REAL-TSE 2026 challenge: system description of the ChuEst submissions

Participant name (as on the leaderboard): `chuest_kts`
Team name (as on the leaderboard): `ChuEst`
Tracks: Track 1 (Online) and Track 2 (Offline)
Date: July 2, 2026 (version 1.0). Revised July 7, 2026 (version 1.1).

Revision note (version 1.1)

This revision fixes the data usage description, after our email exchange with the organizing committee on July 6-7, 2026. The systems, the submissions, and all scores are not changed. Only the text is fixed. The changes are:

1. Section 3.2: the version 1.0 text wrongly wrote the `usef_realt_scp` pairs as Track 2 stage-1 training data. Those pairs were a validation input only, never training data (change 5 below explains what that split actually is). The real stage-1 training data was AliMeeting training-partition data (simulated near-field mixtures and far-field training recordings). The v3combo set was used in a later fine-tuning stage of the Track 1 line, not in stage 1. Section 3.2 now says this correctly.
2. Section 4 (dataset table): the REAL-T / REAL-TSE DEV row is fixed in the same way: validation and checkpoint selection only, on both tracks. Never used for training or fine-tuning in our ranked entries; the withdrawn #296 is covered in change 4.
3. Section 2.4: the words “fine-tuned on REAL-T at 16 kHz” came from our internal naming confusion (our internal file names use “realt” as a project prefix). That stage was trained on AliMeeting training-partition data (section 3.2 is about the same run). No REAL-T corpus audio and no REAL-TSE DEV audio was in any training input. The wording is fixed.
4. Section 3.6: version 1.0 wrote that the DEV fine-tuning of our fallback candidate #296 was “explicitly permitted”. This was our wrong reading of the DEV data usage rule (validation, model comparison, and hyper-parameter tuning only). We fixed the sentence, we withdraw #296 as a fallback, and we do not name a new fallback. Our ranked entry #188 did not train or fine-tune on DEV, so it does not need a fallback.
5. Section 3.3: version 1.0 called the validation split `usef_realt_scp/valid` a held-out split of the REAL-TSE DEV set. It is actually our internal synthetic AliMeeting validation split. The official DEV set was used only for checkpoint selection and model comparison. The text is fixed.

1. Overview

This document describes the two systems behind our leaderboard entries in the REAL-TSE 2026 challenge. Both are target-speaker extraction systems built on USEF-TFGridNet. They share one early 16 kHz warm-start ancestor (sections 2.4 and 3.2), but after that point they were developed separately and the final systems are quite different.

Our Track 1 (Online) entry is submission #279, `usef_b5_dnsmos17s_iter1000`, a causal block-online USEF-TFGridNet with 15.64 M parameters and a measured algorithmic latency of 85.6 ms. On the leaderboard it stands at rank 8 with SCORE 8.2500, and its speaker similarity of 0.533 is the highest of all participating teams on that track.

Our Track 2 (Offline) entry is submission #188, `submit_reuse_5`, a two-stage pipeline: a USEF-TFGridNet extractor followed by the released nvidia/RE-USE speech enhancer, with a per-file routing step that decides for each utterance whether the enhanced or the raw extracted signal is kept. It is at rank 6 with SCORE 9.7500. All leaderboard figures in this paper are the values displayed on July 2, 2026; positions remain provisional until the organizers finalize the results.

For each track we cover the model architecture, training objectives, data usage, validation strategy, and post-processing, as requested in the organizers' instructions. Section 2.8 contains the end-to-end latency report required for Track 1. Section 3.6 discusses the routing step of our Track 2 system in light of the clarification the organizers sent on June 27, because we want this paper to be a complete and faithful description of the submitted systems.

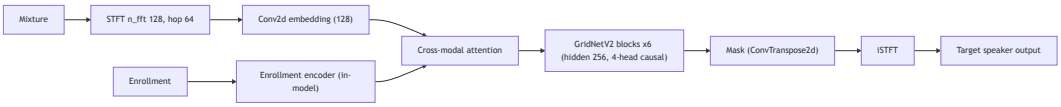
2. Track 1 (Online) system

2.1 Model architecture

The model is `Tar_Model`, a USEF-TSE target-speaker extractor built on a TF-GridNet separator with cross-modal attention over the enrollment utterance. It has 15,641,885 parameters (15.64 M across 290 tensors). We note this explicitly because it is not the 33.5 M BSRNN baseline provided with the challenge.

The mixture is transformed by an STFT with n_fft 128 and hop 64 at 16 kHz, which gives a 4 ms frame rate and 65 frequency bins, and a Conv2d layer embeds the spectrogram into 128 channels. The enrollment utterance is encoded by the model's own encoder and fused with the mixture representation through cross-modal attention. USEF-TSE is speaker-embedding-free: no external speaker-embedding network is part of the deployed model. The fused representation then passes through six GridNetV2 blocks, each combining an intra-frame BiLSTM, a block-online inter-frame BiLSTM, and 4-head causal self-attention with hidden size 256. A ConvTranspose2d layer produces the mask, and an inverse STFT reconstructs the target-speaker waveform. We submit this waveform as-is; no post-processing is applied on Track 1.

Figure 1. Track 1 model (USEF-TFGridNet, submission #279)



2.2 Causality and latency

The submitted configuration ("b5", W = 24) runs the time axis as a 24-frame block-online window with causal patches enabled throughout (`p_rnn`, `p_attn`, `p_conv`, `p_norm`, `p_inorm` all true). The system latency as submitted is 92 ms, and the algorithmic latency measured with the official verification script is 85.6 ms. Both are within the 100 ms budget for Track 1. A 6.4 ms value appears in our earlier submission history, but it belongs to a different submission (#199) and does not apply to #279.

The DNSMOS surrogate described in the next section is a training-time loss only. It is not part of the shipped weights, so it has no effect on the architecture or on latency.

We confirm that our Track 1 submission fully satisfies the requirements specific to the Online track.

2.3 Training objectives and recipe

Training warm-started from `usef_lookahead_b5/temp_best.pth.tar` (all 290 tensors loaded, DNSMOS 1.731 at that anchor point). We used Adam with learning rate 2e-5, gradient clipping at max_norm 5, batch size 1, and 17-second crops at 16 kHz. Runs were deterministic (seed 3407, `cudnn.deterministic`).

The loss has four terms:

1. SI-SNR reconstruction loss.
2. An energy penalty (weight 250) that acts as a proxy for the timing F1 metric.
3. An ECAPA-TDNN speaker-cosine loss (weight 5.0) as a proxy for speaker similarity. The ECAPA network is used only inside this loss, not in the deployed model.
4. A differentiable DNSMOS surrogate (weight 2.0): the official `sig_bak_ovr.onnx` converted with onnx2torch and frozen, with the objective of raising the mean OVRL score. This term is masked to active speech regions so the network cannot raise its score by outputting silence.

The official metrics were used only in this way, as training-time losses, which the challenge rules permit. Evaluation was a single pass with fixed weights. The checkpoint that produced submission #279 is `iter1000`.

2.4 Data usage

The model was trained only on open-source corpora (their training splits) plus standard augmentation. The REAL-T and REAL-TSE DEV and EVAL audio was used for validation and testing only, never as training data. One exception should be stated clearly: we used utterance-length statistics of the evaluation set (metadata only, no

audio) to length-match our re-synthesized training mixtures at 17 seconds. That re-synthesized set contains 7,600 mixtures of 17 seconds each, built with real room impulse responses (RT60 0.5 to 0.8 s), signal-to-interference ratios drawn uniformly from -5 to +5 dB, and 3 to 4 speakers per mixture; the source material is AliMeeting (40%), CHiME-6 (27%), AMI (23%), and AISHELL-4 (10%), training splits only.

The warm-start history, from the original public model to the submitted model, is: the public USEF-TSE model (USEF-TFGridNet, arXiv:2409.02615, CC BY-NC 4.0), first adapted at 16 kHz on AliMeeting training-partition data (this run is the same checkpoint that serves as the Track 2 stage-1 extractor; see section 3.2), then fine-tuned for timing F1 on our v3combo set, then converted from bidirectional to block-online (W = 24, weights transferred verbatim), then trained as "b5" on Libri2Mix train-clean-100, and finally trained with the length-matched 17-second recipe plus the DNSMOS, ECAPA, and energy losses described above, which produced #279.

A wording fix from version 1.0: we called that first 16 kHz adaptation stage "fine-tuned on REAL-T". This came from our internal file naming. Our file names use "realt" as a project prefix (for example `usef_realt_scp_v3combo`). No REAL-T corpus audio and no REAL-TSE DEV audio was in any training input, in this stage or any later stage. In the whole warm-start history above, DEV data appears only in validation and checkpoint-selection roles, never as training input. During development we kept the training sets separate from the DEV and EVAL sessions: we checked the session ranges, and we applied an exclusion list of 4,644 utterance IDs covering DEV/EVAL sessions when we built the training sets.

Section 4 lists the datasets and pretrained models with licenses.

2.5 Validation strategy

We validated on the REAL-TSE DEV set with the official metrics and selected checkpoints by their DEV scores. The recorded DEV scores of the selected `iter1000` checkpoint are TER 0.583, SIM 0.564, DNSMOS OVRL 1.762, and timing F1 0.860.

2.6 Submission and results

The leaderboard entry is a single best-counted submission: #279, `usef_b5_dnsmos17s_iter1000_track1.zip`, submitted June 24, 21:25 AoE. Its official scores are TER 0.749 (rank 8), TIMING F1 0.842 (rank 8), SIM 0.533 (rank 1 of all teams), DNSMOS OVRL 1.877 (rank 16), SCORE 8.2500, shown as rank 8 on the platform (dense ranking; two teams tie at rank 5). Our later submissions (#297, #300, #338) did not beat #279, so #279 is the version on the final leaderboard, following the rule that only each team's best submission appears there.

The scored artifact (5,000 wav files and 6 csv files covering EVAL1 and EVAL2, 16 kHz mono PCM16) is preserved with md5 `026545b5d66b7ae4e0ed38963765a01b`.

2.7 Checkpoint provenance

We also want to report one operational mistake transparently. The original `iter1000.pth.tar` was deleted on June 25 by an automatic keep-8 retention policy (saved 13:10, pruned 18:05, per the training log). We regenerated it deterministically on the same machine and training harness (all 290 warm-start tensors reproduced; the archived file has md5 `193e32a2`). The consistency check is complete: on the DEV set the regenerated checkpoint reproduces the recorded `iter1000` scores within 0.003 to 0.017 on all four metrics (TER 0.580 vs 0.583, SIM 0.571 vs 0.564, DNSMOS OVRL 1.745 vs 1.762, timing F1 0.865 vs 0.860). This difference is consistent with GPU floating-point nondeterminism, and for the same reason the regenerated weights are not byte-identical to the original ones. The originally scored audio output was never lost and is preserved separately (md5 above). The same training harness re-created another checkpoint of ours bit-exactly before.

2.8 Latency report (Track 1 requirement)

Model: USEF-TFGridNet b5, STFT window 128, hop 64 at 16 kHz, block-online window W = 24.

Algorithmic latency: 85.6 ms, measured with the official `utils/latency_verification.py`. The nominal value is (W - 1) x hop = 23 x 4 ms = 92 ms. The script reports Max at threshold 1e-5, and two threshold settings agree, so we consider the 85.6 ms a real causal latency, not a numerical artifact.

Buffering latency: about 4 ms, one STFT hop of framing, and it is subsumed within the measured algorithmic latency.

End-to-end: approximately 85.6 ms, which is within the 100 ms Track 1 budget.

3. Track 2 (Offline) system

3.1 System architecture

The Track 2 entry (#188, `submit_reuse_5`, submitted June 17) is simple: two deterministic stages and a per-file routing rule. No fine-tuning is involved in the second stage.

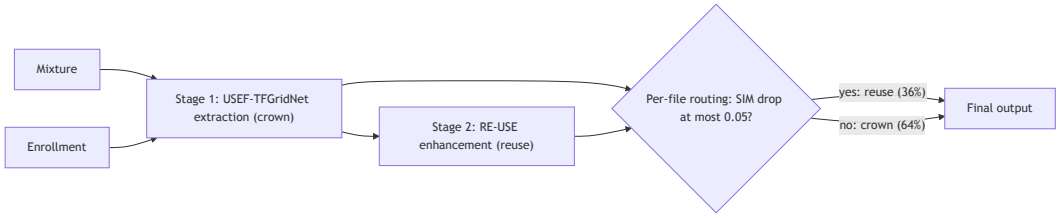
Stage 1 is a USEF-TFGridNet target-speaker extractor conditioned on the enrollment utterance. The TF-GridNet mask network uses STFT n_fft 128 and hop 64, hidden size 256, embedding dimension 128, 4 attention heads, and 6 layers. Its input is the mixture plus the enrollment; its output is the raw extracted target, which we call “crown” internally.

Stage 2 is the nvidia/RE-USE enhancer exactly as released, without fine-tuning. RE-USE is a USEMamba/SEMamba-style time-frequency Mamba universal speech enhancer: 30 TF-Mamba blocks with d_state 16, d_conv 4, expand 4, hid_feature 64, a

two-channel magnitude-plus-phase input, a mapping head, and sampling-frequency-independent operation (n_fft 320 to 640 at 16 kHz). It takes the stage 1 output and produces an enhanced version, which we call “reuse”.

The routing step then chooses, per utterance, between the two candidate outputs. Section 3.4 describes the rule and section 3.6 discusses its compliance status.

Figure 2. Track 2 system (two stages plus per-file routing, submission #188)



3.2 Training objectives and data

Only stage 1 was trained. Its objective is SI-SDR (the standard USEF-TSE negative SI-SDR loss), optimized with Adam at learning rate 1e-4 and batch size 4. Training warm-started from the released 8 kHz pretrained USEF-TFGridNet (loaded with `strict=False`), and resumed from epoch 13 through epoch 16. The best checkpoint, epoch 16, is the one in the submitted system.

The stage-1 training data was built only from the AliMeeting training partition: simulated near-field mixtures, and the AliMeeting far-field training recordings. No AMI data, no REAL-T corpus audio, and no REAL-TSE DEV or EVAL audio was used for training in this run. The validation and test path in the config (`usef_realt_scp/valid`) is our internal synthetic AliMeeting validation split, not the official DEV set. The official REAL-TSE DEV set was used only for checkpoint selection and model comparison (section 3.3), never for training or fine-tuning.

Stage 2 used no training and no training data: the RE-USE weights are the released ones, downloaded from Hugging Face (`nvidia/RE-USE`).

No official evaluation model, score, gradient, or embedding was used to train stage 1. Its objective is SI-SDR on training data, which the organizers permit.

3.3 Validation strategy

We validated stage 1 during training on our internal synthetic AliMeeting validation split (`usef_realt_scp/valid`). For checkpoint selection and model comparison we used the official REAL-TSE DEV set with the official metrics, and we selected the checkpoint by DEV speaker similarity and TER. (Version 1.0 called `usef_realt_scp/valid` a held-out split of the official DEV set. That was wrong naming from our side; it is an internal synthetic split. The official DEV set was used only in the checkpoint selection and model comparison step.)

3.4 Post-processing: per-file routing

For each of the 5,000 evaluation utterances we computed a speaker-similarity estimate of both candidates (crown and reuse) against the enrollment, using the publicly available WeSpeaker speaker-embedding models (via `wespeakerruntime`, the same package and default models the official evaluation toolkit uses for its speaker-similarity metric). The rule: take the enhanced output only if it does not lower the similarity estimate by more than 0.05 relative to crown (sim_crown minus sim_enh at most 0.05); otherwise keep crown. The recorded routing decisions confirm the boundary sits exactly at 0.0500, and the split is 3,210 utterances on crown (64%) and 1,790 on reuse (36%). The rule is designed to protect speaker similarity: enhancement is applied only where it does not reduce the similarity, so the DNSMOS benefit of RE-USE is limited to that 36% subset.

3.5 Submission and results

The leaderboard entry is #188, `submit_reuse_5`. As of July 2, 2026 it stands at rank 6 (dense ranking; one other team ties at the same score) with SCORE 9.7500: TER 0.710 (rank 8), TIMING F1 0.831 (rank 13), SIM 0.532 (rank 4), DNSMOS OVRL 2.173 (rank 14). Speaker similarity is our strongest axis on this track as well.

For completeness, our later fine-tuned candidates scored strictly worse and never displaced #188: #296 (f1aware1700), #309 (depthDEV), and #335 (reuse_frcrn). The standing best is the earlier and simpler system described here.

3.6 Compliance discussion of the routing step

On June 27, 22:29 KST, about 25 hours after the submission deadline (June 25, 23:59 AoE), the organizers sent a clarification stating that “for evaluation submissions, teams must not use the official evaluation models… to… select among candidates… for individual evaluation utterances”, and that “candidate selection with respect to DNSMOS, speaker similarity, or any other official metric, will not be accepted for the official leaderboard or final ranking”, inviting affected teams to contact them to identify a valid prior submission. The earlier organizer emails did not contain this clause, and we give that faithful description in this paper.

Our position on #188 is the following. The routing rule was fixed before submission and is fully deterministic: identical input always yields identical output, and no leaderboard feedback entered any decision. The similarity estimates were computed locally with the publicly available WeSpeaker models (via `wespeakerruntime`), the same package and default models the official evaluation toolkit uses for its speaker-similarity metric. We did not query the official scoring pipeline or use leaderboard scores per utterance. However, we state clearly that the routing is a per-utterance selection between two candidate outputs based on a speaker-similarity estimate. No such restriction existed in the challenge materials when #188 was submitted (June 17) or when the submission window closed (June 25 AoE); how the June 27 clarification applies retroactively is for the organizers to decide.

A correction from version 1.0 about the fallback. Version 1.0 wrote here that fine-tuning with DEV-computed objectives was "explicitly permitted", and named #296 (`f1aware1700`) as our valid prior submission. That was our wrong reading of the rules. The DEV set may be used for validation, model comparison, and hyper-parameter tuning only, and must not be used for training or fine-tuning. Our records show that the fine-tuning of #296 used DNSMOS, speaker-similarity, and VAD objectives computed on the DEV set, so #296 does not follow that rule. So we withdraw #296, and we do not name any fallback submission. Our ranked entry #188 is not affected: its stage-1 training used no DEV data (section 3.2), and its stage-2 had no training at all.

3.7 Weights and artifacts

All items below are md5-verified on our compute node.

- Stage 1 checkpoint: `epoch16.pth.tar` (188 MB), md5 `3c330b7ec87b8677a926f2fd7be1581b`, intact from the June 17 training run, with an identical backup copy. Config: `USEF-TSE-weights/chkpt/USEF-TFGridNet/config.yaml`.
- Stage 2: `model.safetensors` (38.6 MB), md5 `ee3845cd2db6c72c88bab53da31531f2`, identical to the Hugging Face `nvidia/RE-USE` release.
- Routing decisions: 22 csv files (`reuse_route_EVAL1_*.csv`, `reuse_route_EVAL2_*.csv`) with columns split, corpus, utterance, sim_raw, sim_enh, chosen.
- Routed submission output: 5,000 wav files (EVAL1 2,000 plus EVAL2 3,000, PCM16, 16 kHz, mono).

4. Datasets and pretrained models

The two tables below report every dataset and pretrained model used in either track, with citations or links and licenses.

Datasets (training splits only, unless noted):

Dataset	Role in our systems	Source / citation	License
AliMeeting (M2MeT)	40% of the Track 1 re-synthesized fine-tune mixtures; source of the stage-1 training set (both tracks' warm-start root) and, with AMI, of v3combo	Yu et al., ICASSP 2022; OpenSLR SLR119	CC BY-SA 4.0
CHiME-6	27% of the Track 1 re-synthesized mixtures	Watanabe et al., CHiME-6 Challenge 2020	research use via the CHiME challenge
AMI Meeting Corpus	23% of the Track 1 re-synthesized mixtures; also in v3combo	Carletta et al., 2006	CC BY 4.0

Dataset	Role in our systems	Source / citation	License
AISHELL-4	10% of the Track 1 re-synthesized mixtures	Fu et al., Interspeech 2021; OpenSLR SLR111	CC BY-SA 4.0
Libri2Mix (train-clean-100)	earlier Track 1 fine-tune stage (“b5”)	Cosentino et al., 2020, arXiv:2005.11262	CC BY 4.0 (LibriSpeech-derived)
MUSAN	additive noise/music augmentation (speech subset excluded)	Snyder et al., 2015, arXiv:1510.08484	CC BY 4.0
RIRS_NOISES (simulated RIRs)	reverberation augmentation	Ko et al., ICASSP 2017; OpenSLR SLR28	Apache 2.0
WHAM!	noise corpus of the pretrained base (inherited via USEF-TSE)	Wichern et al., Interspeech 2019	CC BY-NC 4.0
AliMeeting-derived stage-1 set (ours, derived)	Track 2 stage-1 training, i.e. the 16 kHz adaptation at the root of both tracks (sections 3.2 and 2.4): simulated near-field mixtures plus far-field training recordings, AliMeeting training partition only	this work (derivative)	inherits the source license
v3combo (ours, derived)	Track 1 fine-tune stage for timing F1 (section 2.4): 493 speakers / 6,000 mixtures built from AliMeeting + AMI training splits	this work (derivative)	inherits the source licenses
REAL-T / REAL-TSE DEV	validation and checkpoint selection only (both tracks); never used for training or fine-tuning in the systems described in this report (for the withdrawn fallback #296, see section 3.6)	challenge organizers	challenge terms
REAL-TSE EVAL	producing submissions only; never used for training (Track 1 used utterance-length statistics, metadata only, no audio, for length matching)	challenge organizers	challenge terms

To say our DEV data usage in one place: the REAL-TSE DEV set was used for validation, model comparison, and checkpoint selection only, in both ranked entries (#279 and #188). No DEV audio was used for training or fine-tuning of any model in those systems. The withdrawn fallback candidate #296 did use DEV-computed objectives in its fine-tuning (section 3.6), and we do not rely on it. During development we kept the training sets separate from the DEV and EVAL sessions: we checked the session ranges, we applied the exclusion list of 4,644 utterance IDs, and we verified the manifests (for example, the v3combo manifests have zero overlap with the DEV set in mixtures, enrollments, target references, speakers, and source segments).

Pretrained models:

Model	Use	Source	License
USEF-TSE / USEF-TFGridNet	warm-start base of both tracks' extractors	Zeng et al., arXiv:2409.02615; HF `ZBang/USEF-TSE`	CC BY-NC 4.0
ECAPA-TDNN	training-time speaker-cosine loss only (Track 1); not part of any deployed model	Desplanques et al., Interspeech 2020; HF `speechbrain/spkrec-ecapa-voxceleb`	Apache 2.0
DNSMOS P.835 (`sig_bak_ovr.onnx`)	training-time differentiable surrogate loss only (Track 1); not part of any deployed model	Reddy et al., ICASSP 2022; `microsoft/DNS-Challenge`	Microsoft research license (DNS Challenge)
WeSpeaker (via `wespeakerruntime`)	Track 2 per-file routing similarity estimate, computed locally (section 3.4)	Wang et al., ICASSP 2023; `wenet-e2e/wespeaker`	Apache 2.0
nvidia/RE-USE	Track 2 stage-2 enhancer, used exactly as released (no fine-tuning)	HF `nvidia/RE-USE`	research and development only

We want to emphasize one license point: the nvidia/RE-USE license is research and development only, and our Track 2 leaderboard entry depends on that model, so we state this openly here.

5. Reproducibility and distribution

We provide a public repository with the materials for verifying and reproducing our systems:

https://github.com/kts12345/real-tse-2026-chuest

The repository contains the Track 1 model code (`model_USEF_TFGridNet.py`, `TFgridnet.py`, utilities, and the block-online inference script), the run configuration (`config.yaml`: dnsmos_w 2.0, inter_block 24, num_layers 6, n_fft 128), and the

dataset and pretrained-model declaration; the Track 2 inference and routing scripts with the full per-utterance routing decisions (22 csv files, 3,210 crown / 1,790 reuse); and this paper. These files are the contents of our internally verified reproduction package (assembled and md5-verified on June 28).

The model checkpoints are attached to the repository release because of file size limits: the Track 1 weights `usef_b5_dnsmos17s_iter1000.pth.tar` (188 MB, md5 `193e32a2`, 290 tensors, 15.64 M parameters) and the Track 2 stage-1 weights `epoch16.pth.tar` (188 MB, md5 `3c330b7e`). The RE-USE weights are not redistributed in the repository because of their research-and-development-only license; the exact file we used is identified by md5 in section 3.7 and is available from `nvidia/RE-USE` on Hugging Face.

The original `#279` evaluation waveforms are preserved separately (md5 `026545b5`). If any part of this description needs more evidence (training logs, leaderboard exports, latency script output), please contact us and we will provide it.