL-2 (Unseen, 3,000 pairs): Newly collected real-world conversational recordings specifically captured for the REAL-TSE Challenge. EVAL-2 covers diverse acoustic scenarios including meeting rooms, cafés, home environments, and in-vehicle conversations. To make the evaluation reflect practical device mismatch, each conversation was synchronously recorded with multiple devices, including two high-quality microphones (H1 and H2, different distances), a mobile phone, and speaker-worn close-talk headset microphones. Besides the enrollment utterances derived from the recordings similar to the REAL-T dataset, each recording participant also provided extra single-speaker utterances outside the conversation, which are used as external enrollment audio. EVAL-2 therefore provides a comprehensive assessment of model generalization to previously unseen acoustic environments, recording devices, and enrollment–mixture channel conditions.

Table III summarizes the EVAL-2 channel pairing design. The split includes both matched- and cross-channel mixture-enrollment pairs, enabling subsequent analysis of channel-mismatch robustness upon receiving participant submissions.

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

To ensure a fair evaluation, metadata such as scenario labels, speaker identities, and recording device information has been removed from the released evaluation set. Participants are required to process each mixture–enrollment pair *independently*.

E. Training Data Policy

The challenge does **not** define an official training set. Participants are free to exploit any open-source data assets and public pretrained checkpoints, under the strict condition that all assets, pipelines, and models are transparently documented in their system descriptions to ensure reproducibility.

However, the following corpora's *development and test splits* must **not** be used at any stage, including pre-training, training, or data augmentation: AliMeeting, AISHELL-4, AMI, DipCo, and CHiME6. The official *training splits* of these corpora are permitted. For the AMI corpus, which has multiple partition schemes, participants should follow the Full-corpus-ASR partition; the excluded development and test sessions are listed in the challenge FAQ.

This open-training design is deliberate: by removing constraints on training resources, the challenge aims to uncover the practical upper bounds of current technologies while encouraging diverse solutions that span data engineering, novel model architectures, and innovative training strategies.

V. BASELINE SYSTEM

A. Model Variants

We provide four BSRNN-based [4] baseline systems [22] (Table IV) developed within the WESEP framework [16]. All systems are trained on the fully-overlapped **Libri2Mix-100** dataset [6] at 16,kHz for 150 epochs, utilizing an exponential learning rate decay from 10^{-3} to 2.5×10^{-5}. These baselines evaluate two distinct modeling strategies for incorporating target-speaker enrollment information, compared across both causal and non-causal extractor configurations.

a) Embedding-based conditioning.: `BSRNN_EMB` uses a fixed utterance-level speaker embedding extracted by a pretrained ECAPA-TDNN [18], [23] speaker encoder, serving as the conventional high-level speaker identity cue for TSE.

b) TF-Map + Context.: `BSRNN_TFMAP` uses a multi-level conditioning method that combines a spectral-level TF-Map feature with a contextual embedding feature [22]. This method retains lower-level time-frequency information from the enrollment utterance in addition to neural speaker information.

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	0.693	0.501	2.82	1.66	0.841
	BSRNN_EMB_CAUSAL	0.705	0.492	2.69	1.63	0.829
	BSRNN_TFMAP	0.703	0.521	2.72	1.50	0.838
	BSRNN_TFMAP_CAUSAL	0.652	0.535	2.76	1.56	0.844
EVAL-1	BSRNN_EMB	0.816	0.507	2.91	1.91	0.827
	BSRNN_EMB_CAUSAL	0.806	0.500	2.76	1.78	0.807
	BSRNN_TFMAP	0.837	0.535	2.80	1.75	0.822
	BSRNN_TFMAP_CAUSAL	0.801	0.553	2.84	1.79	0.837
EVAL-2	BSRNN_EMB	0.838	0.357	2.85	1.80	0.830
	BSRNN_EMB_CAUSAL	0.811	0.370	2.72	1.67	0.831
	BSRNN_TFMAP	0.839	0.382	2.73	1.52	0.833
	BSRNN_TFMAP_CAUSAL	0.808	0.391	2.75	1.59	0.835

B. Causal Configuration

The aforementioned conditioning methods naturally support causal inference, as the enrollment features can be precomputed before processing the mixture, and the conditioning fusion is applied frame by frame. Consequently, the causal baselines retain the original speaker-conditioning modules and impose causality exclusively within the BSRNN extractor. Specifically, global normalization is replaced with feature-axis normalization, and the bidirectional LSTM layers along the time axis are replaced with unidirectional LSTM layers.

C. Baseline Results

Table V reports the official baseline results on DEV, EVAL-1, and EVAL-2. Since the released EVAL-1 and EVAL-2 datasets do not include ground-truth references, these reported scores serve as the official benchmark and cannot be evaluated locally using the validation script.

These numbers are provided as reference points for verifying the evaluation pipeline and calibrating future system descriptions. Since all baselines are trained exclusively on synthetic Libri2Mix mixtures, their performance should be interpreted as a practical lower bound rather than a ceiling for REAL-TSE.

VI. METRIC ADJUSTMENT IN THE FINAL EVALUATION

Near the final stage of the challenge, several teams reported anomalous or severely inconsistent behavior of DNSMOS OVRL. Following the completion of official submissions, we conducted a rigorous post-submission analysis encompassing the submitted systems, automatic metrics, and supplementary human listening tests. This investigation revealed that DNSMOS OVRL, which was originally designated as the perceptual quality metric, no longer reliably reflected the perceived quality of the extracted target speech in this challenge. Specifically, several systems achieved inflated DNSMOS OVRL scores that failed to consistently align with the subjective quality observed in human mean opinion score (MOS) evaluations. This mismatch strongly suggests that DNSMOS OVRL was substantially over-optimized in this competitive setting, rendering it ineffective as a reference-free perceptual metric.

The same issue is also visible from the relationship between DNSMOS OVRL and DNSMOS P808 on participant submissions. Since both metrics are intended to estimate perceptual speech quality, genuine improvements in extracted-speech quality should generally lead to consistent variation in the two scores across systems. However, Fig. 1 shows clear outliers: some systems obtain unusually high DNSMOS OVRL scores without corresponding gains in DNSMOS P808. This pattern indicates metric-specific over-optimization rather than a uniform improvement in perceptual quality.

We then compared the DNSMOS variants against human MOS to determine which metric should replace OVRL. Table VI reports the Linear Correlation Coefficient (LCC) and Spearman's Rank Correlation Coefficient (SRCC) between human MOS and the three DNSMOS variants on the post-submission listening set. DNSMOS OVRL exhibits negligible correlation with human MOS in Track 1 and remains weak overall. While DNSMOS PRO improves upon OVRL, DNSMOS P808 yields the highest and most robust correlations with human MOS across both tracks. These results provide direct empirical justification for substituting OVRL with P808 in the final perceptual quality evaluation.

TABLE VI
CORRELATION BETWEEN HUMAN MOS AND DNSMOS VARIANTS (LCC / SRCC).

		DNSMOS-OVRL		DNSMOS-P808		DNSMOS-PRO	
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

Together, these observations highlight a broader limitation of reference-free neural evaluation metrics. When such metrics are public and can be used directly during system development, they may be unintentionally over-optimized or even exploited by model training, post-processing, or selection strategies. As a result, the metric can become less correlated with human perception, even if systems achieve higher automatic scores. This issue is particularly important for real-world TSE, where perceptual quality must be balanced with speaker consistency, intelligibility, and target-speaker activity.

After repeated discussions within the organizing committee and consultation with external experts, we decided to replace DNSMOS OVRL with DNSMOS P808 in the final evaluation. DNSMOS P808 is based on a different evaluation model and is independent from DNSMOS OVRL. More importantly, among the DNSMOS variants examined in our listening analysis, it showed the highest agreement with human MOS. We therefore believe that it provides a more reasonable and robust assessment of perceptual quality for the final leaderboard.

We emphasize that utilizing DNSMOS OVRL during system training did not violate any challenge regulations, and systems

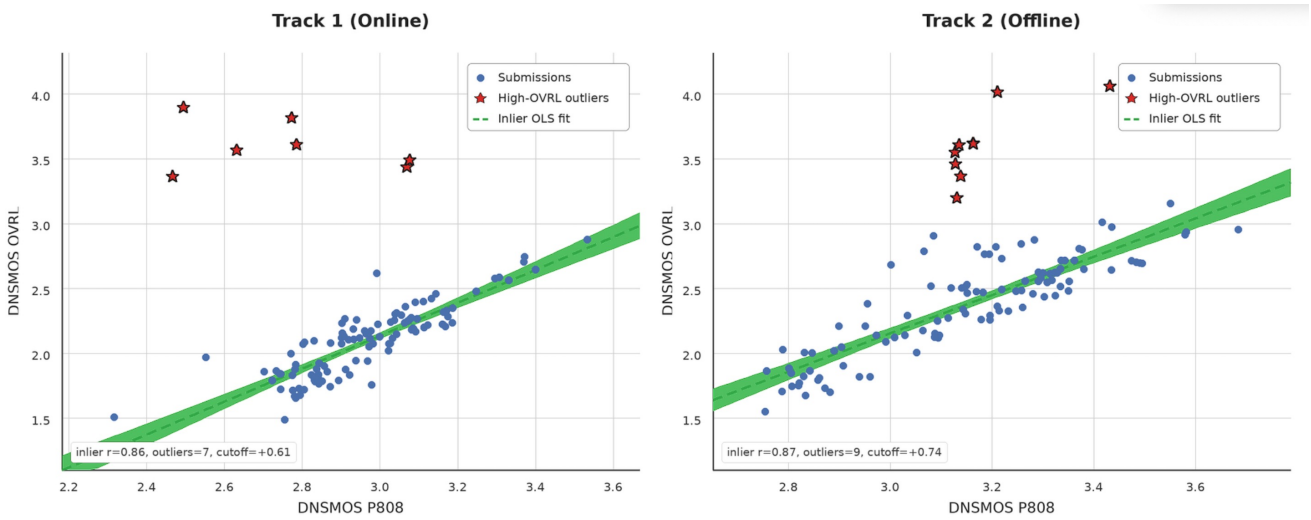

Fig. 1. Relationship between DNSMOS OVRL and DNSMOS P808 on participant submissions. The high-OVRL outliers do not show corresponding gains in P808, indicating metric-specific over-optimization rather than a consistent improvement in perceptual quality.

optimized for this objective achieved exceptional performance under the originally announced criterion. We also acknowledge that altering an evaluation metric post-submission is suboptimal and may shift the relative rankings among participating teams. Nevertheless, the primary objective of this challenge is to evaluate the authentic perceptual quality of TSE systems as fairly and meaningfully as possible, rather than to reward over-optimization toward a metric that has proved unreliable for this specific task. Consequently, DNSMOS P808 has been adopted as the official perceptual metric for the final leaderboard.

VII. Conclusions and Future Analysis

The REAL-TSE Challenge advances the field by bridging the gap between simulated laboratory benchmarks and real-world target speaker extraction. Grounded in authentic bilingual recordings spanning diverse acoustic environments, the challenge provisions both online and offline tracks to enable a rigorous, multi-faceted evaluation of modern TSE systems. Its open-training policy and multi-dimensional scoring framework effectively drive the development of creative, practically deployable solutions.

This overview paper delineates the task definition, dataset design, evaluation protocol, baseline recipes, validation tools, and the final metric adjustment. Subsequent analyses will comprehensively dissect participant submissions and challenge results, focusing on modeling architectures, speaker-conditioning paradigms, training data formulations, causal versus non-causal processing designs, and robustness techniques under real conversational conditions. Furthermore, we will benchmark system behaviors across different languages, acoustic scenarios, overlap patterns, and evaluation metrics, with particular emphasis on trade-offs where improvements in one dimension conflict with another.

Beyond system-level findings, subsequent publications will document methodological insights gained from the official evaluation process itself, encompassing metric reliability, latency verification, speech activity scoring, subjective listening test designs, data release policies, and the engineering challenges of benchmarking real-world TSE systems. We expect these reflections to uncover both the limitations of our current setup and concrete pathways for refining future editions of REAL-TSE and related real-recorded speech evaluation frameworks. Ultimately, we hope the REAL-TSE benchmark will catalyze the speech community toward developing TSE systems that are robust, highly generalizable, and seamlessly deployable in real conversational scenarios.

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] 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.

[5] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1406–1410.

[6] 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.

[7] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

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

[9] 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.

[10] 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*, 2025.

[11] 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. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2022, pp. 6167–6171.

[12] 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*, 2021, pp. 3665–3669.

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

[14] 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.

[15] 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*, 2020, pp. 434–436.

[16] 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 *Interspeech 2024*, 2024, pp. 4273–4277.

[17] 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 *International Conference on Learning Representations*, vol. 2024, 2024, pp. 44 440–44 455.

[18] 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. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.

[19] 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. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 886–890.

[20] 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.

[21] 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.

[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 *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (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*, 2020, pp. 3830–3834.