# CARTSE Submission to the REAL-TSE Challenge Track 2: Offline Target Speaker Extraction

Li Li, Shogo Seki
CyberAgent, Inc., Japan

***Abstract*—This paper describes the CARTSE system submitted to the offline track (Track 2) of the Real-World Target Speaker Extraction (REAL-TSE) Challenge. The system is a band-split RNN (BSRNN) separator conditioned on a fine-tuned ECAPA-TDNN speaker encoder through three complementary mechanisms: a time-frequency enrolment map, frame-level context fusion, and an intermediate-block fusion (deep fusion). Training proceeds through simulated pre-training, then fine-tuning on real meetings using quality-filtered pseudo-targets. These pseudo-targets are then improved by regenerating them from the model's own output (self-training), and the model is refined through multiple fine-tuning stages with complementary objectives. The final system is an equal-weight parameter average (weight soup) of four diversely-weighted runs. On the official leaderboard our system reaches a token error rate (TER) of 0.651, a speaker similarity (SIM) of 0.544, a DNSMOS P.835 overall (OVRL) score of 3.364, and a voice-activity timing F1 of 0.857.**

***Index Terms*—target speaker extraction, band-split RNN, pseudo-labeling, data filtering, weight averaging**

## I. Introduction

Target speaker extraction (TSE) isolates the speech of an enrolled speaker from a mixture, guided by an enrolment utterance of that speaker [1]. The Real-World Target Speaker Extraction (REAL-TSE) Challenge moves TSE from simulated mixtures to genuine far-field meeting and in-the-wild recordings, where the enrolment and the mixture are frequently captured by *different devices* (microphone array, headset, or personal phone), and where the target speaker may be silent for long intervals. These conditions break the clean-reference assumptions of classical TSE training. Track 2, the offline track, allows the separator to use the full context of the mixture.

Systems are ranked by the unweighted mean of the dense ranks of four metrics: token error rate (TER), enrolment-to-output speaker-embedding cosine similarity (SIM), DNSMOS P.835 overall quality (OVRL) [6], and a voice-activity timing ratio F1 (RATIO-F1).

Our system is a band-split RNN (BSRNN) separator conditioned on a fine-tuned ECAPA-TDNN speaker encoder, with an additional intermediate-block fusion (deep fusion) that re-injects the speaker embedding inside the separator. Our contributions are three-fold: (i) pre-training on simulated mixtures and speaker-verification data; (ii) multi-stage, multi-objective fine-tuning on real recorded meetings using self-trained, quality-filtered pseudo-targets; and (iii) a final model soup, an equal-weight parameter average of four diversely-weighted fine-tuning runs.

Participant name: **CARTSE**; team name: **CARTSE** (as registered on the leaderboard).

## II. System Architecture

The system comprises two modules: a separator that extracts the target speech, and a speaker encoder that turns the enrolment $e$ into a conditioning embedding. Both follow the official challenge baseline, a band-split RNN (BSRNN) [2] separator with target-speaker conditioning and an ECAPA-TDNN [3] speaker encoder, initialized from the released baseline weights and adapted by fine-tuning.

**Separator:** The mixture $x$ is transformed by a short-time Fourier transform (STFT) with window length of 512 and hop of 128 into a real/imaginary spectrogram of shape $(2, F, T)$, where $F{=}257$ is the number of frequency bins and $T$ the number of time frames. A time-frequency enrolment map [31] is concatenated to it, and the result is band-split into 32 sub-bands (100/200/500/2000-Hz widths) with sub-band normalization, giving a feature tensor $(32, 128, T)$. The frame-level speaker embedding then modulates these features. The separator stacks 6 repeats of a band-RNN followed by a temporal RNN, emits a complex band mask, and inverts it by an inverse STFT (iSTFT) to produce the target speech estimate $\hat{s}$. The pipeline is shown in Fig. 1.

**Speaker encoder:** The conditioning embedding is produced by the baseline's ECAPA-TDNN [3] speaker encoder, which is further fine-tuned on English and Chinese datasets then adapted to the real recordings.

**Conditioning:** The enrolment conditions the separator through three paths. At the input, the time-frequency enrolment map (TF-Map) is an enrolment-magnitude $\times$ mixture-magnitude attention map concatenated to the spectrogram, making the input shape of $(3, F, T)$. Within the separator, the frame-level ECAPA embedding is linearly projected from 512 to the feature dimension 128 and applied by element-wise multiplication. The embedding is injected a second time, deeper in the separator, at block 3 of the 6 repeats. We call this second injection the intermediate-block fusion, or deep fusion.

## III. Data

Our data pipeline combines large-scale simulation with real meeting recordings. The separator is first pre-trained on *simulated* mixtures generated on-the-fly from clean speech, then fine-tuned on *real* far-field mixtures. Because these real mixtures have no clean per-speaker reference, their targets

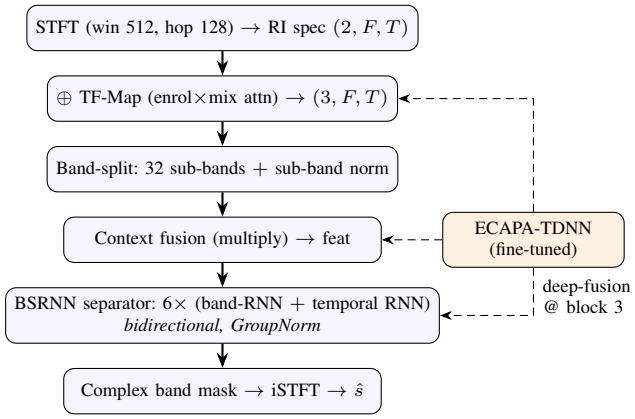


Fig. 1. CARTSE separator. Solid arrows are the signal path; dashed arrows are the three conditioning injections.

are *generated as pseudo-targets* by a teacher model and then quality-filtered. The speaker-conditioning encoder is trained separately on verification corpora and domain-adapted to the challenge recordings. This section describes each of these ingredients in turn: the simulated pre-training corpora, the speaker-encoder data, the real fine-tuning corpora and their self-training rounds, and the shared data augmentations including noise, reverberation, and channel-gap.

### A. Simulated pre-training corpora

The base separator is pre-trained with on-the-fly mixing of clean English speech: LibriSpeech train-100 [11] with the urgent2025 speech set (EARS [14], VCTK [13], WSJ [15], LibriVox/DNS5 [16], LibriTTS [12]), containing 4,240 speakers / 498,665 utterances. Mixtures draw 1/2/3 speakers of $[0.1, 0.6, 0.3]$ at $[-5, 5]$ dB, with a reverberant target and added noise, where room impulse responses (RIRs) are simulated by setting $RT_{60}$ at 0.05–0.3 s and signal-to-noise (SNR) is set at $[5, 15]$ dB. The signal lengths are set as 4 s for mixture and 3 s for enrolment with reverb/noise augmentation (0.5/0.5).

A fraction of mixtures are made *target-absent* so the separator learns to distinguish target speaker in mixtures and output silence if target speaker is absent, via two independently-drawn mechanisms: *distractor enrolment* (prob 0.35, a random absent speaker) and *noise-only* (prob 0.05, pure noise), making about $38.25\%$ of target-absent in total. Validation uses a static held-out simulated set.

### B. Speaker-encoder data

The conditioning encoder is a bilingual ECAPA-TDNN [3] trained for verification on VoxCeleb2-dev [18] (EN) and CN-Celeb1/2 train [19] (ZH), which consist of 8,987 speakers and 1,743,328 utterances in total, with RIR and 0.9/1.0/1.1 speed-perturbation augmentation ($\times 3 \rightarrow$ 26,961 classes, prob 0.6). A domain fine-tune is then conducted on real recordings using enrolment clips and target-solo no-overlap segments from the far-field sessions, mixed 1:1 with the verification data to avoid forgetting. To generate speaker embedding robust against device properties, we introduce equalizer (EQ) filtering as another data augmentation.

TABLE I
REAL PSEUDO-TARGET CORPUS ROUNDS. THE MEDIAN COLUMN IS THE PER-CLIP MEDIAN OVRL / PRECISION / TER OVER EACH ROUND'S KEPT CORPUS.

| Round | Teacher | # clips (hours) | Median |
|---|---|---|---|
| round-1 | raw GSS | 1,624 (7.2 h) | 2.58 / 0.95 / 0.18 |
| round-2 | prior submission | 7,149 (34.6 h) | 2.86 / 0.97 / 0.22 |
| round-3 | self-trained | 7,710 (37.9 h) | 3.06 / 0.97 / 0.22 |

### C. Real fine-tuning corpora and self-training

Following REAL-T [10], we draw target-speaker training data from real conversational meeting corpora (AISHELL-4 [26], AliMeeting [27], AMI [28], CHiME-6 [29]). DipCo [30] is excluded since there is no official train split. These real mixtures have no clean per-speaker reference, so targets are *generated as pseudo-targets* by a teacher TSE model and quality-filtered. For every corpus we use only its official *train* split. The dev and eval splits of each corpus are never used for training. Each training example takes the mixture and the original enrolment as input and the generated target as the signal to reconstruct; roughly 3.6 enrolments per (mixture, speaker) are sampled at random. A target is kept only if it satisfies all four criteria:

- $\Delta$OVRL $> 0$: a positive OVRL gain over the mixture;
- OVRL $\geq 2.2$: the absolute OVRL clears a quality floor;
- prec $\geq 0.80$: precision from FireRedVAD timing;
- TER $< 0.6$: token error rate from the official zipformer ASR.

The pseudo-targets are refined over three successive *rounds*, each re-generated by a stronger teacher than the last so that target quality rises from one round to the next (Table I). Round-1 targets are produced by guided source separation (GSS) [9], and round-2 by an earlier version of our own system (an early challenge submission); in both, the targets come from a *different* model than the one later trained on them. Round-3 is the *self-training* round: the same model that is trained on the data also generated its targets (see below).

The round-3 corpus is produced by *self-training*: the separator we fine-tuned on the round-2 corpus is used to re-generate the training targets, and that same model is then trained further on them. Rounds 1 and 2 instead take their targets from a different, earlier model. Because this round-3 teacher is stronger than the earlier submission that produced round-2, re-generating with it gives cleaner targets. The same checkpoint also initialises the subsequent multi-stage fine-tune (Section IV-C).

### D. Noise, RIR, and channel-gap augmentation

Three augmentation families broaden the training conditions. A large noise dataset used for all stages consists of 180,264 clips from the URGENT Challenge 2025

(DNS5+FSD50k [23]+FMA [24], WHAM! [21]), and MUSAN [20]. Moreover, an 8-environment DEMAND [25] second source (vehicle/home/café) is added as weighted auxiliary noise (prob 0.3) for out-of-domain generalization in the real data fine-tuning stage since the real mixtures are all meeting-room scenes. Reverberation uses a 40,248-wav RIR pool (SLR28 [22]+RWCP+AIR). Channel-gap augmentation models the enrolment–mixture device mismatch: a *light* form adds reverb/noise to the enrolment (each prob 0.2) in real-data fine-tuning, while a smooth, RMS-preserving EQ filter (Appendix A) is used only in the speaker-encoder fine-tune. On real data the mixture is genuine audio, so no synthetic RIR or noise-only mixtures are added and mixture noise is reduced to prob 0.4 (SNR $[5, 20]$ dB dB).

## IV. Training

Throughout, $x$ is the mixture, $\hat{s}$ the output, $s$ the target, $e$ the enrolment, $y$ the per-frame target-activity label, $s_t = (\langle \hat{s}, s\rangle / \|s\|^2)\, s$ the scaled target projection, and $\tilde{s}$ the target $s$ with its non-speech frames (those with $y$=0) set to zero.

### A. Simulated pre-training

The base separator is first pre-trained on simulated mixtures with the **zero-target masked SI-SDR loss**, split by target presence with floor $\tau$=$10^{-3}$:

$$L_{\text{pres}} = -10 \log_{10} \frac{\|s_t\|^2}{\|\hat{s} - s_t\|^2 + \tau \|s\|^2}, \tag{1}$$

$$L_{\text{abs}} = \eta \cdot 10 \log_{10} \big( \|\hat{s}\|^2 + \tau \|x\|^2 \big), \quad \eta = 2.0. \tag{2}$$

Target-present chunks maximise SI-SDR ($L_{\text{pres}}$), while target-absent chunks are driven to silence ($L_{\text{abs}}$).

### B. Fine-tuning

The pre-trained separator is then fine-tuned on the real, pseudo-labeled data. All real-data fine-tuning adds mixture noise (probability 0.4, SNR $[5, 20]$), the DEMAND auxiliary source (0.3), and light channel-gap (0.2), uses uniform language batching, and is optimised with Adam (gradient clipping 5.0); the fine-tuned ECAPA encoder is grafted into the simulated-base checkpoint beforehand. The main objective is the **scenario-aware loss**, evaluated over target-active (TA) and target-silent (TS) intervals:

$$\begin{aligned} L_{\text{TA}} &= -\alpha_1 \,\text{SI-SDR}_{\text{TA}} + w_{\text{m}}\, \text{MRSTFT}(\hat{s} \odot m_{\text{TA}}, s \odot m_{\text{TA}}), \\ L_{\text{TS}} &= \alpha_2\, 10 \log_{10}(\|\hat{s}\|^2_{\text{TS}} + \epsilon), \end{aligned} \tag{3}$$

where $m_{\text{TA}}$ is the binary target-active mask (1 on TA frames, 0 elsewhere) that restricts the magnitude term to the active regions, $\alpha_1$=0.5, $\alpha_2$=0.01, $w_{\text{m}}$=1.0, and MRSTFT a multi-resolution STFT magnitude loss (fft/win $[512, 1024, 2048]$, hop $[128, 256, 512]$). $L_{\text{TA}}$ is applied on target-active frames and $L_{\text{TS}}$ on target-silent frames, the two sets being determined by the per-frame activity labels $y$. Weighted auxiliary losses are summed on top of this objective; each is defined below and its per-stage weight is given in Table II. All auxiliary networks are frozen (no gradient update) and, unless stated otherwise, operate on the waveform; $w$ denotes the loss weight.

**speaker-consistency (wespeaker) loss:** Cosine distance between WeSpeaker ResNet34 embeddings of the output and the enrolment,

$$L = w\big(1 - \cos(\text{WeSpk}(\hat{s}), \text{WeSpk}(e))\big), \tag{4}$$

where WeSpk is a frozen WeSpeaker ResNet34 model [4], the EN or ZH model selected by the utterance language. The enrolment embedding WeSpk$(e)$ is detached, and the optional interferer-repulsion term is disabled.

**mel-filterbank loss:** $L_1$ distance between per-utterance CMVN-normalized 80-dimensional Kaldi log-mel features (25 ms window, 10 ms hop) of the output and of $\tilde{s}$,

$$L = w\, L_1\big(\text{CMVN}(\text{fbank}(\hat{s})),\ \text{CMVN}(\text{fbank}(\tilde{s}))\big). \tag{5}$$

**DNSMOS loss:** Direct maximization of the P.835 OVRL predicted by a frozen DNSMOS CNN [6], with MOS$(\hat{s})$ the predicted OVRL:

$$L = w\,(5 - \text{MOS}(\hat{s})). \tag{6}$$

**mel-pvad loss:** A frozen single-layer logistic VAD applied to the output log-mel produces a per-frame speech logit, supervised by binary cross-entropy against the reference activity labels $y$:

$$L = w\, \text{BCE}(\text{logit}, y). \tag{7}$$

**zipformer-ASR loss:** Feature-space MSE between the frozen encoder of the TER-evaluation Zipformer ASR [7] (the EN or ZH encoder per language) applied to the output and to $\tilde{s}$,

$$L = w\, \text{MSE}(\text{Zip}(\hat{s}), \text{Zip}(\tilde{s})). \tag{8}$$

**FireRedVAD feature loss:** Feature-space MSE between a frozen FireRedVAD DFSMN (CMVN filterbank input) applied to the output and to $\tilde{s}$.

**whisper-CE loss:** Teacher-forced token-level cross-entropy of the ground-truth transcript through a frozen Whisper-medium [5]: a differentiable log-mel of the output is passed through the encoder–decoder,

$$L = w\, \text{CE}(\text{Whisper}(\text{logmel}(\hat{s})), \text{transcript}), \tag{9}$$

with a per-sample ZH/EN language prefix whose tokens are masked from the loss. To save memory, we only use samples shorter than 30 s for computing the whipser-CE loss.

**soft-F1 loss:** A differentiable RATIO-F1 surrogate. The frozen FireRedVAD DFSMN produces a per-10 ms speech probability $p$ from the output; with the reference labels $y$ and soft counts TP $= \sum p\, y$, FP $= \sum p(1-y)$, FN $= \sum (1-p) y$,

$$L = w\,(1 - 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN} + \epsilon)). \tag{10}$$

TABLE II

OFFLINE TRAINING STAGES. *Init* NAMES THE STAGE WHOSE DEV-BEST CHECKPOINT INITIALIZES THE STAGE; *Sched.* IS CHUNK LENGTH / BATCH / EPOCHS / LEARNING RATE ("→" A DECAY). ALL REAL-DATA STAGES (S1–S4) USE THE SCENARIO-AWARE OBJECTIVE; THE LISTED WEIGHTS ARE THE AUXILIARY LOSSES ADDED ON TOP.

| Stage | Init | Data | Sched. | Auxiliary losses (weight) |
|---|---|---|---|---|
| S0 (sim base) | baseline | speech pool | 4 s / 12 / 200 / 1e-4 | zero-target masked SI-SDR ($\eta$=2.0) |
| S1 | S0 | round-2 | 3 s / 12 / 80 / 1e-4→1e-5 | wespeaker 1.5, dnsmos 0.1, zipformer-asr 1.0, fireredvad 0.05 |
| S2 | S1 | round-3 | 2–30 s / 2 / 80 / 1e-4→1e-5 | wespeaker 3.0, mel-fbank 3.0, dnsmos 1.0, mel-pvad 1.0, whisper-ce 1.0 |
| S3 | S2 | round-3 | 2–30 s / 2 / 40 / 3e-5→1e-6 | wespeaker 5.0, mel-fbank 1.0, dnsmos 0.5, mel-pvad 1.0, whisper-ce 2.0 |
| S4 (×4) | S3 | round-3 | 2–30 s / 2 / 80 / 5e-5→1e-5 | mel-fbank 1.0, dnsmos 0.5, soft-f1 5.0, whisper-ce + wespeaker per run |
| S5 (soup) | S4×4 | — | — | equal-weight parameter average |

### C. Training Recipe

**Staged fine-tuning.** Our system is trained as a sequence of stages; the per-stage data, schedule, and loss weights are listed in Table II. It starts with simulated pre-training under the zero-target masked SI-SDR loss, followed by a first real-data fine-tune on the round-2 corpus. That round-2 model also serves as the self-training teacher: we average its best-5 (by DEV) checkpoints into one model, use it to regenerate the round-3 targets, and initialise the next stage from it. Fine-tuning then proceeds on round-3: one stage adds the mel-pvad and whisper-CE losses, and the next continues with a heavier speaker-consistency weight. Each stage starts from the DEV-best checkpoint of the previous one.

**Model soup.** The final stage launches four parallel fine-tunes from the same checkpoint, with identical initialization, data, length, and learning rate; they differ only in the *loss weights* of two auxiliary terms, the whisper-CE and speaker-consistency (wespeaker) losses, set to $(\text{whisper-CE}, \text{wespeaker}) = (2,5), (2,8), (3,8), (2,11)$ across the four runs. Each run keeps the average of its best-3 checkpoints by DEV, and the final system is the equal-weight parameter average (a model soup [8]) of the four resulting models.

## V. RESULTS AND VALIDATION

Table III reports the official DEV and EVAL results against the two official non-causal baselines: `BSRNN_TFMAP`, from which our separator is initialized, and `BSRNN_EMB`. On both splits our system improves over both baselines on all four metrics, most markedly in perceptual quality (EVAL OVRL $1.61 \rightarrow 3.36$) and intelligibility (EVAL TER $0.84 \rightarrow 0.65$).

TABLE III

OFFLINE RESULTS ON THE OFFICIAL DEV AND EVAL (TRACK-2 LEADERBOARD) SETS. `TFMAP` AND `EMB` ARE THE OFFICIAL NON-CAUSAL BASELINES (THE FORMER IS OUR INITIALIZATION); SIM IS THE ENROLMENT-TO-OUTPUT SIMILARITY.

| Set | System | TER↓ | SIM↑ | OVRL↑ | F1↑ |
|---|---|---|---|---|---|
| DEV | `TFMAP` | 0.703 | 0.521 | 1.500 | 0.838 |
| | `EMB` | 0.693 | 0.501 | 1.660 | 0.841 |
| | CARTSE | 0.462 | 0.614 | 3.348 | 0.895 |
| EVAL | `TFMAP` | 0.838 | 0.443 | 1.612 | 0.829 |
| | `EMB` | 0.829 | 0.417 | 1.844 | 0.829 |
| | CARTSE | 0.651 | 0.544 | 3.364 | 0.857 |

## VI. CONCLUSION

The CARTSE system couples a triple-conditioned band-split RNN separator with a pseudo-labeling pipeline for unlabeled real meetings, a self-training round, a multi-objective fine-tune, and a final weight soup across diversely-weighted runs. All evaluation models used as training objectives were applied only to eligible training data; the evaluation set was processed by a single fixed system version, without per-utterance metric-driven optimization.

## APPENDIX A
## CHANNEL-GAP AUGMENTATION

In the real recordings the enrolment and mixture often come from different devices, and thus differ in frequency response, bandwidth, and recording chain. On DEV the learned mixture–enrolment similarity is the strongest TER predictor (Spearman $\rho = -0.52$) and the out-of-domain evaluation condition shows the largest enrolment–mixture gap, yet synthetic training draws both from the same chain, so the conditioning path is never trained to be device-invariant.

The augmentation passes each training enrolment through a random simulated device response (enrolment channel $\neq$ mixture channel on almost every sample): a smooth, RMS-preserving dB gain curve applied by rFFT multiply, combining a spectral tilt ($\pm 4$ dB/oct), a cubic-spline random EQ (6 log-spaced nodes in $[80, 7600]$ Hz, $\pm 6$ dB), and Butterworth band limits (low-pass $U(3.4$–$7.5$ kHz) p0.4, high-pass $U(50$–$150$ Hz) p0.5), in source–room–device order. TSE fine-tuning uses this light enrolment-side setting (reverb/noise-enroll 0.2); the speaker-encoder fine-tune uses an "extreme" preset (prob 0.7, tilt $\pm 10$ dB/oct, 8-node $\pm 12$ dB EQ, wider band limits).

## APPENDIX B
## DATA SOURCES

Per challenge rules, no official training set is defined; participants may use any open-source data and pre-trained models with documented sources. Speech corpora: LibriSpeech [11], LibriTTS [12], VCTK [13], EARS [14], WSJ [15], DNS5/LibriVox [16], VoxCeleb1/2 [17], [18], CN-Celeb1/2 [19]. Noise/RIR: MUSAN [20], WHAM! [21],

RIRS_NOISES/SLR28 [22], FSD50k [23], FMA [24], DEMAND [25]. Challenge corpora (training splits only): AISHELL-4 [26], AliMeeting [27], AMI [28], CHiME-6 [29]. Pre-trained models: ECAPA-TDNN and WeSpeaker ResNet34 [3], [4], Whisper-medium/large-v2 [5], FireRedASR-AED-L / FireRedVAD, DNSMOS P.835 [6], and zipformer-zh/en [7].

## REFERENCES


[1] K. Žmolíková *et al.*, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," IEEE J. Sel. Topics Signal Process., vol. 13, no. 4, pp. 800–814, 2019.

[2] Y. Luo and J. Yu, "Music source separation with band-split RNN," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 31, pp. 1893–1901, 2023.

[3] 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.

[4] H. Wang *et al.*, "WeSpeaker: A research and production oriented speaker embedding learning toolkit," in Proc. ICASSP, 2023.

[5] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023, pp. 28492–28518.

[6] 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.

[7] Z. Yao *et al.*, "Zipformer: A faster and better encoder for automatic speech recognition," in Proc. ICLR, 2024.

[8] M. Wortsman *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in Proc. ICML, 2022, pp. 23965–23998.

[9] C. Boeddeker *et al.*, "Front-end processing for the CHiME-5 dinner party scenario," in Proc. CHiME Workshop, 2018.

[10] 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.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in Proc. ICASSP, 2015, pp. 5206–5210.

[12] H. Zen *et al.*, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in Proc. Interspeech, 2019, pp. 1526–1530.

[13] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Univ. Edinburgh, 2019.

[14] J. Richter *et al.*, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in Proc. ICASSP, 2024.

[15] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in Proc. Workshop on Speech and Natural Language, 1992, pp. 357–362.

[16] H. Dubey *et al.*, "ICASSP 2023 deep noise suppression challenge," in Proc. ICASSP, 2023.

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

[18] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in Proc. Interspeech, 2018, pp. 1086–1090.

[19] Y. Fan *et al.*, "CN-Celeb: A challenging Chinese speaker recognition dataset," in Proc. ICASSP, 2020, pp. 7604–7608.

[20] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," arXiv:1510.08484, 2015.

[21] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in Proc. Interspeech, 2019, pp. 1368–1372.

[22] T. Ko *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in Proc. ICASSP, 2017, pp. 5220–5224.

[23] E. Fonseca *et al.*, "FSD50K: An open dataset of human-labeled sound events," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 829–852, 2022.

[24] M. Defferrard *et al.*, "FMA: A dataset for music analysis," in Proc. ISMIR, 2017, pp. 316–323.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND)," in Proc. Meetings on Acoustics, 2013.

[26] Y. Fu *et al.*, "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization" in conference scenario," in Proc. Interspeech, 2021, pp. 3665–3669.

[27] F. Yu *et al.*, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in Proc. ICASSP, 2022, pp. 6167–6171.

[28] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in Proc. MLMI, 2006, pp. 28–39.

[29] S. Watanabe *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in Proc. CHiME Workshop, 2020.

[30] M. Van Segbroeck *et al.*, "DiPCo — dinner party corpus," arXiv:1909.13447, 2019.

[31] K. Zhang *et al.*, "Multi-level speaker representation for target speaker extraction," in Proc. ICASSP, 2025, pp. 1–5.