CE-TSE: CROSS-EMOTION AND PHYSICAL ACOUSTICS FOR ROBUST TARGET SPEAKER EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Extracting the target speaker's voice from noisy mixtures environment is a core challenge for intelligent auditory systems. Target Speaker Extraction (**TSE**) aims to mimic the human "Cocktail Party Effect". When humans extract target speech in noisy environments, they rely on not only speaker identity, but also spatial acoustic cues and emotional information. However, existing TSE benchmark datasets lack both realistic acoustic conditions such as room reverberation and spatial geometry, and complete emotional annotations. Therefore, we manually annotate emotional labels for over **2,300 utterances** from the REAL-TSE dataset, consuming approximately 58 person-hours of annotation effort. Then we incorporate real room impulse responses and noise to perform physical spatial simulation to build the Cross-Emotion Target Speaker Extraction dataset (**CE-TSE Dataset**). Building upon this dataset, we propose the Cross-Emotion Target Speaker Extraction Network (**CETSE-Net**). CETSE-Net is equipped with an emotion disentanglement module (**CE**), which decouples emotional features from speaker identity representations. Experimental results show that CETSE-Net achieves consistent improvements across all evaluation metrics: +1.844 dB on SI-SNR, +1.736 dB on SDR and +0.42 on PESQ.

1 INTRODUCTION

Target Speaker Extraction (TSE) isolates a specific target voice from multi-speaker mixtures. TSE serves as a core technique for real-world applications such as teleconferencing, in-vehicle voice interaction, and hearing aid systems (Žmolíková et al., 2019). Humans can extract target speech and suppress interfering sounds in complex acoustic scenarios, which relies on speaker identity, emotional features, and spatial acoustic information. This capability is known as the Cocktail Party Effect.

Current state-of-the-art TSE models are trained on classic mixture datasets such as WSJ0-2mix and LibriMix (Hershey et al., 2016; Cosentino et al., 2020). These TSE models use reference speech only to identify the target speaker for speech extraction. But these models ignore **emotional features** and **real-world spatial acoustic information**. This limitation hinders the study of cross-emotion TSE. Cross-emotion TSE research remains limited for two main reasons. Most mainstream TSE datasets lack emotion labels. Researchers cannot analyze emotion mismatch between reference speech and target speech without these labels. Meanwhile, existing datasets with emotion labels, such as IEMOCAP (Busso et al., 2008), CREMA-D (Cao et al., 2014), and RAVDESS (Luna-Jiménez et al., 2021), only contain clean speech. These emotional datasets do not include room reverberation, spatial geometry or background noise. The emotional datasets cannot recreate real acoustic scenes and cannot test how emotion affects TSE performance in real use.Prior work confirms that emotion strongly influences TSE extraction results (Švec et al., 2022). In summary, no public TSE benchmark dataset provides both emotion labels and realistic acoustic simulation.

To address these limitations, we start with the REAL-TSE dataset. While REAL-TSE dataset features highly realistic conversational scenes, it does not provide native emotion annotations (Wang et al., 2026). So, we build a web annotation platform (Figure 4) and conduct manual emotion annotations for over **2,300** real-scene audio clips. Furthermore, we construct the CE-TSE dataset for cross-emotion states. Unlike existing benchmark datasets, the CE-TSE dataset focuses on restoring the physical sound field characteristics of real conversational scenes. We integrate multi-source audio resources including clean speech, measured room impulse responses (RIR), and diverse background noises.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

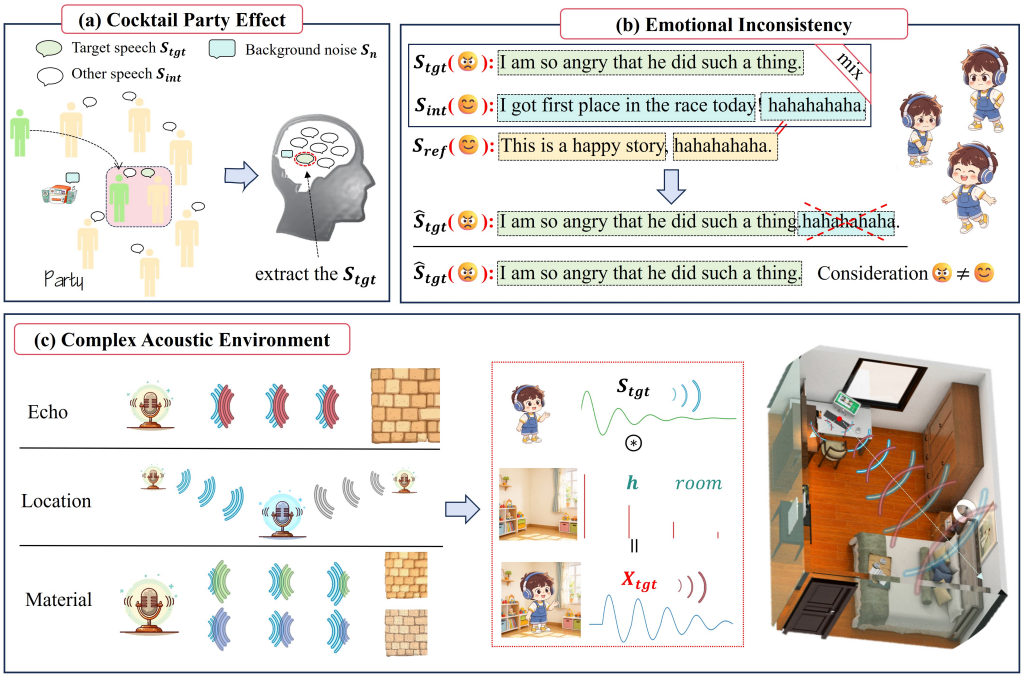

Figure 1: **Problem setup and motivations for TSE in emotional and real-world acoustic environments.**This figure illustrates the problem setup and motivations for TSE: **(a)** mimicking the Cocktail Party Effect; **(b)** revealing the interference of emotional inconsistency; and **(c)** demonstrating that complex acoustic environments require RIR convolution for simulation.

Based on the CE-TSE dataset, we further propose CETSE-Net. CETSE-Net has a dedicated emotion disentanglement module. The module explicitly separates emotional features and speaker identity features in speech and helps the model focus on identity information. This figure shows the problem setup and motivations for TSE: (a) mimicking the Cocktail Party Effect; (b) revealing the interference of emotional inconsistency; and (c) demonstrating that complex acoustic environments require RIR convolution for simulation.Based on the above analysis, the core contributions of this paper are as follows:

REAL-TSE Emotion Annotation: We conduct manual emotion annotations on the public REAL-TSE dataset. This establishes a data foundation for emotion-aware TSE research.

CE-TSE Dataset Construction: We construct the CE-TSE dataset to better reflect complex physical acoustic environments. The dataset incorporates diverse noise interferences and spatial distributions to simulate real conversational scenarios under varying acoustic conditions.

CETSE-Net: We propose a novel Emotion Disentanglement Module.This module separates emotional information from the enrollment speech. It keeps only identity-related representations. This design substantially boosts TSE separation performance.

2 RELATED WORK

TSE Datasets Existing mainstream TSE datasets are mainly simulated. Classic benchmarks including WSJ0-2mix and LibriMix generate mixtures based on clean speech, lacking realistic reverberation, environmental noise, and real emotional speaking scenarios (Hershey et al., 2016; Cosentino et al., 2020). Although the recently proposed REAL-TSE dataset adopts real conversational data (Wang et al., 2026), it does not provide fine-grained emotion annotations and cannot support emotion-aware TSE research. Overall, current public datasets fail to model real-world cross-emotion mismatch problems in TSE tasks. In contrast, our CE-TSE dataset provides fine-grained emotion labels and incorporates physical acoustic information including room reverberation and environmental noise.

Target Speaker Extraction TSE is another important branch of speech separation. It does not separate all sources. Instead, TSE recovers only the speech of a designated target speaker from the mixture. TSE is a classic task for solving the "Cocktail Party Problem" (Zmolikova et al., 2023).Early TSE methods such as VoiceFilter and VoiceFilter-Lite adopt fixed speaker embeddings as conditional information (Wang et al., 2018; 2020).SpeakerBeam, SpEx and SpEx+ design speaker-aware end-to-end TSE frameworks. They fuse reference speaker features and employ multi-scale temporal convolution networks (Žmolíková et al., 2019; Xu et al., 2020; Ge et al., 2020).Latest works continuously optimize TSE architectures. DPCCN improves robustness under noisy and reverberant conditions. X-SepFormer applies Transformer structures to reduce speaker confusion. USEF-TSE abandons fixed speaker embeddings and utilizes cross-attention for flexible reference modeling (Han et al., 2022; Liu et al., 2023; Zeng and Li, 2025).In addition, StyleTSE and FlowTSE expand TSE research paradigms by introducing text-style conditions and generative acoustic modeling (Huo et al., 2025; Navon et al., 2025). WeSep further provides a unified open-source platform for fair TSE evaluation (Wang et al., 2024). (Jin et al., 2025) introduced facial expression embeddings as perceptual cues for TSE, demonstrating that expression features provide complementary information under modality missing. (Mi et al., 2026) proposed a two-stage TSE-SER framework for emotion recognition under human speech noise. However, these works rarely consider the emotion mismatch between reference speech and target speech. Different from them, our CETSE-Net model explicitly disentangles speaker identity and emotion, and we build a corresponding dataset to support cross-emotion TSE research.

3 Task Description

3.1 TSE Problem

TSE aims to extract the speech of the target speaker from multi-speaker mixtures. It uses a reference speech from the target speaker. Formally, given the mixed speech S_{mix}, the target speech S_{tgt}, and the reference speech S_{ref}, the TSE model f_θ estimates the target speech with:

$$\hat{S}_{\mathrm{tgt}} = f_\theta(S_{\mathrm{mix}}, S_{\mathrm{ref}}) \tag{1}$$

where θ denotes the model parameters.

As shown in Figure3,a typical TSE model follows an Encoder-Separator-Decoder architecture. The Encoder first transforms the raw waveform of S_{mix} and S_{ref} into acoustic embeddings E_{mix} and E_{ref}. A Speaker Encoder then extracts the speaker identity embedding Z_{spk} from the reference embedding E_{ref}. The Separator takes E_{mix} as input and Z_{spk} as a condition. The Separator fuses the target speaker's identity into the mixture representation. The Separator then generates a mask or a filtered representation for the target speaker. Representative backbones include Conv-TasNet (Luo and Mesgarani, 2019) and X-SepFormer (Liu et al., 2023).Finally, the Decoder reconstructs the estimated target waveform S_{tgt} from the separated representation. However, most existing TSE methods do not explicitly incorporate emotion into their frameworks. These methods treat the reference speech purely as an identity cue and ignore emotional information embedded within this reference signal.

3.2 Cross-Emotion Problem in TSE

In real conversational scenarios, the reference speech and the target speech often have different emotional states. That is $e(S_{\mathrm{ref}}) \neq e(S_{\mathrm{tgt}})$,where $e()$ denotes the emotion label. This ideal assumption is frequently violated in practice. For example, the reference speech may be neutral, while the target speech is angry.This mismatch leads to extraction errors, as shown in Figure 1 (b).

This practical inconsistency motivates us to focus on cross-emotion extraction. This work targets a challenging and under-explored scenario. In this scenario, S_{ref} and S_{tgt} may carry different emotions. In this state, the emotional information in the reference features contaminates the speaker identity representation. This degrades extraction performance. We believe this is a new task. We call it Cross-Emotion Target Speaker Extraction (CE-TSE).

3.3 Evaluation Metrics

We use three mainstream metrics to evaluate the extracted speech quality: Scale-Invariant Signal-to-Noise Ratio (SI-SNR), Signal-to-Distortion Ratio (SDR), and Perceptual Evaluation of Speech Quality (PESQ).

(1) SI-SNR is defined as:

$$\mathrm{SI-SNR}(\hat{S}, S) = 10\log_{10}\frac{\|s_{\mathrm{proj}}\|_2^2}{\|\hat{S} - s_{\mathrm{proj}}\|_2^2} \qquad s_{\mathrm{proj}} = \frac{\langle \hat{S}, S\rangle}{\|S\|_2^2} S \tag{2}$$

where $\hat{S}$ and S are the zero-mean normalized estimated and target signals. We adopt SI-SNR to measure the similarity between estimated speech. SI-SNR decomposes the estimated speech into target speech component and error component. It calculates the energy ratio between these two parts. SI-SNR is scale-invariant and measures the similarity between estimated speech and ground-truth target speech. A higher SI-SNR value indicates better speech extraction performance(Luo and Mesgarani, 2018).

(2) SDR measures the ratio between the target speech energy and the interference error energy. It is a classic metric in speech separation. It is defined as:

$$\mathrm{SDR}(\hat{S}, S) = 10\log_{10}\frac{\|S\|_2^2}{\left\|\hat{S} - S\right\|_2^2} \tag{3}$$

SDR computes the energy ratio of target speech to total error energy. The error includes residual interference and waveform distortion. SDR evaluates the overall distortion of extracted speech. A higher SDR value indicates less distortion in the extracted target speech (Vincent et al., 2006). (3) PESQ is an objective speech-quality metric (Rix et al., 2001). Designed to mimic human subjective listening, it measures the perceptual difference between estimated speech and ground-truth target speech, and outputs a score ranging from 0.5 to 4.5 (often simplified to 1.0–4.5). Higher PESQ scores correspond to better perceptual quality and more natural-sounding speech.

4 Construction of the CE-TSE Dataset

To simulate complex real-world acoustic conditions, we construct a Target Speaker Extraction dataset: **the CE-TSE Dataset**. This dataset is multi-corpus, fine-grained emotion-annotated, and incorporates real physical acoustic factors. Specifically, the CE-TSE Dataset uses physical simulation techniques. These techniques simulate noisy backgrounds and reverberant acoustic conditions in practical scenarios. Figure2 shows an overview of the dataset construction.

4.1 Data Collection

Unlike the scripted acted utterances used in most existing TSE benchmarks, our dataset covers a wide range of scenarios and incorporates authentic physical environmental factors. We collect raw audio from several sources, including speech, room impulse responses (RIRs), and background noise subsets. All collected audio files are preprocessed and amplitude-normalized to unify the waveform level. Notably, REAL-TSE does not provide native emotion labels (Wang et al., 2026). To address this, we design a dedicated web-based annotation UI and manually label more than 2,300 audio samples to obtain emotion annotations.

To investigate the cross-emotion mismatch problem in target speaker extraction, we construct a dedicated CE-TSE dataset. Specifically, we adopt three widely used emotional speech corpora, including IEMOCAP (Busso et al., 2008), CREMA-D (Cao et al., 2014), and RAVDESS (Luna-Jiménez et al., 2021), to build emotion-mismatched TSE samples.

4.2 physical acoustic augmentation

To simulate target speaker extraction scenarios under real acoustic environments, we follow the mainstream synthesis paradigm in speech separation to construct mixture audio samples with reverberation,

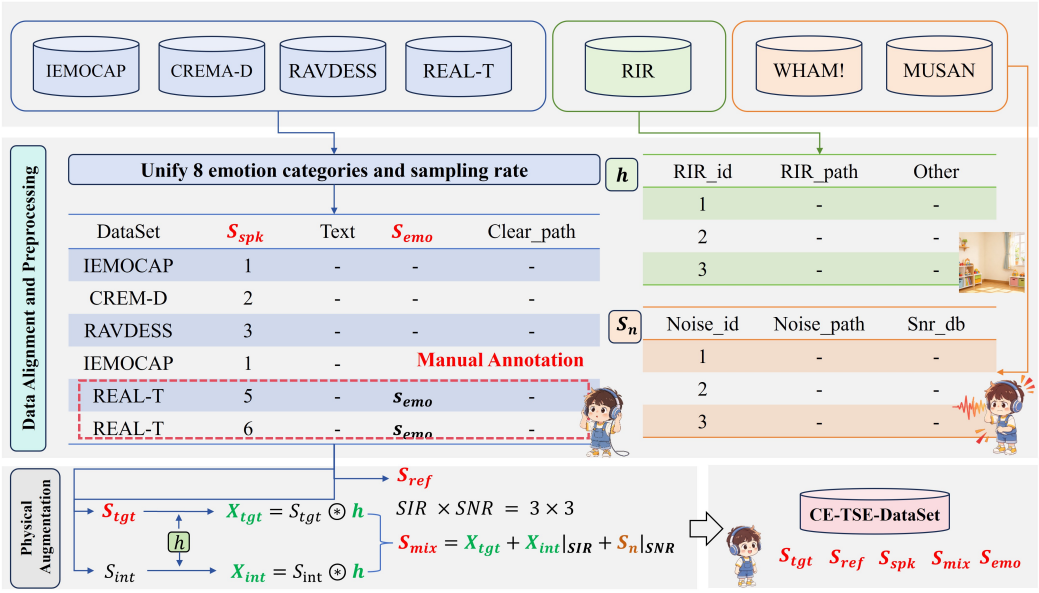

Figure 2: **The construction pipeline of the CE-TSE Dataset.** First, we unify 8 emotion categories and the sampling rate. And we also perform manual emotion annotation on the REAL-TSE dataset. This step yields speaker identities, emotion labels, and reference speech. Next, we introduce real room impulse responses and noise. We use physical simulation to generate the mixed speech with reverberation and noise. Finally, we construct the CE-TSE Dataset.

interfering speakers and background noise. Instead of shallow waveform-level augmentations such as random amplitude scaling and time shifting, our physical acoustic augmentation follows the physical law of sound propagation. It simulates the full process: sound emission from the source, spatial reflection and attenuation, and finally arrival at the microphone, introducing reverberation distortion caused by real environments during signal generation.

Our room impulse responses (RIRs) come from two sources: (1)DNS Challenge rirs noises dataset (Reddy et al., 2021): This public benchmark RIR corpus is widely adopted in speech enhancement and speaker extraction tasks. It contains both simulated and real-recorded RIRs, covering small, medium and large rooms. (2)RIR simulation via the Image Source Method (ISM) implemented in Pyroomacoustics (Scheibler et al., 2018): ISM is a classic geometric acoustic algorithm. ISM models sound reflections on room boundaries by introducing virtual image sources symmetric to the walls. The signal captured by the microphone is the superposition of contributions from the original source and all image sources. Using this method, we programmatically generate RIRs for diverse environments including car cabins, small/medium/large rooms, offices and corridors, to create reverberation conditions of varying strengths.

Reverberation is the temporal tail formed when sound waves reflect multiple times off walls and obstacles in an enclosed space. The Room Impulse Response (RIR) characterizes the time-domain signal captured by a microphone when a point source emits an impulse. The RIR describes the multipath acoustic propagation from the source to the microphone. This propagation includes the direct sound path, early wall reflections, and the late reverberation tail.

Following the image-source method , the discrete RIR $h[n]$ is expressed as:

$$h[n] = \sum_{k \in \mathcal{K}} \frac{\beta_k}{4\pi d_k} \cdot \delta\left[n - \left\lfloor F_s \frac{d_k}{c} \right\rfloor\right] \tag{4}$$

Here, $\mathcal{K}$ is the set of all image sources. d_k is the propagation distance of the k-th path. β_k is the cumulative wall reflection coefficient. The term $\frac{\beta_k}{4\pi d_k}$ gives the amplitude attenuation of this path. F_s is the sampling rate. c is the speed of sound. $\delta[\cdot]$ is the discrete Kronecker delta function. Each peak

in $h[n]$ corresponds to one propagation path. The peak position is determined by the delay. The peak height is determined by the attenuation.

Reverberant speech is obtained by convolving clean speech with the RIR:

$$X_{\text{tgt}} = S_{\text{tgt}} \circledast h \qquad X_{\text{int}} = S_{\text{int}} \circledast h \tag{5}$$

Here, S_{tgt} and S_{int} are clean target and interfering speech. X_{tgt} and X_{int} are the reverberant speech after room propagation. $\circledast$ denotes convolution. The physical meaning of convolution is that each sample of clean speech is delayed, attenuated, and summed according to the paths described by h.

The final noisy reverberant mixture is formulated as:

$$Y[n] = \big(S_{\text{tgt}} \circledast h[n]\big)[n] + \big(S_{\text{int}} \circledast h[n]\big)[n] + \nu[n] \tag{6}$$

Here, $Y[n]$ is the mixture signal received by the microphone. $\nu[n]$ is additive environmental noise sampled from the MUSAN corpus. This equation means that reverberant target speech, reverberant interfering speech, and environmental noise are superimposed. This superposition forms the final mixture audio used to train the TSE model.

4.3 Mixture Audio Synthesis

The mixture audio is synthesized in three steps: physical scene augmentation, multi-speaker mixing, and background noise addition.

Step 1: Physical scene augmentation. A room impulse response (RIR) h is randomly selected to represent the current physical environment. The clean target speech S_{tgt} and the interfering speech S_{int} are convolved with the same RIR h, ensuring that both sources are placed in the same room acoustic environment.

Step 2: Multi-speaker mixing. Given the target signal-to-interference ratio SIR (Vincent et al., 2006):$\text{SIR} = 10\log_{10}\frac{\mathbb{E}[X_{\text{tgt}}^2]}{\mathbb{E}[X_{\text{int}}^2]}$.the scaling factor α applied to the reverberant interfering speech is: $\alpha = \sqrt{\frac{\mathbb{E}[X_{\text{tgt}}^2]}{\mathbb{E}[X_{\text{int}}^2]} \cdot 10^{-\frac{\text{SIR}}{10}}}$ The reverberant target speech is then mixed with the scaled interfering speech, yielding the interference-containing signal.

Step 3: Background noise addition. Background noise X_n is randomly sampled from a noise corpus (e.g. WHAM! and MUSAN (Wichern et al., 2019; Snyder et al., 2015)). Given the target signal-to-noise ratio SNR:$\text{SNR} = 10\log_{10}\frac{\mathbb{E}[X_{\text{tgt}}^2]}{\mathbb{E}[X_n^2]}$ the scaling factor β is: $\beta = \sqrt{\frac{\mathbb{E}[X_{\text{tgt}}^2]}{\mathbb{E}[X_n^2]} \cdot 10^{-\frac{\text{SNR}}{10}}}$ The final input mixture is:$S_{\text{mix}} = X_{\text{tgt}} + \alpha\, X_{\text{int}} + \beta\, X_n$. After audio mixture generation, our dataset outperforms existing benchmarks in multiple dimensions, especially in physical acoustic augmentation and real-world acoustic scenario simulation, as summarized in Table 1. This demonstrates that our dataset not only provides fine-grained emotion annotations but also better reflects real-world environments.

5 CETSE-Net

5.1 Model Architecture

Our CETSE-Net is shown in Figure 3. This model disentangles reference speech features into latent variables of speaker identity and emotion. In the training stage, multi-task supervision regularizes the disentangled feature space. In the inference stage, we only use identity features to perform cross-emotion target speech extraction.

The **Disentanglement Encoder** consists of convolutional blocks, which map the reference features $E_{\text{ref}} \in \mathbb{R}^{C\times T}$ into two parallel latent streams: $Z_{\text{spk}} = \text{head}_{\text{spk}}(\phi(E_{\text{ref}})), Z_{\text{emo}} = \text{head}_{\text{emo}}(\phi(E_{\text{ref}}))$ where ϕ is a shared convolutional feature extractor and head_{spk} and head_{emo} are two parallel projection heads. The two latent variables share the same frame-level temporal resolution: Z_{spk},$Z_{\text{emo}} \in \mathbb{R}^{C\times T}$.

Table 1: Comparison of datasets. Our CE-TSE dataset supports bilingual recordings, spontaneous conversational mixtures, emotion annotations and physical acoustic augmentation. It also covers noise, speaker overlap and diverse SNR/SIR settings. Our dataset fills the gap for cross-emotion target speech extraction.

Dataset	IEMOCAP	CREMA-D	RAVDESS	REAL-TSE	Libri2Mix	WHAM!	WHAMR!	Ours
Noise?	×	×	×	✓	✓	✓	✓	✓
Real spontaneous conversation?	×	×	×	✓	×	×	×	✓
Bilingual (Chinese & English)?	×	×	×	✓	×	×	×	✓
Speaker overlap?	×	×	×	✓	✓	✓	✓	✓
Multiple SNR/SIR?	×	×	×	✓	✓	✓	✓	✓
Physical acoustic augmentation?	×	×	×	×	×	×	×	✓
Emotion labels?	✓	✓	✓	×	×	×	×	✓

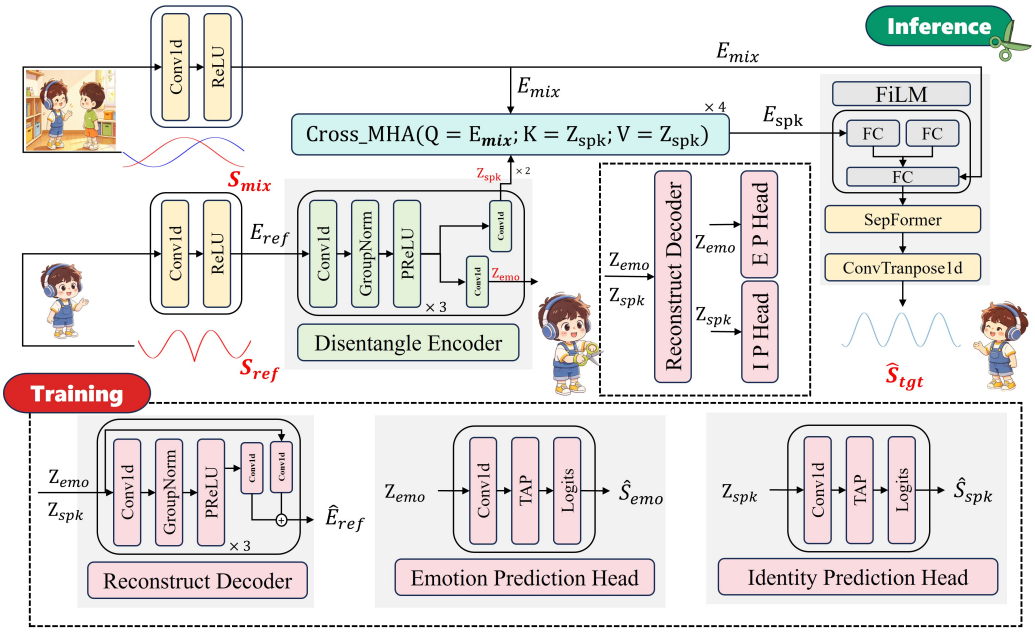

Figure 3: **The framework of our CETSE-Net model.** The training stage disentangles reference speech into speaker identity and emotion latent representations with multi-task supervision. During inference, only speaker identity features are adopted for cross-emotion target speech extraction.

The **Reconstruction Decoder** concatenates the two latent variables and reconstructs the reference features, with residual connections:$\hat{E}_{\text{ref}} = \text{out}_{\psi}([\, Z_{\text{spk}};\, Z_{\text{emo}}\,]) + \text{shortcut}([\, Z_{\text{spk}};\, Z_{\text{emo}}\,])$ where ψ denotes the parameters of the decoder, and $[\,\cdot\,;\,\cdot\,]$ denotes channel-wise concatenation. The Reconstruction Decoder works together with the Disentanglement Encoder to produce purified speaker and emotion latent representations. The joint reconstruction constraint discourages encoder shortcut learning, in which all acoustic information is compressed into a single latent stream, and preserves complete reference speech information after disentanglement. At inference time, only Z_{spk} is adopted to avoid emotion leakage from the reference speech.

5.2 Loss Function

Separation Loss: The separation loss is defined as the negative SI-SNR (minimizing it is equivalent to maximizing the SI-SNR):$\mathcal{L}_{\text{tse}} = -\,\text{SI-SNR}(S_{tgt}, \hat{S}_{tgt})$ where S_{tgt} and $\hat{S}_{tgt}$ denote the target and estimated signals, respectively.

Reconstruction Loss: The reconstruction loss constrains the output of the aggregation decoder to approximate the original reference features, ensuring that disentanglement does not discard

information:$\mathcal{L}_{\text{rec}} = \left\| E_{\text{ref}} - \hat{E}_{\text{ref}} \right\|_2^2$ where E_{ref} and $\hat{E}_{\text{ref}}$ are the reference and reconstructed features, respectively.

Classification Loss: The frame-level speaker classification head predicts the speaker identity from Z_{spk}, and the cross-entropy loss ensures that Z_{spk} retains speaker information,We adopt two cross-entropy losses for supervision.$\mathcal{L}_{\text{spk}} = -\sum_{c=1}^{C_{\text{spk}}} y_c \log p_c,\ p = \text{softmax}(f_{\text{cls}}(Z_{\text{spk}}))$ is the speaker classification loss. It constrains Z_{spk} to encode speaker identity. $\mathcal{L}_{\text{emo}} = -\sum_{c=1}^{C_{\text{emo}}} y_c \log p_c,\ p = \text{softmax}(f_{\text{cls}}(Z_{\text{emo}}))$ is the emotion classification loss. It constrains Z_{emo} to capture emotional information. where C_{spk} and C_{emo} denote the numbers of speaker and emotion classes, respectively. $y_c \in \{0, 1\}$ is the ground-truth label of class c (speaker identity for $\mathcal{L}_{\text{spk}}$ and emotion category for $\mathcal{L}_{\text{emo}}$), and p_c is the corresponding predicted probability.

Adversarial Loss: The adversarial loss prevents the emotional information Z_{emo} from being inferred from the speaker representation Z_{spk}. The discriminator is trained to predict the emotion category from Z_{spk}, and its loss is defined as: $\mathcal{L}_{\text{disc}} = -\sum_{c=1}^{C_{\text{emo}}} y_c \log e_c, e = \text{softmax}(f_{\text{disc}}(Z_{\text{spk}}))$ where y_c is the ground-truth emotion label and e_c is the predicted probability of emotion class c.

The adversarial loss, minimized by the main network to make Z_{spk} emotion-invariant, is defined as:$\mathcal{L}_{\text{adv}} = -\mathcal{L}_{\text{disc}}$.

Total Loss:The total loss of the main network is a weighted sum of four terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tse}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cls}}\left(\mathcal{L}_{\text{spk}} + \mathcal{L}_{\text{emo}}\right) + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \tag{7}$$

where λ_{rec}, λ_{cls}, and λ_{adv} are hyperparameters that balance the loss terms. Specifically, $\mathcal{L}_{\text{tse}}$ optimizes the core cross-emotion target speech extraction performance. The reconstruction loss $\mathcal{L}_{\text{rec}}$ ensures that the combined speaker and emotion latent streams retain all essential acoustic information of the reference speech after disentanglement. The classification losses $\mathcal{L}_{\text{spk}}$ and $\mathcal{L}_{\text{emo}}$ impose explicit supervision to separate speaker identity and emotional attributes into their respective latent branches. The adversarial loss $\mathcal{L}_{\text{adv}}$ further discourages emotion information from leaking into the speaker latent representation Z_{spk}.

5.3 TRAINING AND TESTING STAGE

During training, the model is trained together with the aggregation decoder and the two classification heads. At the test stage, only forward computation is performed, and neither the loss functions nor the aggregation decoder are involved. Therefore, these modules can be pruned, as shown in Figure 3. Specifically, given the mixed speech and the reference speech, the model directly outputs the estimated target speech, which is then compared with the clean target S_{tgt} using SI-SNR, SDR and PESQ as evaluation metrics.

6 EXPERIMENTS

We conduct two sets of experiments. The first studies the effect of the CE module, which disentangles the input to the backbone. The second ablates the loss weights λ_{rec}, λ_{cls}, and λ_{adv}.

6.1 EFFECT OF THE CE MODULE

Table 2 presents the objective results in cross-emotion, noisy, and reverberant scenarios. We integrate the proposed CE module into four typical TSE backbones. Experimental results show that the CE module consistently improves the extraction performance of all baseline models.The SI-SNR improvements range from 1.57 dB to 2.32 dB. These results demonstrate that emotional information significantly affects TSE performance. Furthermore, the disentanglement of speaker identity and emotion is effective for robust target speech extraction.

6.2 ABLATION STUDY ON LOSS WEIGHTS

Table 3 presents the ablation results for the three loss weights, λ_{rec}, λ_{cls}, and λ_{adv}. Using only the classification loss λ_{cls} causes performance drops, while the adversarial loss λ_{adv} alone brings obvious

Table 2: **Performance of four TSE backbones with or without the CE module.** The suffix CE represents the model equipped with the proposed cross-emotion module; red numbers denote improvements over the baseline.

Method	SI-SNR (dB)		SDR (dB)		PESQ	
	Baseline	CE ↑	Baseline	CE ↑	Baseline	CE ↑
SpEx+	16.401	18.724 +2.32	14.286	16.537 +2.25	2.71	2.82 +0.11
SpeakerBeam	14.653	16.219 +1.57	12.541	14.078 +1.54	2.48	2.75 +0.27
X-SepFormer	17.835	19.471 +1.64	15.692	17.284 +1.59	2.33	2.46 +0.13
CETSE-Net	22.312	24.156 +1.84	20.157	21.893 +1.74	2.42	2.84 +0.42

improvements over the baseline. The model reaches peak SI-SNR, SDR and PESQ when all three loss terms are adopted. Removing any single loss component leads to degraded speech extraction performance. These observations verify that explicitly introducing emotion-related constraints is necessary to realize effective cross-emotion target speech separation.

Table 3: **Ablation study of loss weights.** Red numbers denote improvements and blue numbers denote degradations over the USEF-TSE baseline.

λ_{rec}	λ_{cls}	λ_{adv}	SI-SNR (dB)	SDR (dB)	PESQ
✓	×	×	22.314 +0.002	20.157	2.42
×	✓	×	21.627 -0.685	19.483 -0.674	2.28 -0.140
×	×	✓	22.892 +0.580	20.674 +0.517	2.51 +0.090
✓	✓	×	23.251 +0.939	21.038 +0.881	2.63 +0.210
✓	×	✓	24.073 +1.761	21.756 +1.599	2.78 +0.360
×	✓	✓	23.648 +1.336	21.392 +1.235	2.71 +0.290
✓	✓	✓	**24.156 +1.844**	**21.893 +1.736**	**2.84 +0.420**

7 CONCLUSION

Emotion-Annotated REAL-TSE Dataset: The public REAL-TSE dataset features authentic real-world conversational scenarios, yet it lacks fine-grained emotion annotations, which severely limits its application in emotion-aware target speaker extraction. Therefore, we conduct comprehensive manual emotion labeling and establish a standardized emotion annotation system for REAL-TSE. This work provides credible emotional supervision resources and lays a solid data foundation for exploring cross-emotion TSE challenges.

CE-TSE Dataset Construction: Most TSE datasets fail to faithfully simulate real physical acoustic environments. we construct the CE-TSE dataset o close this simulation-to-reality gap. We employ rigorous physical acoustic simulation. The dataset covers diverse reverberation, hierarchical noise and natural multi-speaker overlap. The dataset reconstructs complex real acoustic scenarios. This method offers a viable path for audio to simulate physical environments and improves the real-world generalization of trained TSE models.

CETSE-Net: Using the CE-TSE dataset, we propose CETSE-Net. The model explicitly introduces emotion-oriented constraints and decouples emotion variations from speaker identity representations. We observe consistent gains: **+1.844 dB SI-SNR, +1.736 dB SDR, and +0.42 PESQ**. These results confirm that emotional supervision mitigates cross-emotion performance degradation. This outcome demonstrates our emotion annotation and dataset construction are valuable for robust TSE task.

AI USE STATEMENT

Large language models were used to polish sentence expression, refine wording, and check grammar in manuscript writing. The AI tools did not participate in research idea design, model architecture design, dataset construction, experiment conduction or result analysis. All text, experimental data, figures and references were carefully reviewed and verified by the authors, who take full responsibility for the entire submission.

REFERENCES

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., and Vincent, E. (2020). Librimix: An open-source dataset for generalizable speech separation.

Ge, M., Xu, C., Wang, L., Chng, E. S., Dang, J., and Li, H. (2020). Spex+: A complete time domain speaker extraction network. *arXiv preprint arXiv:2005.04686*.

Han, J., Long, Y., Burget, L., and Černockỳ, J. (2022). Dpccn: Densely-connected pyramid complex convolutional network for robust speech separation and extraction. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7292–7296. IEEE.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 31–35. IEEE.

Huo, M., Jain, A., Huynh, C. P., Kong, F., Wang, P., Liu, Z., and Bhat, V. (2025). Beyond speaker identity: Text guided target speech extraction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jin, Z., Zeng, B., Yang, P., Du, J., Ju, W., Tian, Y., Liu, J., and Li, M. (2025). Robust audio-visual target speaker extraction with emotion-aware multiple enrollment fusion. *arXiv preprint arXiv:2509.12583*.

Liu, K., Du, Z., Wan, X., and Zhou, H. (2023). X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., and Fernández-Martínez, F. (2021). Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22):7665.

Luo, Y. and Mesgarani, N. (2018). Tasnet: time-domain audio separation network for real-time, single-channel speech separation.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266.

Mi, J., Shi, X., Ma, D., He, J., Fujimura, T., and Toda, T. (2026). Robust speech emotion recognition under human speech noise. *Computer Speech & Language*, 100:101987.

Navon, A., Shamsian, A., Segal-Feldman, Y., Glazer, N., Hetz, G., and Keshet, J. (2025). Flowtse: Target speaker extraction with flow matching. *arXiv preprint arXiv:2505.14465*.

Reddy, C. K., Dubey, H., Gopal, V., Cutler, R., Braun, S., Gamper, H., Aichner, R., and Srinivasan, S. (2021). Icassp 2021 deep noise suppression challenge. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6623–6627. IEEE.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 351–355. IEEE.

Snyder, D., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469.

Wang, Q., Moreno, I. L., Saglam, M., Wilson, K., Chiao, A., Liu, R., He, Y., Li, W., Pelecanos, J., Nika, M., et al. (2020). Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition. *arXiv preprint arXiv:2009.04323*.

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., Saurous, R. A., Weiss, R. J., Jia, Y., and Moreno, I. L. (2018). Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. *arXiv preprint arXiv:1810.04826*.

Wang, S., Qian, Z., Zhang, K., Han, J., Liu, Z., Yu, X., Li, H., Delcroix, M., Yu, K., Xie, L., et al. (2026). Slt 2026 real-tse challenge: Real-world target speaker extraction from conversational recordings. *arXiv preprint arXiv:2607.15198*.

Wang, S., Zhang, K., Lin, S., Li, J., Wang, X., Ge, M., Yu, J., Qian, Y., and Li, H. (2024). Wesep: A scalable and flexible toolkit towards generalizable target speaker extraction. *arXiv preprint arXiv:2409.15799*.

Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., Manilow, E., and Roux, J. L. (2019). Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*.

Xu, C., Rao, W., Chng, E. S., and Li, H. (2020). Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM transactions on audio, speech, and language processing*, 28:1370–1384.

Zeng, B. and Li, M. (2025). Usef-tse: Universal speaker embedding free target speaker extraction. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2110–2124.

Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., and Černockỳ, J. (2019). Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814.

Zmolikova, K., Delcroix, M., Ochiai, T., Kinoshita, K., Černockỳ, J., and Yu, D. (2023). Neural target speech extraction: An overview. *IEEE Signal Processing Magazine*, 40(3):8–29.

Švec, J., Žmolíková, K., Kocour, M., Delcroix, M., Ochiai, T., Mošner, L., and Černocký, J. H. (2022). Analysis of impact of emotions on target speech extraction and speech separation. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5.

A REAL-TSE Dataset annotation

We manually annotate emotions for the REAL-TSE dataset, which originally has no emotion labels. The annotated samples provide speaker, emotion and reference speech pairs for constructing CE-TSE as shown in Figure 4.

B CE-TSE Dataset Discription

Table 4 lists all audio resources adopted in our experiments, covering emotional speech, noise and room impulse responses with their statistics and properties.We construct the CE-TSE dataset from these raw materials through physical simulation and speech mixing.

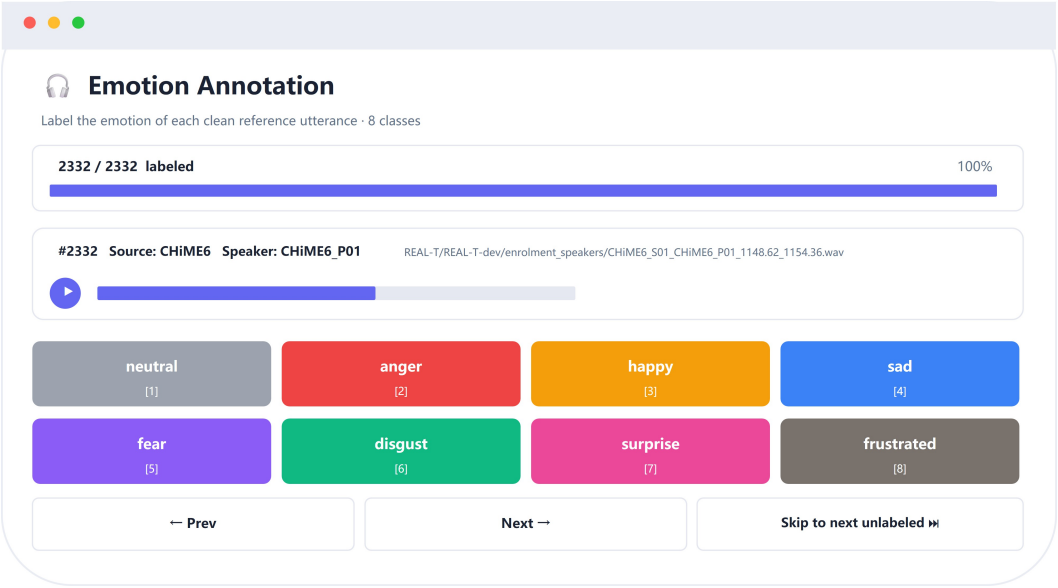

Figure 4: **Data annotation Web UI.**We build a web UI for data annotation. Annotators listen to each speech sample and assign emotion labels manually.

Table 4: **Summary of audio datasets and assets**

Category	Scale	Speaker / Source	Emotion Labels	Description
IEMOCAP	7,529	10 speakers · 5 sessions	8 classes; original 9 labels	Dyadic improvised and scripted dialogues with acted emotions and utterance-level transcription
CREMA-D	7,442	91 actors	6 classes: ANG, DIS, FEA, HAP, NEU, SAD	12 fixed sentences × 6 emotions × 4 intensity levels
RAVDESS	1,440	24 actors	8 classes: neutral, calm, happy, sad, angry, fearful, disgust, surprised	2 fixed sentences × 8 emotions × 2 intensity levels
REAL-T	2,332	190 speakers	8 classes:manual revision	Spontaneous dialogues from real meetings and dinners with natural emotions, overlaps and background sounds
WHAM!	28,000	wham noise	—	16 kHz environmental noise
MUSAN	930	free-sound 843;sound-bible 87	—	MUSAN noise subset, 16 kHz
RIR · simulated	60,000	DNS rirs-noises	—	Simulated room impulse responses, 16 kHz mono
RIR · real	417	DNS rirs-noises	—	Measured real binaural impulse responses

C HYPERPARAMETER

Table 5 summarizes the training hyperparameters adopted for all baseline models and their CE-augmented variants. We use the Adam optimizer and train all models for 100 epochs.

Table 5: **Hyperparameter settings for all models**

Model	Learning rate	Batch size	Workers	Optimizer	Epochs
USEF-TSE	1×10^{-3}	2	2	Adam	100
CETSE-Net	1×10^{-3}	2	2	Adam	100
SepFormer	7×10^{-4}	8	4	Adam	100
SepFormer (CE)	7×10^{-4}	8	4	Adam	100
SpeakerBeam	7×10^{-4}	2	4	Adam	100
SpeakerBeam (CE)	7×10^{-4}	2	4	Adam	100
X-SepFormer	1×10^{-3}	4	8	Adam	100
X-SepFormer (CE)	1×10^{-3}	32	8	Adam	100

D THEORY OF THE CE MODULE

Let S denote speaker identity; E_r the emotion of the reference enrollment utterance; E_t the emotion of the target speech. E_{ref} and E_{mix} are acoustic embeddings of reference and mixture speech, respectively. Z_{spk} denotes the disentangled speaker-related latent representation, and Z_{emo} denotes the emotion-related latent representation.

Without the CE module (original USEF-TSE, no disentanglement). The reference embedding E_{ref} encodes both speaker identity S and reference emotion E_r; the mixture embedding E_{mix} encodes S together with target emotion E_t.

$$E_{\text{ref}} = f(S, E_r),\ E_{\text{mix}} = g(S, E_t) \quad \text{sim}(E_{\text{ref}}, E_{\text{mix}}) \propto P(S \mid E_{\text{ref}}, E_{\text{mix}}) \tag{8}$$

The similarity between embeddings implicitly compares the two conditional distributions: $P(E_{\text{ref}} \mid S, E_r)$ and $P(E_{\text{mix}} \mid S, E_t)$.

Same emotion ($E_r = E_t$): When reference and target share identical emotion, the two conditional distributions match. The embedding similarity yields unbiased posterior estimation for speaker identity.

$$P(E_{\text{ref}} \mid S, E_r) = P(E_{\text{mix}} \mid S, E_t) \tag{9}$$

$$\text{sim}(E_{\text{ref}}, E_{\text{mix}}) \ \Rightarrow\ P(S \mid E_{\text{mix}}) \text{ unbiased} \tag{10}$$

Cross emotion ($E_r \neq E_t$): Under emotion mismatch, the two distributions diverge. Since E_{ref} carries unwanted emotional information from enrollment speech, similarity matching is disturbed and produces biased speaker posterior.

$$P(E_{\text{ref}} \mid S, E_r) \neq P(E_{\text{mix}} \mid S, E_t) \tag{11}$$

$$\text{sim}(E_{\text{ref}}, E_{\text{mix}}) \ \Rightarrow\ P(S \mid E_{\text{mix}}) \text{ biased} \tag{12}$$

With the CE module. The CE disentanglement encoder strips emotional components from reference features. Only the speaker-centric latent Z_{spk} is fed into cross-multi-head-attention (CMHA), which performs matching against mixture-derived speaker representation Z_{mix}.

$$\text{CMHA}: \ Z_{\text{spk}}\, Z_{\text{mix}} \longrightarrow \text{Fusion/Apart} \tag{13}$$

Now similarity operates over speaker-only distributions $P(Z_{\text{spk}} \mid S)$ and $P(Z_{\text{mix}} \mid S)$.

Same emotion ($E_r = E_t$):

$$P(Z_{\text{spk}} \mid S, E_r) = P(Z_{\text{mix}} \mid S, E_t) \tag{14}$$

$$\text{sim}(Z_{\text{spk}}, Z_{\text{mix}}) \propto P(S \mid Z_{\text{mix}}) \text{ unbiased} \tag{15}$$

Cross emotion ($E_r \neq E_t$): If disentanglement is imperfect, residual emotion leakage still shifts latent statistics, leading to representation drift and potential bias.

$$P(Z_{\text{spk}} \mid S, E_r) \neq P(Z_{\text{spk}} \mid S, E_t) \ \Rightarrow\ \text{representation drift} \tag{16}$$

$$\text{sim}(Z_{\text{spk}}, Z_{\text{mix}}) \ \Rightarrow\ P(S \mid Z_{\text{mix}}) \text{ biased} \tag{17}$$

Disentanglement objective (conditional independence). Our multi-task and adversarial losses aim to enforce conditional independence: given speaker identity S, Z_{spk} should be statistically independent of reference emotion E_r. Under this ideal condition, emotion mismatch no longer contaminates speaker matching.

$$Z_{\text{spk}} \perp E_r \mid S \iff P(Z_{\text{spk}} \mid S, E_r) = P(Z_{\text{spk}} \mid S) \tag{18}$$

$$\Rightarrow \ \text{sim}(Z_{\text{spk}}, Z_{\text{mix}}) \propto P(S \mid Z_{\text{mix}}) \tag{19}$$

In practice, the model approximates rather than strictly satisfies this independence constraint.