Detect, Attend and Extract: Keyword Guided Target Speaker Extraction

Haoyu Li[1,2,*], **Yu Xi**[2,4,*], **Yidi Jiang**[3], **Shuai Wang**[1,6,†], **Kate Knill**[4], **Mark Gales**[4], **Haizhou Li**[5,6] and **Kai Yu**[2,†]

[1]School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[2]X-LANCE Lab, MoE Key Lab of Artificial Intelligence, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[3]National University of Singapore, Singapore
[4]ALTA Institute, Machine Intelligence Lab, Department of Engineering, University of Cambridge, UK
[5]The Chinese University of Hong Kong, Shenzhen, China
[6]Shenzhen Loop Area Institute, Shenzhen, China
{haoyu.li.cs,yuxi.cs}@sjtu.edu.cn, shuaiwang@nju.edu.cn, kai.yu@sjtu.edu.cn

Abstract

Target speaker extraction (TSE) aims to extract the speech of a target speaker from mixtures containing multiple competing speakers. Conventional TSE systems predominantly rely on speaker cues, such as pre-enrolled speech, to identify and isolate the target speaker. However, in many practical scenarios, clean enrollment utterances are unavailable, limiting the applicability of existing approaches. In this work, we propose DAE-TSE, a keyword-guided TSE framework that specifies the target speaker through distinct keywords they utter. By leveraging keywords (i.e., partial transcriptions) as cues, our approach provides a flexible and practical alternative to enrollment-based TSE. DAE-TSE follows the Detect-Attend-Extract (DAE) paradigm: it first detects the presence of the given keywords, then attends to the corresponding speaker based on the keyword content, and finally extracts the target speech. Experimental results demonstrate that DAE-TSE outperforms standard TSE systems that rely on clean enrollment speech. To the best of our knowledge, this is the first study to utilize partial transcription as a cue for specifying the target speaker in TSE, offering a flexible and practical solution for real-world scenarios. Our code (https://github.com/GnafiY/DAE-TSE) and demo page (https://gnafiy.github.io/DAE-TSE_demo) are now publicly available.

1 Introduction

Target speaker extraction (TSE) aims to isolate the speech of the target speaker from a mixture of multiple speakers, also known as the cocktail party problem [Zmolikova *et al.*, 2023]. In doing so, we need to find ways to inform the system who the target speaker is by providing a reference cue. There have been studies on different reference cues, such as pre-enrolled reference speech [Yang *et al.*, 2023; Zhang *et al.*, 2025; Peng *et al.*, 2024; Liu *et al.*, 2024; Zhang *et al.*, 2024; Heo *et al.*, 2024], target-speaker lip movements [Lin *et al.*, 2023; Tao *et al.*, 2024; Wu *et al.*, 2024; Li *et al.*, 2024], acoustic spatial information [He *et al.*, 2024; Wang *et al.*, 2024b; Sun *et al.*, 2024; Pandey *et al.*, 2024], and language information [Borsdorf *et al.*, 2021; Borsdorf *et al.*, 2022; Yıldırım *et al.*, 2025].

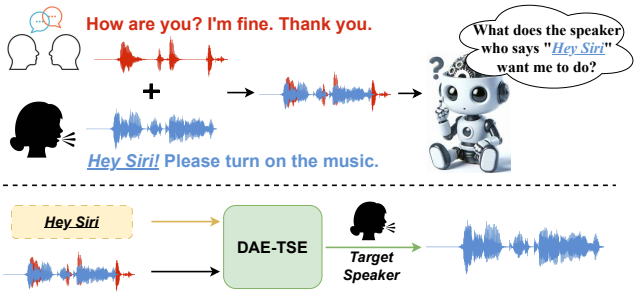


Figure 1: Illustration of the application scenario and objectives of the proposed DAE-TSE framework. In multi-talker scenarios, DAE-TSE aims to extract the speech of the target speaker who uttered the given keywords, such as "Hey Siri," from the mixture.

However, in dynamic environments such as ad-hoc meetings or voice assistant interactions, obtaining pre-registered voiceprints for all potential speakers is often unfeasible. Furthermore, relying on sustained auxiliary cues poses significant challenges; it is impractical to require a target speaker to maintain a fixed spatial location for reliable Direction of Arrival (DOA) estimation or to ensure that their lip movements remain constantly visible to a camera. In such unconstrained contexts, text-based enrollment offers a compelling solution. Prior research has explored extraction using *natural language descriptions*, known as *Language-queried Audio Source Separation (LASS)* [Liu *et al.*, 2022; Dong *et al.*, 2023; Hao *et al.*, 2026; Ma *et al.*, 2024; Shi *et al.*, 2025]. These systems isolate sources based on descriptive attributes (e.g., "female speaker", "the loudest speaker"). Conversely, *keyword-guided extraction*, which uses *partial transcription* as a ref-

[†] indicates the corresponding authors.

erence, remains underexplored despite its immense potential. As illustrated in Figure 1, a user may wish to retrieve speech from a participant based on salient keywords they mentioned (e.g., specific topics). This framework offers greater flexibility by enabling full speech extraction using only partial transcription cues, thereby eliminating the need for cumbersome pre-enrollment procedures. It is crucial to distinguish between natural language descriptions and keyword cues. LASS systems use descriptive prompts as semantic attributes. In contrast, keyword-guided TSE must identify the target speaker by aligning the acoustic content with the specific words they speak. Existing instruction-following systems, such as LLM-TSE [Hao *et al.*, 2026], treat text as a semantic instruction rather than explicitly performing keyword-guided alignment and identification.

The primary challenge in keyword-guided TSE lies in leveraging local linguistic information (keywords) to infer a global speaker representation. Unlike traditional TSE, which derives global embeddings from enrolled speech, our method must bridge content-speaker alignment to generate speaker-aware representations. To address this, we investigate the design of a partial keyword-guided TSE system that incorporates an effective cue encoder to derive target speaker embeddings conditioned on the provided transcription. Specifically, we propose the Detect-Attend-Extract (DAE) keyword-guided TSE framework, which consists of two main components: (1) a Keyword-guided Cue Encoder (KCE) that generates target speaker embeddings from the speech mixture and the keyword transcription, and (2) a TSE backbone that extracts the complete target speech. The KCE is trained via a joint Automatic Speech Recognition (ASR) and Speaker Verification (SV) objective, employing a cross-attention mechanism to align the mixture with keywords. Such a design enables DAE-TSE to determine the presence of keywords and localize their timestamps in the mixture when they exist.

Overall, DAE-TSE first detects the presence of the keywords in the mixture, then attends to the target speaker based on the keyword context, and finally extracts the target speech using the TSE backbone. Our core contributions are:

- We propose the Detect-Attend-Extract TSE (DAE-TSE) framework, which extracts target-speaker speech using only short keyword transcripts, eliminating the need for pre-enrolled speech.
- We propose a Keyword-guided Cue Encoder (KCE) under joint ASR-SV training, aligning textual and acoustic features via cross-attention to enable keyword detection, localization, and global speaker embedding extraction.
- Experimental results demonstrate that DAE-TSE surpasses competitive baselines while utilizing only 28.4% of the complete transcription, attaining a keyword localization error of $\sim$100 ms.

2 DAE-TSE: Keyword Guided Target Speaker Extraction

2.1 Target Speaker Extraction Fundamentals

Consider a mixture $\mathbf{x} \in \mathbb{R}^L$ consisting of L samples of the target speaker's waveform $\mathbf{y} \in \mathbb{R}^L$ and N interfering sources $\mathbf{n}_i \in \mathbb{R}^L$, i.e.,

$$\mathbf{x} = \mathbf{y} + \sum\nolimits_{i=1}^{N} \mathbf{n}_i, \tag{1}$$

TSE aims to recover $\mathbf{y}$ from $\mathbf{x}$. Target speaker extraction relies on a conditioning cue $\mathbf{q}$ that encodes speaker identity or content, typically formulated as a fixed vector or temporal sequence. Mainstreaming TSE systems derive an embedding $\mathbf{q}$ from a clean enrollment utterance via a pre-trained speaker encoder. Alternatively, categorical cues such as spatial layout (near vs. far-field) or gender (male vs. female) are presented as learnable embeddings. In audio-visual TSE, $\mathbf{q}$ is a latent stream extracted from lip motion or facial features.

A standard TSE system typically consists of two major components: a cue encoder and a speech extraction module. The latter further contains a mixture encoder, a fusion module, and a target decoder [Zmolikova *et al.*, 2023]. In general, the TSE model extracts the target speech $\hat{\mathbf{y}}$ from the mixture $\mathbf{x}$ conditioned on the cue $\mathbf{q}$:

$$\hat{\mathbf{y}} = \mathcal{F}_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{q}\right), \tag{2}$$

where $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$ denotes the TSE model parameterized by $\boldsymbol{\theta}$. The right panel of Figure 2 depicts the overall architecture of DAE-TSE, comprising a keyword-guided cue encoder and a speech extraction backbone.

2.2 Keyword-guided Cue Encoder for DAE-TSE

Architecture

In this section, we introduce the Keyword-guided Cue Encoder (KCE) for DAE-TSE, which leverages a few keywords to extract target speaker embeddings from the mixture. The KCE adopts a Transformer-based architecture that takes keywords and the speech mixture as dual inputs. Through cross-attention mechanisms, it aligns textual and acoustic representations to produce a compact cue embedding for TSE.

As illustrated in Figure 2, the input keywords are first converted into phoneme-based embeddings $\mathbf{PE}^{\text{Keyword}} \in \mathbb{R}^{L_{\text{kw}} \times D_{\text{kw}}}$, where L_{kw} denotes the length of the phoneme sequence, and D_{kw} is the phoneme embedding dimension. Keyword embeddings are then passed through a Transformer-based encoder to obtain the latent representations $\mathbf{LE}^{\text{Keyword}} \in \mathbb{R}^{L_{\text{kw}} \times D}$, where D is the output dimension of the Transformer. The overall operation is formulated as:

$$\mathbf{LE}^{\text{Keyword}} = \text{Transformers}(\mathbf{PE}^{\text{Keyword}}). \tag{3}$$

For the speech input, the mixture is first transformed into log-Mel filterbank (FBank) features, denoted as $\mathbf{E}^{\text{FBank}} \in \mathbb{R}^{T \times D_{\text{fb}}}$, where T is the number of frames and D_{fb} is the acoustic feature dimension. These features are combined with positional embeddings to form the initial input to the Transformer-based speech encoder, represented as $\mathbf{E}_0^{\text{Speech}}$. To incorporate keyword information, each Transformer block is augmented with a cross-attention module that fuses acoustic and textual representations. Specifically, in the i-th Transformer block, the speech features serve as the queries:

$$\mathbf{Q}_i^{\text{Speech}} = \text{Attention}(\mathbf{E}_{i-1}^{\text{Speech}}, \mathbf{E}_{i-1}^{\text{Speech}}, \mathbf{E}_{i-1}^{\text{Speech}}), \tag{4}$$

and the output of the block is computed as:

$$\mathbf{E}_i^{\text{Speech}}, \mathbf{M}_i = \text{FFN}(\text{Attention}(\mathbf{Q}_i^{\text{Speech}}, \mathbf{LE}^{\text{Keyword}}, \mathbf{LE}^{\text{Keyword}})), \tag{5}$$

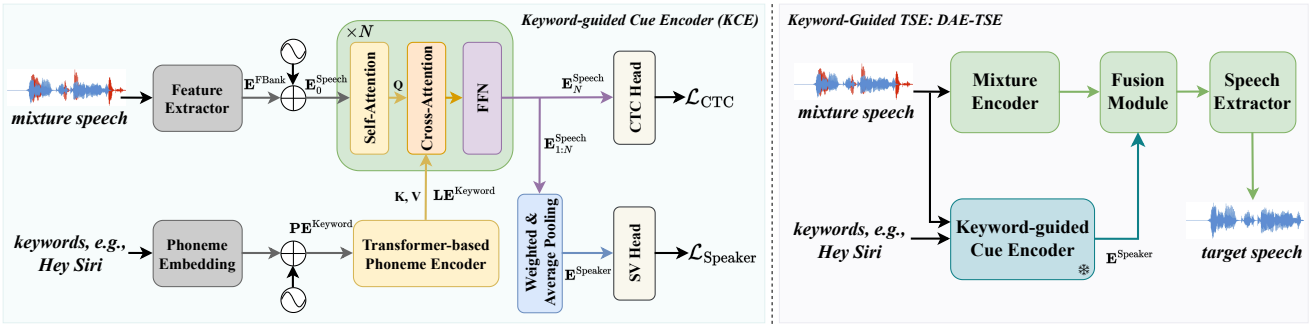


Figure 2: Overview of the proposed DAE-TSE framework. (1) The left panel depicts the architecture of the Keyword-guided Cue Encoder (KCE), which is jointly optimized via ASR and SV losses and processes the mixture speech and keywords to yield target-speaker cue embeddings. (2) The right panel shows the DAE-TSE training pipeline leveraging the pretrained KCE, which extracts the target speech from mixture inputs guided by cue embedding $\mathbf{E}^{\text{Speaker}}$.

where FFN denotes a feed-forward network, $\mathbf{M}_i$ is the attention map, and $\mathbf{LE}^{\text{Keyword}}$ represents the keyword-derived key and value embeddings used for cross-attention. After processing through all N layers of the speech encoder, we obtain the final output:

$$\mathbf{E}^{\text{Speech}} = [\mathbf{E}_1^{\text{Speech}}, \mathbf{E}_2^{\text{Speech}}, \cdots, \mathbf{E}_N^{\text{Speech}}] \in \mathbb{R}^{N\times T\times D}. \quad (6)$$

Training Criteria

To train KCE, we adopt a multi-task learning paradigm that combines Automatic Speech Recognition (ASR) and Speaker Verification (SV). Specifically, the ASR task predicts the transcription containing the keywords, while the SV predicts the speaker identity who utters the keyword.

Given the mixture speech $\mathbf{x}$, assuming the corresponding target transcription is $\mathbf{y}^{\text{Trans}}$, and the target speaker identity is $\mathbf{y}^{\text{Spk}}$, the Connectionist Temporal Classification (CTC)-based ASR loss can be represented as:

$$\mathcal{L}_{\text{CTC}}\left(\mathbf{x}, \mathbf{y}^{\text{Trans}}\right) = -\log p\left(\mathbf{y}^{\text{Trans}}|\mathbf{x}\right) \quad (7)$$

$$= -\log \sum_{\pi_{\text{CTC}}\in\mathcal{B}_{\text{CTC}}^{-1}(\mathbf{y}^{\text{Trans}})} p\left(\boldsymbol{\pi}_{\text{CTC}}|\mathbf{x}\right). \quad (8)$$

Here, $\mathcal{B}_{\text{CTC}}$ maps valid CTC alignments $\boldsymbol{\pi}_{\text{CTC}}$ (including the blank token) to the label sequence $\mathbf{y}^{\text{Trans}}$, while $\mathcal{B}_{\text{CTC}}^{-1}$ represents its inverse mapping. In contrast to ASR, which typically uses the final layer representation $\mathbf{E}_N^{\text{Speech}}$ to calculate the CTC loss, we used a weighted layer pooling to obtain the speaker information from each layer of the speech encoder dynamically. Specifically, a trainable weight vector $\mathbf{w} = [w_1, w_2, \cdots, w_N] \in \mathbb{R}^N$ is applied to fuse the keyword-biased speech representations across Transformer layers. The fused feature is computed as:

$$\mathbf{E}_{\text{sum}}^{\text{Speech}} = \sum_{i=1}^{N} \mathbf{w}_i \mathbf{E}_i^{\text{Speech}} \in \mathbb{R}^{T\times D}. \quad (9)$$

To obtain a condensed speaker representation, average pooling is applied along the time axis:

$$\mathbf{E}^{\text{Speaker}} = \text{AvgPooling}(\mathbf{E}_{\text{sum}}^{\text{Speech}}) \in \mathbb{R}^{D}. \quad (10)$$

The speaker verification loss consists of two components: (1) a Cross Entropy (CE) loss $\mathcal{L}_{\text{CE}}$ for speaker identification and (2) a regularization term $\mathcal{L}_{\text{Reg}}$ that bounds $\|\mathbf{w}\|$:

$$\mathcal{L}_{\text{Speaker}} = \mathcal{L}_{\text{CE}}\left(\text{Linear}(\mathbf{E}^{\text{Speaker}}), \mathbf{y}^{\text{Spk}}\right) + \beta \underbrace{(\|\mathbf{w}\| - 1)^2}_{\mathcal{L}_{\text{Reg}}}, \quad (11)$$

where $\mathbf{y}^{\text{Spk}}$ is the ground-truth speaker label, and $\beta = 0.01$ is a regularization coefficient.

The total loss for training KCE combines the CTC loss for the target-speaker ASR task and the CE loss for the SV task:

$$\mathcal{L}_{\text{KCE}} = \mathcal{L}_{\text{CTC}} + \alpha\mathcal{L}_{\text{Speaker}}, \quad (12)$$

where $\alpha = 0.5$ balances the ASR and SV objectives.

2.3 Speech Extraction Backbone for DAE-TSE

The speech extraction module estimates the target speech from the mixture, conditioned on the target speaker embedding from the cue encoder. Specifically, we adopt the powerful Band-Split RNN (BSRNN) [Luo and Yu, 2023] as the extraction backbone. BSRNN operates in the time-frequency domain by estimating a complex-valued spectral mask to extract the target speech from the mixture. Given the input waveform $\mathbf{x} \in \mathbb{R}^L$, the short-time Fourier transform (STFT) yields:

$$\mathbf{X} = \text{STFT}(\mathbf{x}) \in \mathbb{C}^{F\times T}, \quad (13)$$

where F and T represent the frequency bins and time frames, respectively.

BSRNN explicitly splits the mixture spectrogram into sub-bands and performs interleaved band-level and sequence-level modeling using two types of residual RNN layers. Specifically, the input spectrogram $\mathbf{X} \in \mathbb{C}^{F\times T}$ is divided into K sub-band spectrograms $\mathbf{B}_s \in \mathbb{C}^{F_s\times T}$, for $s \in \{1, 2, \ldots, K\}$, where F_s is the width of each sub-band. Independent RNNs are applied along the time and frequency dimensions in a sequential manner to model temporal and inter-band dependencies. The network then estimates a complex-valued mask $\mathbf{M} \in \mathbb{C}^{F\times T}$, which is applied to the input spectrogram via element-wise multiplication to obtain the predicted target spectrogram:

$$\hat{\mathbf{Y}} = \mathbf{X} \odot \mathbf{M} \in \mathbb{C}^{F\times T}, \quad (14)$$

where $\hat{\mathbf{Y}}$ denotes the estimated spectrogram of the target speech. Subsequently, the inverse STFT (ISTFT) is applied to reconstruct the time-domain waveform:

$$\hat{\mathbf{y}} = \text{ISTFT}(\hat{\mathbf{Y}}) \in \mathbb{R}^L, \tag{15}$$

where $\hat{\mathbf{y}}$ is the estimated target signal. We utilize the negative scale-invariant signal-to-noise ratio (SI-SNR) [Luo and Mesgarani, 2019] as the objective function to train the extraction model:

$$\mathcal{L}_{\text{SI-SNR}}(\mathbf{y}, \hat{\mathbf{y}}) = -10 \log_{10} \left(\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2} \right), \tag{16}$$

where

$$\mathbf{s}_{\text{target}} = \frac{\langle \hat{\mathbf{y}}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2} \mathbf{y}, \quad \mathbf{e}_{\text{noise}} = \hat{\mathbf{y}} - \mathbf{s}_{\text{target}}. \tag{17}$$

The BSRNN backbone was originally proposed for blind speech separation (BSS), where no information about the target speaker is provided. To support personalized tasks such as personalized speech enhancement (PSE) or TSE, previous work [Yu *et al.*, 2023] and the open-source WeSep toolkit [Wang *et al.*, 2024a] extend BSRNN by inserting a fusion module, allowing the network to condition its predictions on speaker embeddings. In this work, we adopt the BSRNN-based TSE implementation from WeSep to incorporate speaker information from the cue encoder into the backbone. As mentioned before, unlike conventional methods that rely on pre-enrolled utterances, our approach derives content-based speaker representations from the keywords of interest.

2.4 DAE Paradigm: Detect, Attend and Extract

The proposed DAE-TSE follows a three-stage detect-attend-extract formulation: the KCE first *detects* the presence of keywords; if confirmed, it *attends* to the target speaker; finally, the TSE backbone *extracts* the target speaker's speech.

1. **Detect.** DAE-TSE first detects the presence of keywords and pinpoints their temporal span through a lightweight search of the mixture. If the keyword is absent in the detection stage, the system outputs silence; otherwise, it proceeds. The cross-attention mechanism between speech and transcriptions naturally establishes frame-level correspondences between the acoustic sequence and the text-based keywords, effectively transforming detection and localization into a search problem. Leveraging this property, we develop a lightweight dynamic programming algorithm that traverses the attention matrix to efficiently detect keyword presence and localize their temporal positions in the mixture. Algorithm 1 processes the phoneme-level cross-attention map $\mathbf{M}_N \in \mathbb{R}^{L_{\text{kw}} \times T}$ from the final KCE layer (Equation 5) to compute the maximal path score S, the start and trigger frame indices (i, j) , and a detection flag d. The flag d is determined by thresholding S against a predefined threshold τ. With $O(L_{\text{kw}}T)$ complexity, the procedure efficiently handles typical keyword lengths.
2. **Attend.** If keyword presence is confirmed, the DAE-TSE encoder extracts a fixed-dimension speaker embedding $\mathbf{E}^{\text{Speaker}}$ in Equation 10 through speech-text cross-attention and pooling, as described in Section 2.2.

Algorithm 1 Detection and Localization of Keywords

Input: Cross-attention map $\mathbf{M}_N \in \mathbb{R}^{L_{\text{kw}} \times T}$, threshold τ
Output: Max path score S, start frame i, trigger frame j, detection flag d
$(K, T) \leftarrow \mathbf{M}_N$.shape, dp $\leftarrow \mathbf{0}^{L_{\text{kw}} \times T}$, prev $\leftarrow \mathbf{0}^{L_{\text{kw}} \times T \times 2}$
for $t = 0, \cdots, T-1$ **do**
  dp$[0, t] \leftarrow \mathbf{M}_N[0, t]$, prev$[0, t] \leftarrow (0, t)$
end for
for $k = 1, \cdots, K-1$ **do**
  for $t = 1, \cdots, T-1$ **do**
    if dp$[k-1, t-1] >$ dp$[k, t-1]$ **then**
      dp$[k, t] \leftarrow$ dp$[k-1, t-1] + \mathbf{M}_N[k, t]$
      prev$[k, t] \leftarrow (k-1, t-1)$
    else
      dp$[k, t] \leftarrow$ dp$[k, t-1] + \mathbf{M}_N[k, t]$
      prev$[k, t] \leftarrow (k, t-1)$
    end if
  end for
end for
$S \leftarrow \max_t$ dp$[K-1, t]$
$t \leftarrow \arg\max_t$ dp$[K-1, t]$, $k \leftarrow K-1$
while $k = K-1$ **do**
  $(k, t) \leftarrow$ prev$[k, t]$
end while
$j \leftarrow t+1$
while $k > 0 \wedge t > 0$ **do**
  $(k, t) \leftarrow$ prev$[k, t]$
end while
$i \leftarrow t$, $d \leftarrow (S \geq \tau)$
return $S, (i, j), d$

3. **Extract.** With the speaker embedding $\mathbf{E}^{\text{Speaker}}$ obtained, the TSE backbone extracts the speech of the target speaker from the mixture.

3 Experimental Setups

DAE-TSE is trained in two successive stages: the KCE is first optimized via the ASR-SV objective, after which the extraction backbone is trained while the KCE remains frozen.

3.1 Data Preparation for KCE Module

Data Simulation

LibriSpeech [Panayotov and others, 2015] is a publicly available English speech corpus consisting of 960 hours of transcribed audio and corresponding speaker labels. For cue encoder pre-training, we use simulated mixtures generated from the train-clean-360 and train-other-500 subsets, covering 2,087 speakers. The train-clean-100 subset is excluded to prevent information leakage, as it is used for mixture generation in the backbone training. To ensure data diversity, we adopt an online mixture generation strategy. Two clean utterances, $\mathbf{s}_1$ and $\mathbf{s}_2$, are randomly sampled and combined as:

$$\mathbf{s}_{\text{mix}} = \gamma_1 \mathbf{s}_1 + \gamma_2 \mathbf{s}_2, \tag{18}$$

where the scaling factors $\gamma_1, \gamma_2 \in [0.1, 0.9]$ are randomly sampled. Online simulation adopts the LibriMix [Cosentino *et al.*, 2020] max protocol: the shorter utterance is zero-padded to match the length of the longer one.

Keyword Selection

During training, we randomly crop 2-6 consecutive words from each target speaker's transcription to form the keyword cue, comprising only 6.5% to 19.5% of the full transcript (≈30.1 words). During evaluation, we restrict the reference to at most 4 successive words—covering $\leq 28\%$ of the 14.1-word average transcript. The selected words are converted to phoneme sequences with a grapheme-to-phoneme toolkit [Lee *et al.*, 2018] and fed to the cue encoder.

Evaluation Dataset Simulation

To comprehensively evaluate the keyword attending and detection capabilities, we consider both scenarios: when the keyword is present and absent in the mixture. Specifically, we construct an evaluation set consisting of 2,620 test samples, each paired with a randomly sampled fixed keyword cue. Approximately 50% of the mixtures contain keywords that do not correspond to any speaker in the mixture, simulating cases where the queried keyword is absent from the speech.

3.2 Data Preparation for Extraction Backbone

To evaluate the performance of the DAE-TSE backbone, we conduct experiments on the Libri2Mix dataset [Cosentino *et al.*, 2020], a two-speaker mixture corpus derived from LibriSpeech. The training set is simulated using the train-clean-100 subset, while the validation and test sets are generated from the dev-clean and test-clean subsets, respectively. Specifically, we adopt the fully overlapped (min) version sampled at 16 kHz.

As mentioned before, the speakers used for cue encoder pre-training are disjoint from those used in backbone training and evaluation, ensuring a fair assessment of the generalization and effectiveness of the proposed DAE-TSE system.

3.3 Training Details

The KCE module is first trained for 150 epochs (10-epoch warm-up) with Adam [Kingma and Ba, 2015] at a learning rate of 1e-3 and a batch size of 64. Input features are 80-dimensional FBank coefficients (25 ms window, 10 ms shift). For the TSE backbone, we adopt the BSRNN implemented in the WeSep framework [Wang *et al.*, 2024a], as it achieves the best overall performance. Full-length mixtures are fed to the BSRNN extractor without chunking. STFT uses a 16 kHz sampling rate, a 512-sample window, and a 128-sample hop, yielding 128 frequency bins. Subsequently, the TSE backbone is trained for 150 epochs while the cue encoder remains frozen; the learning rate decays exponentially from `1e-3` to `2.5e-5` without warm-up. All experiments run on 8 NVIDIA V100 GPUs.

3.4 Baselines

Following the open-source setup in [Wang *et al.*, 2024a] for the BSRNN-based TSE system, all extraction-based baselines adopt the same ECAPA-TDNN [Desplanques *et al.*, 2020] model pretrained on VoxCeleb2 [Chung *et al.*, 2018] in the WeSpeaker toolkit [Wang *et al.*, 2023], which is publicly available[1]. Moreover, since the proposed DAE-TSE leverages keywords as cues, we compare it not only with standard TSE systems but also with a cascaded pipeline that sequentially applies BSS and ASR. Specifically, from the perspective of the type of information utilized, such as speaker identity or content-based keywords, we consider the following baselines.

- **Audio (Standard)**. The standard TSE system utilizes a clean enrollment utterance from the target speaker to extract a speaker embedding, which serves as a conditioning vector for extracting the target speech from a multi-speaker mixture. This method requires clean enrollment data and relies solely on speaker identity, without leveraging any content-related information. It represents the most widely adopted baseline in TSE research.
- **Pipeline**. The pipeline baseline is a cascaded system consisting of a separation front-end and an ASR backend: the mixture is first decomposed into multiple waveforms, each of which is transcribed; the channel with the minimum edit distance to the enrolled keywords is selected, and its separated speech is adopted as the clean enrollment utterance. Specifically, we adopt the powerful open-source MossFormer [Zhao *et al.*, 2024], which achieves state-of-the-art performance on Libri2Mix. Separated sources are transcribed by a high-performance CTC-Transducer-based ASR model[2] from NVIDIA NeMo [Kuchaiev *et al.*, 2019] Team.

3.5 Evaluation Metrics

To comprehensively evaluate the performance of all TSE systems, we adopt a wide range of evaluation metrics from different aspects of evaluation:

- **PESQ and STOI**: Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) are full-reference metrics that estimate perceived speech quality and speech intelligibility under noisy conditions, respectively.
- **DNSMOS**: DNSMOS is a reference-free perceptual quality estimator designed for 16 kHz audio, which produces three scores in the range of 1 to 5: SIG (signal quality), BAK (background noise quality), and OVRL (overall quality). For simplicity, we report only the OVRL score.

  SI-SNR improvement (SI-SNRi): SI-SNRi measures enhancement in speech quality relative to the mixture.
- **Accuracy**: Enhancement accuracy [Zhang *et al.*, 2025] is defined as the percentage of trials whose SI-SNRi exceeds 1 dB, a threshold that designates successful target-speaker extraction.
- **SPK-SIM**: Speaker similarity is computed as the cosine similarity between speaker embeddings extracted from the enhanced and reference signals. We use a pre-trained WavLM model [Chen *et al.*, 2022] for evaluation.
- **Differential Word Error Rate (dWER)**: This metric measures the word error rate (WER) between the ASR

[1]https://wenet.org.cn/downloads?models=wespeaker&version=voxceleb_ECAPA512.zip

[2]https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_fastconformer_hybrid_large_pc

Model	Cue Info.	SI-SNRi (dB)↑	Acc. (%)↑	PESQ↑	STOI↑	DNSMOS↑	SPK_SIM↑	dWER (%)↓
TSE	Pipeline	12.98	89.93	2.84	88.72	**3.17**	0.964	20.81
	Audio	13.52	90.83	2.86	89.51	3.15	0.967	19.84
Multi-Level TSE [Zhang *et al.*, 2025]	Audio	16.08	97.18	2.86	93.71	3.15	0.978	15.14
DAE-TSE	Keywords (k=4)	**16.45**	**98.98**	**2.87**	**95.18**	3.14	**0.982**	**13.62**

Table 1: Comparison of DAE-TSE with representative TSE baselines on Libri2Mix. DAE-TSE uses consecutive keywords, which are randomly sampled from the transcription, as the enrollment cue. The *Audio* setup uses a clean enrollment utterance from the target speaker, while the *Pipeline* baseline refers to a cascaded system of a speech separation front-end and ASR backend. Best results are in **bold**.

transcription of the extracted speech and that of the ground-truth target speech. We employ the base version of Whisper for ASR decoding.

Additionally, for the keyword detection and localization evaluation of KCE, samples containing the keyword are treated as positive examples, while those without the keyword are considered negative examples. Detection performance is reported via precision (Pre.), recall (Rec.), and F1-score; localization accuracy is measured by the ℓ_1 error of the start (S Err.) and end (E Err.) timestamps.

4 Results and Analysis

4.1 Performance of Keyword Guided TSE

Table 1 presents the extraction results for several TSE baselines and various DAE-TSE configurations. Compared to the standard baselines, our DAE-TSE achieves superior performance with only 4 keywords (approximately 28.4% of the full transcription) across nearly all evaluation metrics when compared to our baseline and the powerful Multi-layer TSE. DAE-TSE demonstrates strong performance not only in signal quality metrics (SI-SNRi, PESQ, STOI) but also in acoustic and semantic consistency (Accuracy, SPK-SIM, dWER). These results highlight the effectiveness of DAE-TSE in leveraging local linguistic cues to infer global speaker and contextual information, thereby enhancing the overall extraction capability of the TSE system.

4.2 Keyword Presence Detection and Temporal Localization

DAE-TSE considers a common yet realistic scenario in which the provided keywords may not appear in the target speech. Consequently, DAE-TSE first decides whether the keywords are present in the mixture. If detection is positive, DAE-TSE further attends to and extracts the target speaker; otherwise, it simply outputs silence. When multiple speakers utter the keyword, DAE-TSE randomly selects one as the target. As shown in Table 2, the proposed detection algorithm effectively identifies keyword presence and achieves strong detection performance. Additionally, the localization errors are minimal, demonstrating that the model can accurately determine the temporal positions of the keywords. This capability significantly enhances the practicality and versatility of the proposed DAE-TSE framework.

τ	Detection Metrics (%)			Time Error (ms)	
	Pre.	Rec.	F1	S Err.	E Err.
0.48	**99.00**	93.85	96.36	103.7	100.6
0.36	98.64	97.24	97.93	104.0	100.7
0.33	98.49	97.63	**98.06**	**103.7**	100.4
0.30	97.79	97.71	97.75	103.6	**100.3**
0.23	96.88	**98.11**	97.49	104.3	101.0

Table 2: Keyword presence detection and localization results under different thresholds. τ denotes the threshold. Pre., Rec., and F1 denote Precision, Recall, and F1-score, respectively. S Err. and E Err. represent the absolute time shift error of the start and end point when the keyword is present in the mixture. Best results are in **bold**.

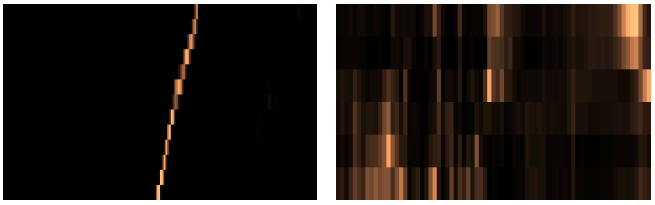


Figure 3: Cross-attention heatmaps for a positive sample (left, keywords present) and a negative sample (right, keywords absent). The horizontal axis denotes speech frame indices, while the vertical axis represents the keyword phoneme sequence. Lighter colors represent stronger alignment between speech frames and keyword phonemes.

We further visualize the cross-attention map from the last layer of KCE in Figure 3. When the keyword is present in the mixture, the attention map exhibits a clear continuous highlight, indicating a strong alignment pattern between the mixture frames and the keyword sequence. In contrast, when the keyword is absent, the attention map appears scattered without a consistent alignment. These observations demonstrate that the cue encoder effectively captures the correlation between the mixture and the provided keywords, thereby bridging keyword-to-speaker alignment. This property enables our framework to perform keyword presence detection and temporal localization.

4.3 Ablation Studies

Training Objectives for Cue Encoder

Table 3 quantifies the impact of each loss of KCE pretraining on the TSE performance. Omitting the regularization term $\mathcal{L}_{\text{Reg}}$ from the speaker loss $\mathcal{L}_{\text{Speaker}}$ incurs minor degradation; completely removing $\mathcal{L}_{\text{Speaker}}$ produces a marked drop in SI-SNRi. These ablations confirm that speaker-oriented losses are indispensable for compelling the encoder to capture reliable speaker signatures. Moreover, removing the ASR loss $\mathcal{L}_{\text{CTC}}$ from keyword-guided cue encoder training precipitates

a pronounced performance drop, evidencing that the original target-speaker embedding encodes indispensable keyword context. This semantic content is as vital as the identity itself for accurate extraction. Collectively, the ablations underscore the core strength of KCE: a speaker encoder that captures the context of keywords through alignment between mixture speech and keyword enrollment.

Model	SI-SNRi (dB)	Acc. (%)
DAE-TSE	**16.45**	**98.98**
w/o $\mathcal{L}_{\text{Reg}}$	15.97	98.02
w/o $\mathcal{L}_{\text{Speaker}}$	14.72	96.13
w/o $\mathcal{L}_{\text{CTC}}$	3.47	70.15

Table 3: Ablation results of removing different components from the training objective of the cue encoder. Specifically, $\mathcal{L}_{\text{Reg}}$ denotes the regularization term in the speaker verification loss, $\mathcal{L}_{\text{Speaker}}$ represents the speaker prediction loss, and $\mathcal{L}_{\text{CTC}}$ corresponds to the keyword-aware ASR objective.

Keyword Lengths for Cue Encoder

As shown in Table 4, we investigate how the number of keywords used as cues influences separation performance. The results show a clear monotonic improvement as the number of keywords increases. Even with just a single keyword, the performance remains strong and surpasses that of the standard TSE system reported in Table 1. Longer keyword sequences provide richer contextual information, further improving extraction quality. Notably, when using approximately 4 keywords, the performance gap compared to using the full transcription becomes negligible. These consistent ablation results confirm that the speaker embedding captures rich global contextual information and that KCE is capable of extracting this information effectively using only a few keywords.

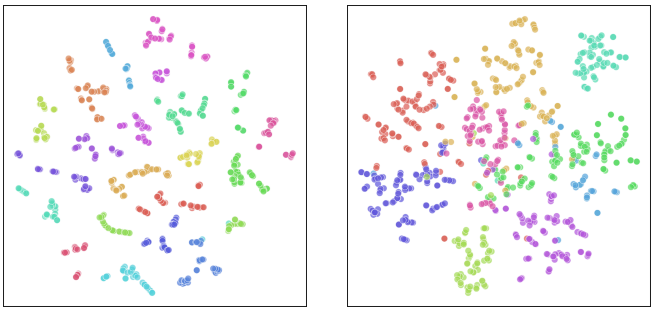
#Keywords	SI-SNRi (dB)	Acc. (%)
1	15.34	96.33
2	15.74	97.18
3	16.37	98.70
4	16.45	**98.98**
Full Trans.	**16.48**	98.78

Table 4: Results for different keyword lengths (#Keywords) used as enrollment cues. Additionally, we evaluate the DAE-TSE system with full transcription enrollment (Full Trans.), where each transcription contains an average of 14.1 words.

4.4 Visualization of Cue Embeddings

In this section, we visualize the distribution of cue embeddings using t-SNE [Maaten and Hinton, 2008]. We analyze two scenarios: in Figure 4(a), points with the same color originate from the same mixture using different keyword cues; in Figure 4(b), points with the same color are extracted from different mixtures.

In both cases, each cluster generally aligns with a single speaker, indicating the speaker-discriminative nature of the cue embeddings. The key distinction lies in contextual consistency: in (a), the embeddings are drawn from the same mixture and thus share a consistent acoustic and linguistic context, whereas in (b), the embeddings lack shared context because they are derived from different mixtures.



(a) (b)

Figure 4: t-SNE scatter plot of speaker embeddings, with each color representing a target speaker. (a) Embeddings in the same color are extracted from the same mixture across different keywords. (b) Embeddings in the same color are extracted from different mixtures.

In Figure 4(a), we observe well-separated clusters with large inter-cluster distances and tight intra-cluster cohesion. This suggests that both the identity of the speaker and the consistent contextual information contribute significantly to the quality of the cue embeddings and the effectiveness of the representation of the target speaker. In contrast, Figure 4(b) shows that even in the absence of shared context, where embeddings are extracted from different mixtures, clusters still largely align with speaker identity. However, some overlap near cluster boundaries indicates that the lack of contextual consistency may introduce ambiguity. These results demonstrate that local keyword-based cues can effectively capture global contextual information, thereby supporting robust target speaker extraction.

5 Conclusion

In this work, we propose DAE-TSE, a novel keyword-guided target speaker extraction framework that leverages partial transcriptions as reference cues. DAE-TSE establishes a three-stage paradigm for keyword-guided TSE: detecting the presence of keywords, attending to the target speaker for representation, and extracting the target speech. The design of DAE-TSE not only enables target speaker extraction but also provides accurate keyword presence detection and temporal localization capabilities via the cross-attention interaction between the mixture and the keyword cues. Extensive experiments on Libri2Mix demonstrate that DAE-TSE outperforms standard TSE baselines that rely on pre-enrolled speech. To the best of our knowledge, this work is the first to explore the use of partial transcriptions to specify the target speaker in TSE, paving the way for more flexible and real-world speech extraction systems. To advance the framework toward practical application, future efforts will focus on evaluating its robustness in challenging environments, such as scenarios with more speakers, sparse overlap, and recordings with background noise and reverberation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62401377) and the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (Grant No. 2024CSJGG1100). This work was also supported by the National Natural Science Foundation of China (Grant No. 92370206).

Contribution Statement

Haoyu Li* and Yu Xi* contributed equally to this work.

References

[Borsdorf *et al.*, 2021] Marvin Borsdorf, Haizhou Li, and Tanja Schultz. Target Language Extraction at Multilingual Cocktail Parties. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 717–724, 2021.

[Borsdorf *et al.*, 2022] Marvin Borsdorf, Kevin Scheck, Haizhou Li, and Tanja Schultz. Experts Versus All-Rounders: Target Language Extraction for Multiple Target Languages. In *Proc. IEEE ICASSP*, pages 846–850, 2022.

[Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[Chung *et al.*, 2018] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Proc. ISCA Interspeech*, pages 1086–1090, 2018.

[Cosentino *et al.*, 2020] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. LibriMix: An Open-Source Dataset for Generalizable Speech Separation. *arXiv preprint arXiv:2005.11262*, 2020.

[Desplanques *et al.*, 2020] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. ISCA Interspeech*, pages 3830–3834, 2020.

[Dong *et al.*, 2023] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian J. McAuley, and Taylor Berg-Kirkpatrick. CLIPSep: Learning Text-Queried Sound Separation with Noisy Unlabeled Videos. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[Hao *et al.*, 2026] Xiang Hao, Jibin Wu, Jianwei Yu, Chenglin Xu, and Kay Chen Tan. Typing to Listen at the Cocktail Party: Text-Guided Target Speaker Extraction. *IEEE Transactions on Cognitive and Developmental Systems*, 18(2):361–372, 2026.

[He *et al.*, 2024] Shulin He, Jinjiang Liu, Hao Li, Yang Yang, Fei Chen, and Xueliang Zhang. 3S-TSE: Efficient Three-Stage Target Speaker Extraction for Real-Time and Low-Resource Applications. In *Proc. IEEE ICASSP*, pages 421–425, 2024.

[Heo *et al.*, 2024] Woon-Haeng Heo, Joongyu Maeng, Yoseb Kang, and Namhyun Cho. Centroid Estimation with Transformer-Based Speaker Embedder for Robust Target Speaker Extraction. In *Proc. ISCA Interspeech*, pages 4333–4337, 2024.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Kuchaiev *et al.*, 2019] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. NeMo: A Toolkit for Building AI Applications Using Neural Modules. *arXiv preprint arXiv:1909.09577*, 2019.

[Lee *et al.*, 2018] Younggun Lee, Suwon Shon, and Taesu Kim. Learning Pronunciation From a Foreign Language in Speech Synthesis Networks. *arXiv preprint arXiv:1811.09364*, 2018.

[Li *et al.*, 2024] Junjie Li, Ruijie Tao, Zexu Pan, Meng Ge, Shuai Wang, and Haizhou Li. Audio-Visual Active Speaker Extraction for Sparsely Overlapped Multi-Talker Speech. In *Proc. IEEE ICASSP*, pages 10666–10670, 2024.

[Lin *et al.*, 2023] Jiuxin Lin, Xinyu Cai, Heinrich Dinkel, Jun Chen, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Zhiyong Wu, Yujun Wang, and Helen Meng. AV-Sepformer: Cross-Attention Sepformer for Audio-Visual Target Speaker Extraction. In *Proc. IEEE ICASSP*, pages 1–5, 2023.

[Liu *et al.*, 2022] Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Separate What You Describe: Language-Queried Audio Source Separation. In *Proc. ISCA Interspeech*, pages 1801–1805, 2022.

[Liu *et al.*, 2024] Yun Liu, Xuechen Liu, Xiaoxiao Miao, and Junichi Yamagishi. Target Speaker Extraction with Curriculum Learning. In *Proc. ISCA Interspeech*, pages 4348–4352, 2024.

[Luo and Mesgarani, 2019] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. ASLP.*, pages 1256–1266, 2019.

[Luo and Yu, 2023] Yi Luo and Jianwei Yu. Music Source Separation With Band-Split RNN. *IEEE/ACM Trans. ASLP.*, 31:1893–1901, 2023.

[Ma *et al.*, 2024] Hao Ma, Zhiyuan Peng, Xu Li, Mingjie Shao, Xixin Wu, and Ju Liu. CLAPSep: Leveraging

Contrastive Pre-Trained Model for Multi-Modal Query-Conditioned Target Sound Extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data Using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Panayotov and others, 2015] Vassil Panayotov et al. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *Proc. IEEE ICASSP*, pages 5206–5210, 2015.

[Pandey *et al.*, 2024] Ashutosh Pandey, Sanha Lee, Juan Azcarreta, Daniel Wong, and Buye Xu. All Neural Low-latency Directional Speech Extraction. In *Proc. ISCA Interspeech*, pages 4328–4332, 2024.

[Peng *et al.*, 2024] Junyi Peng, Marc Delcroix, Tsubasa Ochiai, Oldřich Plchot, Shoko Araki, and Jan Černocký. Target Speech Extraction with Pre-Trained Self-Supervised Learning Models. In *Proc. IEEE ICASSP*, pages 10421–10425, 2024.

[Shi *et al.*, 2025] Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. SAM Audio: Segment Anything in Audio. *arXiv preprint arXiv:2512.18099*, 2025.

[Sun *et al.*, 2024] Tianchi Sun, Tong Lei, Xu Zhang, Yuxiang Hu, Changbao Zhu, and Jing Lu. A Lightweight Hybrid Multi-Channel Speech Extraction System with Directional Voice Activity Detection. In *Proc. IEEE ICASSP*, pages 1486–1490, 2024.

[Tao *et al.*, 2024] Ruijie Tao, Xinyuan Qian, Yidi Jiang, Junjie Li, Jiadong Wang, and Haizhou Li. Audio-Visual Target Speaker Extraction with Reverse Selective Auditory Attention. *arXiv preprint arXiv:2404.18501*, 2024.

[Wang *et al.*, 2023] Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit. In *Proc. IEEE ICASSP*, pages 1–5, 2023.

[Wang *et al.*, 2024a] Shuai Wang, Ke Zhang, Shaoxiong Lin, Junjie Li, Xuefei Wang, Meng Ge, Jianwei Yu, Yanmin Qian, and Haizhou Li. WeSep: A Scalable and Flexible Toolkit Towards Generalizable Target Speaker Extraction. In *Proc. ISCA Interspeech*, 2024.

[Wang *et al.*, 2024b] Yichi Wang, Jie Zhang, Shihao Chen, Weitai Zhang, Zhongyi Ye, Xinyuan Zhou, and Lirong Dai. A Study of Multichannel Spatiotemporal Features and Knowledge Distillation on Robust Target Speaker Extraction. In *Proc. IEEE ICASSP*, pages 431–435, 2024.

[Wu *et al.*, 2024] Tianci Wu, Shulin He, Jiahui Pan, Haifeng Huang, Zhijian Mo, and Xueliang Zhang. Unified Audio Visual Cues for Target Speaker Extraction. In *Proc. ISCA Interspeech*, pages 4343–4347, 2024.

[Yang *et al.*, 2023] Lei Yang, Wei Liu, Lufen Tan, Jaemo Yang, and Han-Gil Moon. Target Speaker Extraction with Ultra-Short Reference Speech by VE-VE Framework. In *Proc. IEEE ICASSP*, pages 1–5, 2023.

[Yıldırım *et al.*, 2025] Mehmet Sinan Yıldırım, Ruijie Tao, Wupeng Wang, Junyi Ao, and Haizhou Li. Leveraging Language Information for Target Language Extraction. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 837–842, 2025.

[Yu *et al.*, 2023] Jianwei Yu, Hangting Chen, Yi Luo, Rongzhi Gu, Weihua Li, and Chao Weng. TSpeech-AI System Description to the 5th Deep Noise Suppression (DNS) Challenge. In *Proc. IEEE ICASSP*, pages 1–2, 2023.

[Zhang *et al.*, 2024] Yiru Zhang, Linyu Yao, and Qun Yang. OR-TSE: An Overlap-Robust Speaker Encoder for Target Speech Extraction. In *Proc. ISCA Interspeech*, pages 587–591, 2024.

[Zhang *et al.*, 2025] Ke Zhang, Junjie Li, Shuai Wang, Yangjie Wei, Yi Wang, Yannan Wang, and Haizhou Li. Multi-level Speaker Representation for Target Speaker Extraction. In *Proc. IEEE ICASSP*, pages 1–5, 2025.

[Zhao *et al.*, 2024] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation. In *Proc. IEEE ICASSP*, pages 10356–10360, 2024.

[Zmolikova *et al.*, 2023] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu. Neural Target Speech Extraction: An Overview. *IEEE Signal Processing Magazine*, pages 8–29, 2023.