Universal Speaker Embedding Free Target Speaker Extraction and Personal Voice Activity Detection

Bang Zeng[a,b], Ming Li[a,b,*]

[a]*School of Computer Science, Wuhan University, Wuhan, China*
[b]*Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Digital Innovation Research Center, Duke Kunshan University, Kunshan, China*

Abstract

Determining "who spoke what and when" remains challenging in real-world applications. In typical scenarios, Speaker Diarization (SD) is employed to address the problem of "who spoke when," while Target Speaker Extraction (TSE) or Target Speaker Automatic Speech Recognition (TSASR) techniques are utilized to resolve the issue of "who spoke what." Although some works have achieved promising results by combining SD and TSE systems, inconsistencies remain between SD and TSE regarding both output inconsistency and scenario mismatch. To address these limitations, we propose a Universal Speaker Embedding Free Target Speaker Extraction and Personal Voice Activity Detection (USEF-TP) model that jointly performs TSE and Personal Voice Activity Detection (PVAD). USEF-TP leverages frame-level features obtained through a cross-attention mechanism as speaker-related features instead of using speaker embeddings as in traditional approaches. Additionally, a multi-task learning algorithm with a scenario-aware differentiated loss function is applied to ensure robust performance across various levels of speaker overlap. The experimental results show that our proposed USEF-TP model achieves superior performance in TSE and PVAD tasks on the LibriMix and SparseLibriMix datasets. The results on the CALLHOME dataset demonstrate the competitive performance of our model on real recordings.

Keywords: Speaker Diarization, Target Speaker Extraction, Personal Voice Activity Detection

1. Introduction

In real-world situations, people can easily recognize when the speaker of interest is talking and accurately understand what they are saying. Researchers have categorized this auditory ability into two tasks: Speaker Diarization (SD) [1–5] and Blind Speech Separation (BSS) [6–11]. SD involves determining the speech activities of multiple speakers from the input audio, answering the question of "who spoke when". In contrast, BSS focuses on separating each speaker's voice from a mixture of audio signals. As crucial front-end processes in speech processing, SD and BSS are widely applied in various real-world speech applications, including speaker verification [12–14] and speech recognition [15–17].

*Corresponding Author: Ming Li
Email addresses: bangzeng@whu.edu.cn (Bang Zeng), ming.li369@dukekunshan.edu.cn (Ming Li)

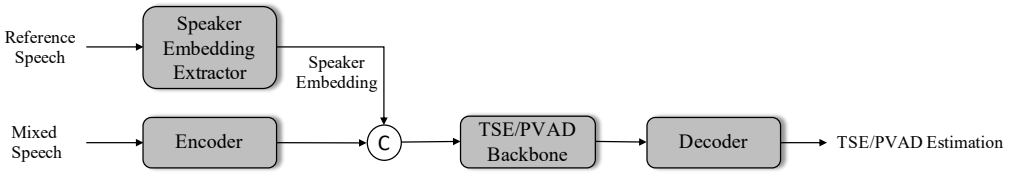

Figure 1: The diagram of a typical target speaker extraction or personal voice activity detection methods. The speaker embedding extractor is typically a pre-trained speaker recognition model. 'C' denotes the concatenation.

Many approaches for SD and BSS have been proposed, contributing significantly to advancements in both areas. However, these methods are generally applied to each task separately, failing to address the challenge of determining "who spoke what and when." Due to the similarities between SD and BSS tasks, some researchers have recently proposed joint SD and BSS methods [18, 19]. Additionally, due to the need for prior knowledge of the number of speakers in speech separation methods for practical use, some studies have focused on constructing joint models for SD and Target Speaker Extraction (TSE) [20]. In our previous work, we have shown that cascading an SD model in front of a TSE model improves the performance of TSE [21]. However, there are some mismatches between SD and TSE, which are categorized as 'output inconsistency' and 'scenario mismatch' in [20]. 'Output inconsistency' refers to the fact that an SD model outputs the speech activity boundaries for all speakers in the mixed audio. In contrast, a TSE model outputs the clean speech estimation of a specific speaker. 'Scenario mismatch' refers to the fact that SD is mainly applied in scenarios with less speaker overlap, whereas TSE focuses more on scenarios with a high degree of speaker overlap. Universal Speaker Extraction and Diarization (USED) [20] addresses output inconsistency and scenario mismatch through the embedding assignment module, supporting speech mixtures with a variable number of speakers and arbitrary overlap ratios. It leverages speaker diarization results to optimize speaker extraction while enhancing speaker diarization accuracy through the extracted speech. However, the USED approach may not be well-suited for scenarios where the number of speakers is smaller than the predefined maximum.

In practical applications, many scenarios only require the speech activity boundaries of a specific speaker rather than those of all speakers in the mixed audio. Personal voice activity detection (PVAD) [22–24] is a technique to determine the target speaker's activity speech segments in multi-speaker scenarios. PVAD may be better suited than SD techniques for constructing joint models with TSE methods. One reason is that the outputs of PVAD and TSE models are the speech activity boundaries and clean speech predictions for a specific speaker, respectively, thus avoiding the issue of 'output inconsistency.' Additionally, PVAD and TSE methods do not require prior knowledge of the number of speakers in the mixed audio, ensuring consistency in the prerequisites. At last, PVAD results enhance the TSE model's ability to suppress interfering speakers when the target speaker is silent.

As shown in Figure 1, typical PVAD and TSE methods first extract the target speaker's embedding from existing reference speech using a pre-trained speaker recognition [14, 25, 26] model. Subsequently, with the assistance of speaker embedding, the PVAD and TSE systems can model the target speaker's components within the mixed audio, producing the final result. However, careful consideration is required when selecting a suitable speaker recognition model for the TSE and PVAD tasks. Moreover, these methods may not fully exploit the information contained in the reference speech. Consequently, relying solely on speaker embeddings for the

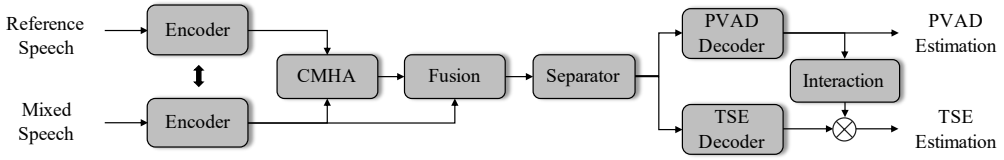

Figure 2: The diagram of the USEF-TP model. 'CMHA' denotes the cross multi-head attention. ⊗ is an operation for element-wise product.

TSE and PVAD tasks may not be optimal.

The speaker embedding-free framework is a viable new solution for the above problem. Our previous work, USEF-TSE [27], demonstrates the effectiveness of a speaker embedding-free framework for the TSE task, achieving state-of-the-art (SOTA) performance on the WSJ0-2mix [28] dataset, a widely used benchmark for monaural speech separation and TSE. In another previous work [24], we show the effectiveness of a speaker embedding-free framework in the PVAD task, achieving remarkably high recall performance.

However, USEF-TSE [27] has only demonstrated its effectiveness in high-overlapping speech scenarios, and its performance in low-overlapping situations has yet to be validated. In [24], the model's performance was only evaluated in short-duration speech scenarios, and its effectiveness in long-duration speech scenarios with varying overlapping ratios was not tested. Moreover, both studies were conducted independently, focusing on the TSE and PVAD tasks in isolation.

To address the aforementioned issues, this paper proposes a Universal Speaker Embedding Free Target Speaker Extraction and Personal Voice Activity Detection model, referred to as USEF-TP. USEF-TP uses the same structure as USEF-TSE [27]. The system framework of USEF-TP is shown in Figure 2. In USEF-TP, a unified encoder processes both the reference and mixed speech signals. The encoded mixed speech features query the reference speech encoding via a cross multi-head attention module. This module outputs frame-level features that preserve the same temporal length as the encoded mixed speech features. Subsequently, USEF-TP utilizes the frame-level output as the target speaker's characteristics and integrates it with the acoustic features of the multi-speaker input speech. After being processed by the backbone network, the fused features are passed through two deconvolution layers to generate the estimated complex spectrogram of the target speaker's speech and the PVAD prediction values. Finally, we refine the estimated complex spectrogram using the PVAD results through an interaction module, thereby obtaining the final predicted TSE results. Although the interaction module and scene-aware loss have been validated in USED [20], they are specifically designed based on the joint PVAD and TSE model, aiming further to enhance the joint modeling between PVAD and TSE. This approach differs from the SD and TSE joint modeling in USED [20]. More importantly, in contrast to USED [20], USEF-TP adopts a speaker-embedding-free framework, eliminating the model's reliance on speaker embeddings and improving its performance on the TSE and PVAD tasks.

This paper is an extension of our previous work [24], and the main new contributions of this work can be summarized as follows:

- We introduce a Universal Speaker Embedding Free Target Speaker Extraction and Personal Voice Activity Detection (USEF-TP) model. The USEF-TP model performs target speaker extraction (TSE) and joint personal voice activity detection (PVAD) with the speaker-embedding-free framework. USEF-TP addresses a unique research problem, i.e., 'the target speaker spoke what and when.'

- Inspired by [20], we designed an interaction module to refine the estimated complex spectrogram based on the PVAD results.
- We propose a multi-task learning algorithm with scenario-aware differentiated loss to optimize USEF-TP. The multi-task learning algorithm ensures temporal alignment and consistency between the PVAD and TSE results. The scenario-aware loss function can smooth the model's performance across different levels of overlap, ensuring consistent effectiveness in both low and high-overlap scenarios.
- Our proposed USEF-TP surpasses the performance of competitive baseline systems for both TSE and PVAD on the LibriMix [29] and SparseLibriMix [29]. The results show that the USEF-TP model achieves superior performance in both highly and sparsely overlapped speech scenarios for the TSE and PVAD tasks. Moreover, the results on the CALLHOME dataset also demonstrate the effectiveness of the USEF-TP model on real recordings.

The rest of the paper is organized as follows. Section 2 discusses the related work. The problem formulation is introduced in Section 3. Section 4 describes the proposed method. We introduce the experimental setup in Section 5. Section 6 presents the experimental results and analysis. Section 7 concludes the study.

2. Related Work

2.1. Speaker Diarization and Personal Voice Activity Detection

Early approaches to Speaker Diarization (SD) [30–32] involve modular systems that use voice activity detection, speech segmentation, and clustering to determine speaker turns. These methods rely on hand-crafted features and traditional machine learning algorithms. However, they often struggle with overlapping speech and require careful tuning for each pipeline stage. The introduction of deep learning has revolutionized SD methods. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, have been used to model the sequential nature of speech, leading to improved performance in detecting speaker changes. End-to-end Neural Diarization (EEND) [1, 2, 33] systems and Target Speaker Voice Activity Detection (TS-VAD) [4, 34] further simplified the process by treating SD as a multi-label classification problem. EEND has become an effective method for consolidating the multiple stages of SD into a single, unified model. These systems are designed to predict speaker activity segments from raw audio directly. By jointly optimizing all stages of the SD process, EEND systems offer improved performance over traditional pipeline-based approaches. TS-VAD is a specialized VAD technique aimed at identifying the speech activity of a group of speakers within a multi-speaker environment. This approach is particularly valuable in applications like meeting transcription, where the number of speakers are high and the duration is long. Recent methods have leveraged sequence-to-sequence prediction models to enhance the efficiency of TS-VAD, particularly in challenging acoustic environments [35].

Unlike SD, Personal Voice Activity Detection (PVAD) [22–24] models enable the detection of speech from a specific speaker while ignoring other speakers or background noise within a multi-speaker environment. This method benefits personal assistants and smart home devices, where the target speaker is known as a priori. PVAD has been approached by using speaker embeddings, such as I-vectors [36] or X-vectors [37], to detect the voice activity of a known speaker in a recording. However, the speaker embedding extraction process can be sensitive to

noise and other environmental factors, leading to false positives in detecting the target speaker. Recent work has introduced the speaker embedding free PVAD framework, eliminating the need for an additional speaker recognition model to extract speaker embeddings for a specific individual [24].

Despite the progress, challenges remain in handling highly overlapping speech for joint SD and PVAD. It remains an area where further improvement is needed, especially in real applications.

2.2. Speech Separation and Target Speaker Extraction

Blind Speech Separation (BSS) involves separating the voice of all sources. Traditional BSS approaches, such as Non-negative Matrix Factorization (NMF) [38, 39] and Computational Auditory Scene Analysis (CASA) [40, 41], typically use spectro-temporal masking to isolate each speaker's component from mixed speech. More recent methods have leveraged deep learning to improve separation quality. Deep neural network-based speech separation algorithms, such as Deep Clustering (DC) [28, 42], Deep Attractor Network (DANet) [43, 44], and Permutation Invariant Training (PIT) [45], have significantly improved the performance of speech separation tasks. However, traditional time-frequency domain methods for BSS face a notable limitation: the inability to accurately reconstruct the phase of clean speech. The introduction of time-domain BSS networks (TasNet) [46] addressed this issue. Several time-domain methods, such as Dual-Path RNN (DPRNN) [8], SepFormer [9], Mossformer [11, 47] further enhance the speech separation model's performance. The time-frequency domain model TF-GridNet [10] have achieved remarkable progress. Despite these advancements, many BSS techniques still require prior knowledge of the number of speakers in the mixture, which is not always available in real-world applications. This problem poses significant challenges for deploying these BSS solutions in practical settings.

Target Speaker Extraction (TSE) [48–52] can extract a specific speaker's voice from the mixed audio by utilizing an auxiliary reference speech of the target speaker. A typical TSE model relays on the target speaker embedding from a pre-trained [53, 54] or joint-learned [55–57] speaker embedding extractor. Because embedding extractors can be fine-tuned or adapted to new domains to improve performance, the multi-task joint training approaches adopted in [55–57] have significantly enhanced the model's performance. However, the aim of training this extractor is usually to maximize speaker recognition performance. Moreover, these methods may only partially utilize some of the information in the reference speech. Recent developments have investigated embedding-free methods that utilize frame-level acoustic features from reference speech, bypassing the need for speaker embeddings. The VE-VE framework [58] employs an RNN-based voice extractor that captures speaker characteristics using RNN states instead of speaker embeddings. This approach effectively handles the feature fusion challenge, though it is constrained to RNN-based extraction networks. SEF-Net [59], CIENet [60] and USEF-TSE [27] employ attention mechanisms to interact with the encoder of both the reference and mixed signals. These methods can guide more effective extraction by leveraging contextual information directly.

While these approaches have shown promising results, they are typically optimized for scenarios with heavily overlapped speech, limiting their effectiveness in less overlapping conditions.

2.3. Joint Speech Separation or Target Speaker Extraction with Speaker Diarization

Given the alignment in objectives between the TSE (or BSS) and SD tasks, recent studies have combined these tasks for multi-task training. EEND-SS [18] introduces a framework that

jointly performs SD, BSS, and speaker counting for a flexible number of speakers. It proposes a multiple 1D convolutional layer architecture for estimating separation masks and a fusion technique to refine separated speech signals using SD information. However, this fusion technique is only performed in the inference stage. USED [20] introduces a unified approach designed to address the challenges of TSE and SD in multi-talker scenarios. The model employs an embedding assignment module to manage outputs for a flexible number of speakers and a multi-task interaction module to leverage complementary information from both tasks. The USED [20] model represents a significant advancement in the field, offering a comprehensive solution for "who spoke what and when." USED incorporates an embedding assignment module that dynamically adjusts the number of speaker outputs based on detected speakers, reducing output inconsistencies. However, it still relies on a preset maximum number of speakers during training. If the number of speakers in an evaluation scenario exceeds this limit, the model's ability to assign embeddings accurately and separate speakers effectively may be affected.

3. Problem Formulation

In real-world speech scenarios, a conversation may involve multiple speakers' voices along with background noise and reverberation:

$$\boldsymbol{m} = \sum_{i=1}^{M} \boldsymbol{x}_i + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{m} \in \mathbb{R}^{1\times T_m}$ denotes the mixed speech, T_m is the length of the mixed speech. $\boldsymbol{x}_i$ denotes the speech of the i^{th} speaker. $\boldsymbol{n}$ denotes the additive noise. M is the number of speakers in the mixed speech.

The objective of the USEF-TP model is to detect the speech activity segments of the target speaker and extract the target speaker's voice from the mixed audio:

$$\hat{\boldsymbol{x}}_{tgt}, \hat{\boldsymbol{p}}_{tgt} = \mathscr{M}(\boldsymbol{m}, \boldsymbol{r}) = \mathscr{M}(\boldsymbol{x}_{tgt} + \sum_{i=1}^{M-1} \boldsymbol{x}_i + \boldsymbol{n}, \boldsymbol{r}) \tag{2}$$

where $\hat{\boldsymbol{x}}_{tgt}$ denotes the extracted speech of the target speaker. $\hat{\boldsymbol{p}}_{tgt}$ denotes the VAD precision of the target speaker. $\mathscr{M}(\cdot)$ denotes operations in the USEF-TP model. $\boldsymbol{x}_{tgt}$ represents the speech component of the target speaker within $\boldsymbol{m}$. $\boldsymbol{r} \in \mathbb{R}^{1\times T_r}$ represents the reference speech, T_r is the length of the reference speech.

It is worth noting that the expected output of the model $\hat{\boldsymbol{x}}_{tgt}$ will vary depending on whether the target speaker is present or absent in the speech clip. In [61], speech clips are classified into four classes: QQ, QS, SS, and SQ. In the QQ scenario, neither the target nor the interfering speakers are active. The QS scenario represents a case where the target speaker is silent while the interfering speakers are active. Conversely, the SQ scenario occurs when the target speaker is speaking, and the interfering speakers remain silent. Lastly, the SS scenario involves both the target and interfering speakers being active simultaneously. Nevertheless, this study focuses on tasks related to only one target speaker. As a result, we simplify the previous four classifications into two categories: Target-speaker Active (TA) and Target-speaker Silence (TS). Figure 3 illustrates an example of different scene clips from a mixed audio recording. In the TA scenario, the target speaker is active, encompassing both the SS and SQ scenarios. The expected TSE output

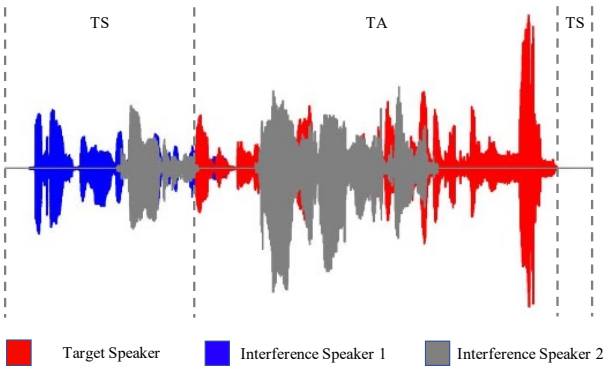

Figure 3: The diagram of different scene clips from a mixed audio recording. 'TA' denotes Target epaker Active. 'TS' denotes Target speaker Silence.

of the model is to accurately restore a prediction that closely aligns with the target speaker's speech components. Conversely, in the TS scenario, the target speaker is quiet, which includes the QS and QQ scenarios. In this scenario, the expected output of TSE is 0, indicating no speech activity from the target speaker:

$$\mathscr{M}(\boldsymbol{m}, \boldsymbol{r})_{tse} = \begin{cases} \hat{\boldsymbol{x}}_{tgt} \rightarrow \boldsymbol{x}_{tgt}, SC \in TA \\ \hat{\boldsymbol{x}}_{tgt} \rightarrow 0, SC \in TS \end{cases} \tag{3}$$

where 'SC' denotes the speech clip. $\mathscr{M}(\cdot)_{tse}$ denotes the TSE output of the USEF-TP model.

4. Methods

The architecture of the USEF-TP model is illustrated in Figure 4. Our proposed USEF-TP model consists of seven modules: the encoder, Cross Multi-Head Attention (CMHA) module, fusion module, separator, TSE decoder, PVAD decoder, and interaction module. This section will provide a detailed description of the USEF-TP model.

4.1. Encoder

Motivated by TF-GridNet [10], USEF-TP uses two weight-sharing convolutional encoders to process the mixed speech $\boldsymbol{m} \in \mathbb{R}^{B\times 1\times T_m}$ and the reference speech $\boldsymbol{r} \in \mathbb{R}^{B\times 1\times T_r}$. T_m and T_r denote the length of the mixed and reference speech, respectively. Similar to USEF-TSE [27], This encoder can be a Short-Time Fourier transform (STFT) or a one-dimensional convolution, depending on whether the model works in the time or time-frequency (T-F) domain. In this work, the USEF-TP model operates in the T-F domain:

$$\boldsymbol{m}_{RI} = STFT(\boldsymbol{m}) \tag{4}$$

$$\boldsymbol{r}_{RI} = STFT(\boldsymbol{r}) \tag{5}$$

where $STFT(\cdot)$ refers to STFT operation. $\boldsymbol{m}_{RI} \in \mathbb{R}^{B\times 2\times F\times L_m}$ and $\boldsymbol{r}_{RI} \in \mathbb{R}^{B\times 2\times F\times L_r}$ represent the stacked real and imaginary components of the STFT features for the mixed and reference speech,

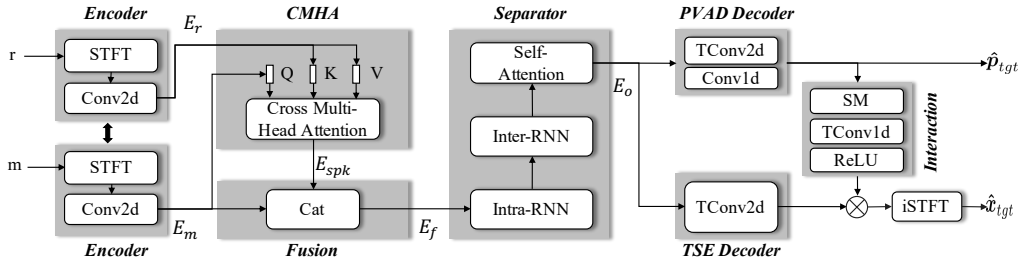

Figure 4: The diagram of USEF-TP model. 'm' and 'r' denote the mixed speech and reference speech, respectively. We use two weight sharing encoder to process the mixed and reference speech separately. $\otimes$ is an operation for element-wise product. The Separator's parameters are set identically to those of the TF-GridNet approach.

respectively. B denotes the batch size, F is the feature dimension for each T-F unit. L_m and L_r is the number of frames of $\boldsymbol{m}_{RI}$ and $\boldsymbol{r}_{RI}$, respectively.

$$\boldsymbol{E}_m = Conv2d(\boldsymbol{m}_{RI}) \tag{6}$$

$$\boldsymbol{E}_r = Conv2d(\boldsymbol{r}_{RI}) \tag{7}$$

where $Conv2d(\cdot)$ refers to 2D convolutions. $\boldsymbol{E}_m \in \mathbb{R}^{B\times C\times F\times L_m}$ and $\boldsymbol{E}_r \in \mathbb{R}^{B\times C\times F\times L_r}$ denotes the encoder results for the mixed and reference speech, respectively. The purpose of $Conv2d(\cdot)$ is to transform the number of feature channels from the original 2 to C.

4.2. CMHA module

The encoder outputs of $\boldsymbol{m}$ and $\boldsymbol{r}$ are passed into the CMHA module, which employs a cross multi-head attention mechanism:

$$\boldsymbol{E}_{spk} = CMHA(q = \boldsymbol{E}_m; k, v = \boldsymbol{E}_r) \tag{8}$$

where $\boldsymbol{E}_m \in \mathbb{R}^{B\times C\times F\times L_m}$ and $\boldsymbol{E}_r \in \mathbb{R}^{B\times C\times F\times L_r}$ represent the encoder outputs of the mixed speech and reference speech, respectively. $\boldsymbol{E}_{spk} \in \mathbb{R}^{B\times C\times F\times L_m}$ represents the output of the CMHA module. CMHA denotes the Cross Multi-Head Attention operation. To align with the separator, the USEF-TP model uses the attention module from TF-GridNet [10] as its CMHA module.

The CMHA module of the USEF-TP model uses the mixed speech encoding $\boldsymbol{E}_m$ as the query and the reference speech encoding $\boldsymbol{E}_r$ as the key and value. Through the Equation (8), the CMHA module produces a frame-level feature $\boldsymbol{E}_{spk}$, which with the same length as $\boldsymbol{E}_m$. In this way, the USEF-TP model allows the lengths of the mixed speech input L_m and the reference speech input L_r to be different.

Unlike the traditional approach shown in Figure 1, The USEF-TP model does not rely on an additional pre-trained speaker recognition model to extract the target speaker's embedding. Instead, it uses the output of the CMHA module as the target speaker's attributes. Compared to fixed-dimensional speaker embeddings, this frame-level feature $\boldsymbol{E}_{spk}$ not only retains target speaker information but also incorporates context-dependent information, which facilitates improved TSE or PVAD performance [27].

4.3. Fusion Module

After the CMHA module, the USEF-TP model uses a fusion module to combine $\boldsymbol{E}_{spk}$ and the mixed speech encoding $\boldsymbol{E}_m$:

$$\boldsymbol{E}_f = F(\boldsymbol{E}_m, \boldsymbol{E}_{spk}) \tag{9}$$

where $\boldsymbol{E}_f$ denotes the output of the fusion module. $F(\cdot)$ denotes the feature fusion operation.

There are generally two approaches to feature fusion. One is to directly concatenate features along the channel dimension, resulting in a fused feature $\boldsymbol{E}_f \in \mathbb{R}^{B\times 2C\times F\times L_m}$. The other approach is to use Feature-wise Linear Modulation (FiLM) [62]:

$$\boldsymbol{E}_f = FiLM(\boldsymbol{E}_m, \boldsymbol{E}_{spk}) \tag{10}$$

$$FiLM(\boldsymbol{E}_m, \boldsymbol{E}_{spk}) = \gamma(\boldsymbol{E}_{spk}) \cdot \boldsymbol{E}_m + \beta(\boldsymbol{E}_{spk}) \tag{11}$$

Here, $\boldsymbol{E}_f \in \mathbb{R}^{B\times C\times F\times L_m}$ denotes the fused features. The scaling and shifting vectors in FiLM are denoted by $\gamma(\cdot)$ and $\beta(\cdot)$, respectively.

In the USEF-TP model, we use concatenation as the feature fusion method. The fused features $\boldsymbol{E}_f \in \mathbb{R}^{B\times 2C\times F\times L_m}$, are then passed into the separator.

4.4. Separator

The USEF-TP model uses TF-GridNet as the backbone of the separator. The separator comprises three modules: the full-band, the sub-band, and the cross-frame self-attention module. The separator first applies zero-padding and unfold operation to the feature $\boldsymbol{E}_f$:

$$\boldsymbol{E}'_f = P(\boldsymbol{E}_f) \tag{12}$$

$$\boldsymbol{U}_{full} = U(RS(\boldsymbol{E}'_f)) \tag{13}$$

where $\boldsymbol{E}'_f \in \mathbb{R}^{B\times 2C\times F'\times L'_m}$ represents the zero-padded version of $\boldsymbol{E}_f$. $P(\cdot)$ denotes the zero-padding operation. The reshape operation, denoted as $RS(\cdot)$, reshapes $\boldsymbol{E}'_f$ into the shape $\mathbb{R}^{(B\times L'_m)\times F'\times 2C}$. Here, $F' = \lceil \frac{F-ks}{hs} \rceil \times hs + ks$. The function $U(\cdot)$ corresponds to the torch.unfold function, with $\boldsymbol{U}_{full} \in \mathbb{R}^{(B\times L'_m)\times(\frac{F'-ks}{hs}+1)\times(2C\times ks)}$ as its output. The parameters ks, hs are the kernel size and stride of the torch.unfold function, respectively.

After the unfolding process, the features pass through the intra-frame full-band module and the sub-band module sequentially, similar to TF-GridNet [10], generating $\boldsymbol{E}_{full}$ and $\boldsymbol{E}_{sub} \in \mathbb{R}^{B\times 2C\times F'\times L'_m}$. The output of the sub-band module $\boldsymbol{E}_{sub}$ is subsequently fed into the attention module. Before this, the zero-padding in the sub-band results must be removed to restore the length to that of the original input:

$$\boldsymbol{E}_{cross} = Cross(\boldsymbol{E}_{sub}[:, :, F, L_m]) \tag{14}$$

$$\boldsymbol{E}_o = \boldsymbol{E}_{cross} + \boldsymbol{E}_{sub}[:, :, F, L_m] \tag{15}$$

where $Cross(\cdot)$ refers to the operations in the cross-frame self-attention module, which applies the multi-head attention on the input. In this way, each T-F unit can attend to other relevant units. $\boldsymbol{E}_{Cross}$, $\boldsymbol{E}_o \in \mathbb{R}^{B\times 2C\times F\times L_m}$ are the outputs of the cross-frame self-attention module and the separator, respectively.

4.5. TSE and PVAD Decoder

4.5.1. TSE Decoder

In the TSE Decoder of the USEF-TP model, a 2D transposed convolution layer is applied to adjust the number of channels in $\boldsymbol{E}_o$ back to 2:

$$\boldsymbol{D}_{tse} = TConv2d_{tse}(\boldsymbol{E}_o) \tag{16}$$

where $\boldsymbol{D}_{tse} \in \mathbb{R}^{B\times 2\times F\times L_m}$ denotes the output of the TSE Decoder. $TConv2d_{tse}(\cdot)$ denotes the 2D transposed convolution layer in the TSE Decoder.

4.5.2. PVAD Decoder

In the PVAD Decoder of the USEF-TP model, a 2D transposed convolution layer is applied to adjust the number of channels in $\boldsymbol{E}_o$ back to 1:

$$\boldsymbol{D}_{pvad} = TConv2d_{pvad}(\boldsymbol{E}_o) \tag{17}$$

where $\boldsymbol{D}_{pvad} \in \mathbb{R}^{B\times 1\times F\times L_m}$ denotes the output of the 2D transposed convolution layer. $TConv2d_{pvad}(\cdot)$ denotes the 2D transposed convolution layer in the PVAD decoder. Then, the PVAD decoder applies a 1D convolution layer transforms the feature dimension of $\boldsymbol{D}_{pvad}$ to 1, while also adjusting its length:

$$\hat{\boldsymbol{p}}_{tgt} = Conv1d_{pvad}(\boldsymbol{D}_{pvad}) \tag{18}$$

where $\hat{\boldsymbol{p}}_{tgt} \in \mathbb{R}^{B\times 1\times L_{pvad}}$ denotes the estimation of the PVAD. L_{pvad} denotes the length of the PVAD labels. $Conv1d_{pvad}(\cdot)$ denotes the 1D convolution layer in the PVAD decoder.

4.6. Interaction Module

Since the training objectives of the TSE and PVAD tasks are highly aligned, the results of these tasks can enhance each other. We can leverage this consistency to refine the results of TSE. Inspired by the USED [20] model, we added an Interaction module to the USEF-TP model. The goal of the Interaction module is to feed the results of PVAD back into the TSE predictions:

$$\hat{\boldsymbol{p}}'_{tgt} = ReLU(TConv1d_{IM}(SM(\hat{\boldsymbol{p}}_{tgt}))) \tag{19}$$

$$\hat{\boldsymbol{x}}_{tgt} = iSTFT(\boldsymbol{D}_{tse} \cdot [\hat{\boldsymbol{p}}'_{tgt}, \hat{\boldsymbol{p}}'_{tgt}]) \tag{20}$$

where $\hat{\boldsymbol{p}}'_{tgt} \in \mathbb{R}^{B\times 1\times L_m}$ denotes the output of the interaction module. $ReLu(\cdot)$ and $SM(\cdot)$ denote the Rectified Linear Unit (ReLU) and softmax function, respectively. $TConv1d_{IM}(\cdot)$ denotes the 1D transposed convolution layer in the interaction module. The purpose of the 1D transposed convolution is to up-sample $\hat{\boldsymbol{p}}_{tgt}$, so that the length of $\hat{\boldsymbol{p}}'_{tgt}$ matches the length of $\boldsymbol{D}_{tse}$. Next, we repeat $\hat{\boldsymbol{p}}'_{tgt}$ along the channel dimension and get $[\hat{\boldsymbol{p}}'_{tgt}, \hat{\boldsymbol{p}}'_{tgt}] \in \mathbb{R}^{B\times 2\times 1\times L_m}$. Then, the element-wise multiplication is performed between $\boldsymbol{D}_{tse} \in \mathbb{R}^{B\times 2\times F\times L_m}$ and $[\hat{\boldsymbol{p}}'_{tgt}, \hat{\boldsymbol{p}}'_{tgt}]$. Finally, the inverse STFT (iSTFT) is applied to obtain the final TSE prediction $\hat{\boldsymbol{x}}_{tgt} \in \mathbb{R}^{B\times 1\times T_m}$.

4.7. Loss Function

The loss function consists of TSE loss and PVAD loss:

$$\mathscr{L} = \lambda_1 \cdot \mathscr{L}_{TSE} + \lambda_2 \cdot \mathscr{L}_{PVAD} \tag{21}$$

where $\mathscr{L}_{TSE}$ and $\mathscr{L}_{PVAD}$ denote the TSE and PVAD loss, respectively. λ_1 and λ_2 represent the weights for TSE and PVAD loss, respectively.

For the TSE loss, we use a scene-aware loss function [20]. In Section 3, we categorized the speech segments in the mixed audio into TA and TS scenarios. In TA, the target speaker is active, while in TS, the target speaker is silent. For TA clips, we use the scale-invariant signal-to-distortion ratio (SI-SDR) [63] as the objective for the TSE estimation:

$$\begin{cases} \boldsymbol{s}_T = \frac{<\hat{\boldsymbol{s}}, \boldsymbol{s}>\boldsymbol{s}}{\|\boldsymbol{s}\|^2} \\ \boldsymbol{s}_E = \hat{\boldsymbol{s}} - \boldsymbol{s}_T \\ \mathscr{L}_{SI-SDR} = -10\lg \frac{\|\boldsymbol{s}_T\|^2}{\|\boldsymbol{s}_E\|^2} \end{cases} \tag{22}$$

where $\hat{\boldsymbol{s}} \in \mathbb{R}^{1\times T}$ represents the estimated target speaker speech, while $\boldsymbol{s} \in \mathbb{R}^{1\times T}$ represents the clean target speech. $< \boldsymbol{s}, \boldsymbol{s} >$ denotes the power of the signal $\boldsymbol{s}$. For TS clips, we use the power loss as the objective for TSE estimation:

$$\mathscr{L}_p = 10\lg\left(|\|\boldsymbol{s}\|^2 - \|\hat{\boldsymbol{s}}\|^2| + \epsilon\right) \tag{23}$$

Unlike USEV [61], we do not directly use power but instead adopt an energy-based loss ($\mathscr{L}_p = 10\lg\left(\|\hat{\boldsymbol{s}}\|^2 + \epsilon\right)$) as the loss function. This change is mainly motivated by two considerations: first, during training, the energy in the TS segments of the target speaker's clean speech labels is nearly zero, so Equation (23) can effectively ensure that the energy in the TS segments of the TSE estimation approaches zero. Second, when the energy of the clean target speech in the labeled TS segments is non-zero (due to label noise), using energy loss as the loss function can help preserve the overall SI-SDR of the TSE estimation as much as possible. The scene-aware TSE loss $\mathscr{L}_{TSE}$ can be redefine as follows:

$$\mathscr{L}_{TSE} = \alpha_1 \cdot \mathscr{L}_{SI-SDR} + \alpha_2 \cdot \mathscr{L}_p \tag{24}$$

where α_1 and α_2 are the weights for the $\mathscr{L}_{SI-SDR}$ and $\mathscr{L}_p$, respectively. We use the binary cross-entropy loss to optimize the model for the PVAD task:

$$\mathscr{L}_{PVAD} = BCE(\hat{\boldsymbol{p}}_{tgt}, \boldsymbol{p}_{tgt}) \tag{25}$$

where $BCE(\cdot)$ denotes the binary cross-entropy loss function. $\boldsymbol{p}_{tgt}$ and $\hat{\boldsymbol{p}}_{tgt}$ represent the ground-truth VAD label of the target speaker and PVAD predicted, respectively.

5. Experimental Setup

5.1. Datasets

5.1.1. LibriMix

The LibriMix [29] dataset is derived from LibriSpeech [64] and WHAM!'s noises [65]. LibriMix is widely used for speech separation, TSE, and speaker diarization tasks. It primarily

includes two subsets: Libri2Mix and Libri3Mix. The dataset offers four variations based on two sampling rates (16 and 8 kHz) and two modes (min and max). In min mode, the mixture ends with the shortest utterance, while in max mode, the shortest utterance is padded to match the length of the longest one. We follow [20] in our experiments to combine the Libri2mix 100h and Libri3mix 100h datasets as the training set. Only the "max" version at a 16 kHz sampling rate is used in the training stage. We evaluate our proposed model on the max modes.

Same as in [20], we follow the scripts [1] to slightly adjust the data split for training and validation and the scripts [2] to prepare the reference speech of the target speaker.

5.1.2. SparseLibriMix

In addition to evaluating the model on the test set of LibriMix, we also tested the model's performance on the SparseLibriMix [29] datasets. Compared to LibriMix, the SparseLibriMix dataset moves towards more realistic, conversation-like scenarios. SparseLibriMix is a sparsely overlapping version of LibriMix. This dataset encompasses more lifelike mixture scenarios, with overlap ratios varying from 0 to 1.0.

5.1.3. CALLHOME

We test the effectiveness of the USEF-TP model on the CALLHOME dataset to evaluate the SD performance of our proposed model on real recordings. We use libri2mix and libri3mix, generated from the LibriSpeech train-clean-100 subset, as the pre-training dataset, totaling approximately 98 hours of data. We then fine-tuned the pre-trained model on the CALLHOME Part 1 dataset and evaluated it on CALLHOME Part 2. It is important to note that we directly use a raw speech segment for enrollment, meaning that higher audio quality benefits our model. In this work, We use the SD results of Diaper [3] [66] to extract the reference speech.

5.2. Network Configuration

5.2.1. PVAD 1.0 and PVAD 2.0

We use the PVAD 1.0 [22] and PVAD 2.0 [23] as the baseline models for the PVAD task. The PVAD 1.0 is a typical PVAD model with concatenation as the speaker embedding fusion module. To align the model complexity with that of the USEF-TP model, we use a 4-layer LSTM network with 256 neurons as the backbone for the PVAD 1.0 model. PVAD 2.0 is a conformer-based model with FiLM as the speaker embedding fusion module. Similarly, to ensure that the PVAD 2.0 model has parameters consistent with the USEF-TP model, we adjusted the size of PVAD 2.0 accordingly. The conformer-based PVAD 2.0 models have 4 Conformer [67] layers in the block, each having a dimension of 256, attention head of 8, feed-forward dimension of 1024, causal 7 $\times$ 7 convolution kernel, and 31 left-context. We set the number of the conformer block to 2.

5.2.2. USEF-TP

The 2D convolution layer in the encoder has a kernel size of (3,3) and a stride of 1, with input and output dimensions of 2 and 128, respectively. The attention block in the CMHA module consists of 1 layer, 4 parallel attention heads, and a 512-dimensional feed-forward network. The kernel size and stride of the torch.unfold function are both 1. In the full-band and sub-band

[1] https://github.com/msinanyildirim/USED-splits

[2] https://github.com/gemengtju/L-SpEx/tree/main/data

[3] https://github.com/BUTSpeechFIT/DiaPer

Table 1: The parameter configuration of the USEF-TP model.

Description	Configuration
The number of the attention layers, attention heads, and dimension of the feed-forward network in the CMHA module	1, 4, 512
Kernel size, stride size, input , and output dimensions for the Conv2d in the Encoder	(3,3), 1, 2, 128
Kernel and stride size for Unfold	1, 1
Number of the TF-GridNet blocks	6
Number of hidden units of BLSTMs in the TF-GridNet blocks	256
Kernel size, stride size, input , and output dimensions for the TConv2d in the TSE Decoder	(3,3), 1, 256, 2
Kernel size, stride size, input , and output dimensions for the TConv2d in the PVAD Decoder	(3,3), 2, 256, 1
Kernel size, stride size, input , and output dimensions for the Conv1d in the PVAD Decoder	2, 1, 161, 1
Kernel size, stride size, input , and output dimensions for the TConv1d in the Interaction module	2, 1, 1, 1

modules, the BLSTM layer has 256 units. The cross-frame self-attention module uses 1 layer, 4 parallel attention heads, and a 512-dimensional feed-forward network. The number of TF-GridNet blocks is 6. The 2D transposed convolution layer in the TSE Decoder has the same kernel size and stride as the 2D convolution layer, with input and output dimensions of 256 and 2, respectively. The 2D transposed convolution layer in the PVAD Decoder has the same kernel size and stride as the 2D convolution layer, with input and output dimensions of 256 and 1, respectively. The kernel size and stride of the 1D convolution layer In the PVAD Decoder are 2 and 1, with input and output dimensions of 161 and 1, respectively. The kernel size and stride of the 1D transposed convolution layer are 2 and 1, with input and output dimensions both of 1. The hyperparameters for the USEF-TP model are summarized in Table 1. Additionally, we have implemented a Speaker-Embedding-Based Target speaker extraction and Personal voice activity detection model with the same backbone as USEF-TP, which we call SEB-TP. The architecture of the SEB-TP model is illustrated in Figure 5. We use ResNet34 [68] to extract speaker embeddings in this work.

5.3. Training Details

For STFT, the window length is 20 ms, and the hop length is 10 ms. A 128-point Fourier transform is used to extract 161-dimensional complex STFT features at each frame. We trained our models using the Adam [69] optimizer, with an initial learning rate set to 1e-4. The learning rate was halved if the validation loss did not improve within 3 epochs. We set the weight of the multi-task loss λ_1, $\lambda_2 = 1$. The weights of the scene-aware TSE loss α_1 and α_2 are 1 and 0.01, respectively. In [61], the experimental results show relatively better performance when the energy loss weight for non-target speaker speech segments is set to 0.01. Moreover, when $\alpha_1 = 1$ and $\alpha_2 = 0.01$, the SI-SDR loss and energy loss are of a similar order of magnitude, with SI-SDR loss still being the predominant factor. Therefore, in this work, we adopt the same

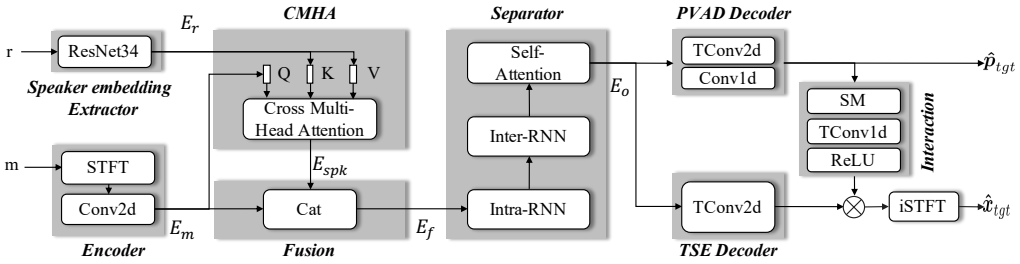

Figure 5: The diagram of SEB-TP model. 'm' and 'r' denote the mixed speech and refernece speech, respectively. $\boldsymbol{E}_r$ denotes the target speaker embedding. $\otimes$ is an operation for element-wise product. The Separator's parameters are set identically to those of the TF-GridNet approach.

hyperparameter settings as [61], $\alpha_1 = 1$ and $\alpha_2 = 0.01$. No dynamic mixing or data augmentation was applied during training. The speech clips were truncated to 4 seconds during training, while the full speech clips were evaluated during inference. In the evaluation phase, each speaker in the mixed speech is considered as the target speaker in turn.

The model's training process is conducted in two stages. In the first stage, a scene-aware multi-task loss function is not applied; instead, the model is trained using SI-SDR and BCE as the primary loss functions. The scene-aware loss function is introduced during the second stage by incorporating energy, SI-SDR, and BCE losses from the first stage.

5.4. Evaluation Metrics

We use SI-SDR or SI-SDR improvement (SI-SDRi) and signal-to-distortion ratio (SDR) or SDR improvement (SDRi) as objective metrics to assess the accuracy of TSE. We use the average energy (dB/s) to evaluate the model's performance on TS speech segments.

We use the Recall (REC), Precision (PRE), F1 score (F1) and Accuracy (ACC) to evaluate the PVAD performance. We use Diarization Error Rate (DER), including Speaker Confusion (SC), False Alarm (FA), and Missed Detection (MS) to evaluate the SD performance. Collar tolerance is set to 0 seconds for LibriMix and SparseLibrimix, 0.25 seconds for CALLHOME dataset.

6. Results and Discussions

We present results from three sets of experiments. The first set is conducted on the LibriMix dataset. The second set focuses on the SparseLibriMix datasets. The third is performed on the CALLHOME dataset. In addition to presenting the TSE and PVAD results, we also integrate each speaker's PVAD outputs to evaluate the model's SD performance. Since our model is optimized only for PVAD and TSE, the SD results show greater variability across checkpoints. The main reason is that SD performance depends on the intermediate outputs of PVAD rather than being directly optimized by the loss function. To ensure a fair and stable evaluation, we mitigate checkpoint fluctuations for SD by averaging results from three nearby checkpoints around the best PVAD or TSE checkpoint.

Table 2: Ablation study of the USEF-TP and SEB-TP model with the individual tasks on the LibriMix dataset for the max mode. 'SEB-TP' refers to the speaker-embedding-based model using the same backbone as USEF-TP.

Personal Voice Activity Detection				
Model	ACC ↑	F1 ↑	PRE ↑	Recall ↑
SEB-TP	0.900	0.921	0.972	0.874
- Only PVAD	0.877	0.905	0.936	0.875
USEF-TP	0.968	0.977	0.983	0.970
- Only PVAD	0.956	0.967	0.964	0.970
Speaker Diarization				
Model	DER (%) ↓	MI (%) ↓	FA (%) ↓	CF (%) ↓
SEB-TP	15.82	13.87	1.89	0.06
- Only PVAD	16.78	12.39	4.14	0.25
USEF-TP	4.92	3.91	0.99	0.02
- Only PVAD	6.48	4.21	2.20	0.06
Target Speaker Extraction				
Model	SDRi (dB) ↑		SI-SDRi (dB) ↑	
SEB-TP	12.38		12.07	
- Only TSE	12.23		11.80	
USEF-TP	17.38		16.82	
- Only TSE	17.11		16.51	

6.1. Results on the LibriMix Dataset

In this section, we report the experimental results of the USEF-TP model on the LibriMix dataset. The LibriMix dataset includes both max and min modes. As the data overlap ratio in min mode is nearly 100%, it is limitedly relevant to the purpose of this study. Therefore, we report only the results in max mode.

6.1.1. Comparison with Speaker Embedding based and Single Task Models

Table 2 shows the comparison results of the USEF-TP and SEB-TP models, which use speaker embedding and the same backbone as USEF-TP. The results in Table 2 show that USEF-TP outperforms SEB-TP in all metrics for the PVAD and SD tasks, with notable improvements in Recall (0.970 vs. 0.874) and DER (4.92% vs 15.82 %). It indicates that the speaker-embedding-free framework in USEF-TP helps improve the model's performance in the PVAD task.

To demonstrate the performance of USEF-TP and SEB-TP on the single tasks of TSE and PVAD, we remove the TSE and PVAD part, respectively. Table 2 shows the ablation study results of the USEF-TP and SEB-TP model with the single tasks on the LibriMix dataset. In the PVAD task, aside from Recall, the USEF-TP model outperforms the single PVAD model in terms of ACC, F1, and PRE. By aggregating the PVAD predictions for each speaker in the mixed speech, we can obtain the corresponding SD results. The USEF-TP model significantly outperforms the single PVAD model in the SD task, particularly regarding DER and MI (DER: 4.92% vs. 6.48% and MI: 3.91% vs. 4.21%). The results of the joint model and single-task model on the PVAD and SD tasks suggest that joint task training may enhance the model's PVAD performance.

For the TSE task, compared to SEB-TP, USEF-TP achieves relative improvements of 40.4% in SDRi (17.38 dB vs. 12.38 dB) and 39.4% in SI-SDRi (16.82 dB vs. 12.07 dB). The results suggest that the speaker-embedding-free framework may enhance the model's TSE performance.

Table 3: Ablation study of the USEF-TP model using the Scene-aware Loss (SL) function with or without the Interaction Module (IM) on the LibriMix dataset for max mode. 'M' denotes the mixed speech. 'IM' and 'SL' denote the Interaction Module and Scene-aware Loss function, respectively. 'TA' and 'TS' denote the target speaker active and target speaker silent speech clips, respectively.

Personal Voice Activity Detection						
Exp	IM	SL	ACC ↑	F1 ↑	Precision ↑	Recall ↑
1	✗	✗	0.954	0.964	0.964	0.946
2	✗	✔	0.965	0.974	0.985	0.962
3	✔	✗	0.969	0.976	0.984	0.966
4	✔	✔	0.970	0.977	0.983	0.970
Speaker Diarization						
Exp	IM	SL	DER (%) ↓	MI (%) ↓	FA (%) ↓	CF (%) ↓
5	✗	✗	5.65	4.30	1.33	0.02
6	✗	✔	5.01	4.00	0.99	0.02
7	✔	✗	5.38	4.46	0.90	0.02
8	✔	✔	4.92	3.91	0.99	0.02
Target Speaker Extraction						
Exp	IM	SL	Overall		TA	TS
			SDRi (dB) ↑	SI-SDRi (dB) ↑	SI-SDRi (dB) ↑	Power (dB/s) ↓
M	-	-	-	-	-	8.47
9	✗	✗	17.13	16.61	15.15	-7.38
10	✗	✔	17.34	16.67	15.37	-15.68
11	✔	✗	17.16	16.57	15.35	-9.73
12	✔	✔	17.38	16.82	15.51	-17.11

The USEF-TP model's SDRi and SI-SDRi are 0.27 dB (17.38 dB vs. 17.11 dB) and 0.31 dB (16.82 dB vs. 16.51 dB) higher than those of the single-task model, respectively. It indicates that joint training may enhance the model's TSE performance.

Overall, the USEF-TP model outperforms the SEB-TP and single-task models (Only PVAD or Only TSE) across all tasks. It indicates that speaker-embedding-free framework and joint task training can enhance performance on each subtask, particularly in the SD task.

6.1.2. Ablation Study for Interaction Module and Loss Function

Table 3 presents the ablation study results for the USEF-TP model with and without the Interaction Module (IM) and Scene-aware Loss function (SL). The experiments are conducted in max mode on the LibriMix dataset. The table is divided into three sections, showing the model's performance on the PVAD, SD, and TSE tasks. When the SL or IM is introduced (Exp 2, 3), the model's ACC, F1, and Recall all improve (ACC: 0.954 →0.965, 0.969, F1: 0.964 → 0.974, 0.976; Recall: 0.946 → 0.962, 0.966). Moreover, when the SL and IM are both introduced (Exp 4), USEF-TP achieves the best performance. For the SD task, after introducing the SL (Exp 6) or IM (Exp 7), the model's DER decrease (DER: 5.65% → 5.01%, 5.38%). With the addition of both SL and IM (Exp 8), the model's DER and MI show a further decrease(DER: 5.65% → 4.92%, MI: 4.30% → 3.91%). The results indicate that the SL and IM provides some benefits for the PVAD task.

For the TSE task, When the SL is introduced (Exp 10), both the SI-SDRi of the entire audio and the TA (target speaker activate) segment increase (16.61 dB → 16.67 dB, 15.13 dB → 15.37 dB). Moreover, the power of the TS (target speaker silent) segments decreases significantly. This indicates that SL improves the model's ability to handle both TA and TS segments. When the

Table 4: Comparison with the previous models on different tasks on the LibriMix dataset for max mode.'†' indicates the model implemented by ourselves.

Personal Voice Activity Detection					
Model	ACC ↑	F1 ↑	PRE ↑	Recall ↑	Parameter (M)
PVAD 1.0† [22]	0.816	0.856	0.893	0.822	2.3
PVAD 2.0† [23]	0.844	0.880	0.906	0.854	10.4
SEB-TP	0.900	0.921	0.972	0.874	15.1
USEF-TP	**0.970**	**0.977**	**0.983**	**0.970**	15.1
Speaker Diarization					
Model	DER (%) ↓	MI (%) ↓	FA (%) ↓	CF (%) ↓	Parameter (M)
PVAD 1.0† [22]	27.69	22.07	4.88	0.74	2.3
PVAD 2.0† [23]	22.16	16.16	5.51	0.49	10.4
SEB-TP	15.82	13.87	1.89	0.06	15.1
TS-VAD [3]	7.28	3.61	2.78	0.89	39.50
USED-F [20]	**4.75**	**2.18**	2.16	0.42	23.12
USEF-TP	4.92	3.91	**0.99**	**0.02**	15.1
Target Speaker Extraction					
Model	SDRi (dB) ↑		SI-SDRi (dB) ↑		Parameter (M)
SpEx+ [55]	10.11		9.06		16.35
SEB-TP	12.38		12.07		15.1
USED-F [20]	13.22		12.70		23.12
USEF-TP	**17.38**		**16.82**		15.1

IM is introduced (Exp 11), the overall SI-SDRi show a slight decrease (16.61 dB → 16.57 dB). In the TA conditions, the model's SI-SDRi improved by 0.2 dB (15.15 dB → 15.51 dB). Under the TS conditions, the decrease in power of the speech segments is not significant. It suggests that the IM mainly enhances performance in overlapped speech segments in the TSE task, with limited improvement in handling TS segments. With joint use of SL and IM, the model's SI-SDRi improved by 0.25 dB overall and by 0.16 dB in TA, with average energy decreasing from the original -9.73 dB/s to -17.11 dB/s. These results indicate that using SL and IM can enhance the model's overall TSE performance, particularly with a notable improvement in TS segments by reducing the inference speakers' artifacts.

6.1.3. Comparison With Previous Models

Table 4 compares the performance of the USEF-TP model across three tasks (PVAD, SD, and TSE). The USEF-TP model is evaluated against multiple baseline models for each task. For the PVAD experimental results, USEF-TP significantly outperforms PVAD 1.0 and PVAD 2.0 in ACC, F1, PRE, and Recall. Specifically, USEF-TP achieves a relative improvement of 14.9 % (0.970 vs. 0.844) in ACC and 13.6 % (0.970 vs. 0.854) in Recall compared to PVAD 2.0. It demonstrates that USEF-TP has a clear advantage over PVAD 1.0 and PVAD 2.0 models. Similarly, we aggregate the PVAD results for each speaker in the mixed speech within the test dataset to compute the model's SD results and compare them with other SD baseline models. The experimental results show that USEF-TP outperforms other models in DER, FA, and CF metrics, with a particularly notable result in CF (0.02 %). It indicates that the USEF-TP model has lower error and confusion rates in the SD task. Although DER and MI is slightly higher than

Table 5: Ablation study of the USEF-TP and SEB-TP model with the individual tasks on the SparseLibriMix dataset. Due to space constraints, we present only the results for overlap ratios of 0% / 20% / 40% in the table. 'SEB-TP' refers to the speaker-embedding-based model using the same backbone as USEF-TP.

Personal Voice Activity Detection (with overlap ratios of 0% / 20% / 40%)				
Model	ACC ↑	F1 ↑	Precision ↑	Recall ↑
SEB-TP	0.896 / 0.886 / 0.888	0.849 / 0.880 / 0.891	0.891 / 0.932 / 0.942	0.810 / 0.834 / 0.845
- Only PVAD	0.763 / 0.810 / 0.827	0.711 / 0.822 / 0.848	0.633 / 0.771 / 0.808	0.809 / 0.881 / 0.892
USEF-TP	0.969 / 0.971 / 0.971	0.956 / 0.971 / 0.973	0.951 / 0.971 / 0.971	0.962 / 0.971 / 0.976
- Only PVAD	0.922 / 0.928 / 0.933	0.896 / 0.930 / 0.939	0.861 / 0.906 / 0.922	0.935 / 0.954 / 0.957
Speaker Diarization (with overlap ratios of 0% / 20% / 40%)				
Model	DER (%) ↓	MI (%) ↓	FA (%) ↓	CF (%) ↓
SEB-TP	28.64 / 23.63 / 22.13	20.60 / 18.57 / 17.90	6.66 / 4.42 / 3.76	1.38 / 0.65 / 0.46
- Only PVAD	53.08 / 33.95 / 28.95	21.38 / 16.75 / 15.80	25.40 / 14.81 / 11.50	6.30 / 2.38 / 1.65
USEF-TP	8.33 / 5.62 / 5.23	3.54 / 2.87 / 2.71	4.57 / 2.64 / 2.46	0.22 / 0.12 / 0.06
- Only PVAD	21.22 / 14.00 / 11.74	3.86 / 3.29 / 3.27	16.40 / 10.33 / 8.22	0.96 / 0.38 / 0.24
Target Speaker Extraction (with overlap ratios of 0% / 20% / 40%)				
Model	SDRi (dB) ↑		SI-SDRi (dB) ↑	
SEB-TP	15.51 / 14.41 / 14.46		14.79 / 13.19 / 13.16	
- Only TSE	14.73 / 14.11 / 13.87		13.26 / 12.41 / 12.21	
USEF-TP	20.39 / 19.27 / 19.28		19.78 / 18.68 / 18.68	
- Only TSE	19.62 / 18.70 / 18.52		19.05 / 18.11 / 17.92	

USED-F (DER:4.92 % vs. 4.75 %; MI: 3.91 % vs. 2.18 %), the overall performance remains superior to other models.

For the TSE results, USEF-TP significantly surpasses other models on SDRi and SI-SDRi metrics, indicating a more substantial improvement in the TSE task. Specifically, compared to the USED-F model, the USEF-TP model achieves improvements of 31.5 % (17.38 dB vs. 13.22 dB) in SDRi and 32.4 % (16.82 dB vs. 12.70 dB) in SI-SDRi. It demonstrates that the USEF-TP model excels in extracting high-quality target speaker signals

The USEF-TP model outperforms other baseline models across all tasks, with a notable advantage in the SDRi and SI-SDRi for the TSE task. Additionally, the USEF-TP model demonstrates significant improvement in all metrics for PVAD, indicating that the joint tasks training enhances the model's performance on both TSE and PVAD tasks.

6.2. Results on the SparseLibriMix Dataset

In this section, we report the experimental results of the USEF-TP model on the SparseLibriMix dataset, which includes multi-speaker mixed test speech with six different overlap ratios: 0%, 20%, 40%, 60%, 80%, and 100%. The primary goal of this set of experiments is to evaluate the proposed model's performance under sparse overlap conditions. The demo are available at this link [3].

6.2.1. Comparison with Speaker Embedding based and Single Task Models

Table 5 presents the comparison results on the SparseLibriMix dataset, comparing the USEF-TP model with single-task models containing USEF-TP and SEB-TP (Only PVAD) or USEF-TP

[3] `https://github.com/ZBang/USEF-TP`

and SEB-TP (Only TSE) across different overlap ratios (0%, 20%, and 40%). The USEF-TP and SEB-TP models outperform the Only PVAD models in ACC, F1, Precision, and Recall across all overlap ratios, indicating that the joint model USEF-TP achieves better overall performance in the PVAD task, with notably stable performance at higher overlap ratios (40%). Additionally, the USEF-TP model achieves significantly lower DER, MI, FA, and CF scores across all overlap ratios, especially in DER at an overlap ratio of 0% (DER: 8.33% vs. 21.22%). It suggests that multi-task joint training can significantly enhance PVAD performance under sparse overlap conditions. By comparing the experimental results of USEF-TP and SEB-TP, we observe that USEF-TP achieves significant improvements in both F1 score and DER across different overlap ratios, with a notable improvement in Recall (overlap ration 0% : 0.962 vs. 0.810). This demonstrates that adopting the speaker-embedding-free framework enhances the model's performance in the PVAD task.

The USEF-TP model achieves higher SDRi and SI-SDRi scores than the Only TSE model, with the largest gap observed at 0% overlap (SDRi: 20.39 dB vs. 19.62 dB, SI-SDRi: 19.78 dB vs. 19.05 dB). It indicates that under sparse overlap conditions, multi-task joint training can significantly enhance TSE performance. Furthermore, the joint model exhibits more stable performance across different overlap ratios compared to the only TSE model. Additionally, when the overlap ratio is 0, USEF-TP achieves a 33.7% (19.78 dB vs. 14.79 dB) improvement in SI-SDRi compared to SEB-TP. These results indicate that the speaker-embedding-free framework enhances the model's performance in the TSE task. Besides speaker identity, other information in the reference speech, such as contextual cues, may also contribute to extracting the target speaker's clean speech.

From the experimental results in Table 5, we can conclude that the USEF-TP model outperforms the SEB-TP model on both tasks. The results demonstrate the effectiveness of the speaker-embedding-free framework in handling mixtures with different overlap ratios. Moreover the USEF-TP model outperforms the single-task models across all tasks, maintaining strong performance at low and high overlap ratios. It indicates that joint task training enhances the model's performance and improves its stability.

6.2.2. Ablation Study for Interaction Module and Loss Function

Table 6 presents the ablation study results on the impact of including the IM and SL on the performance of the USEF-TP model. It examines the model's performance on the SparseLibriMix dataset under different combinations of these two modules. For the PVAD task, incorporating the SL (Exp 3) or IM (Exp 3) leads to improvements in ACC, F1, Precision, and Recall, with further enhancements observed when SL and IM (Exp 4) are both applied. Furthermore, comparing EXP 5,6,7,8 shows that at a 0% overlap ratio, the addition of SL and IM leads to an apparent decrease in DER (9.56% $\rightarrow$ 9.06%, 9.21% $\rightarrow$ 8.33%). These results indicate that the IM and SL can enhance the model's PVAD performance even under sparse overlap conditions.

For the TSE task, as SL and IM are added (from Exp 9 to Exp 12), both SDRi and SI-SDRi improve incrementally. It indicates that the IM and SL positively enhance overall TSE performance. Comparing Exp 9 to Exp 12 reveals that as the overlap ratio increases, overall SI-SDRi tends to decrease, while SI-SDRi under TA conditions shows an upward trend. It suggests the model's capability to handle TS speech segments diminishes with higher overlap ratios. However, Exp 12 maintains a more robust performance in overall SI-SDRi, and compared to Exp 11, Exp 10 and 12 show a more noticeable reduction in average energy under TS conditions. It indicates that SL effectively reduces the power of interfering speakers or noise in silent segments.

Table 6: Ablation study of the USEF-TP model using the scene-aware loss function with or without the Interaction module on the SparseLibriMix dataset. 'M' denotes the mixed speech. 'IM' and 'SL' denote the Interaction module and scene-aware loss function, respectively. 'TA' and 'TS' denote the target speaker active and target speaker silent speech clips, respectively. Due to space constraints, we present only the results for overlap ratios of 0% / 20% / 40% in the table.

Personal Voice Activity Detection

(with overlap ratios of 0% / 20% / 40%)

Exp	IM	SL	ACC ↑	F1 ↑	Precision ↑	Recall ↑
1	✗	✗	0.962 / 0.963 / 0.964	0.946 / 0.962 / 0.967	0.949 / 0.963 / 0.966	0.943 / 0.962 / 0.968
2	✗	✔	0.967 / 0.970 / 0.970	0.955 / 0.965 / 0.972	0.945 / 0.965 / 0.969	0.965 / 0.975 / 0.976
3	✔	✗	0.967 / 0.966 / 0.965	0.954 / 0.966 / 0.965	0.951 / 0.967 / 0.967	0.958 / 0.965 / 0.970
4	✔	✔	0.969 / 0.971 / 0.971	0.956 / 0.971 / 0.973	0.951 / 0.971 / 0.971	0.962 / 0.971 / 0.976

Speaker Diarization

(with overlap ratios of 0% / 20% / 40%)

Exp	IM	SL	DER (%) ↓	MI (%) ↓	FA (%) ↓	CF (%) ↓
5	✗	✗	9.53 / 6.34 / 5.73	3.88 / 2.95 / 2.74	5.34 / 3.25 / 2.92	0.31 / 0.14 / 0.07
6	✗	✔	9.06 / 5.98 / 5.36	3.66 / 2.70 / 2.54	5.09 / 3.17 / 2.76	0.31 / 0.11 / 0.07
7	✔	✗	9.21 / 6.22 / 5.65	3.82 / 3.21 / 2.94	5.07 / 2.87 / 2.64	1.22 / 0.14 / 0.08
8	✔	✔	8.33 / 5.62 / 5.23	3.54 / 2.87 / 2.71	4.57 / 2.64 / 2.46	0.22 / 0.12 / 0.06

Target Speaker Extraction

(with overlap ratios of 0% / 20% / 40%)

Exp	IM	SL	Overall		TA	TS
			SDRi (dB) ↑	SI-SDRi (dB) ↑	SI-SDRi (dB) ↑	Power (dB/s) ↓
M	-	-	-	-	-	3.73 / 5.94 / 6.81
9	✗	✗	20.03 / 18.92 / 18.63	19.37 / 18.29 / 17.96	11.62 / 14.36 / 14.83	-1.14 / -2.70 / -3.89
10	✗	✔	20.15 / 19.11 / 18.86	19.43 / 18.41 / 18.19	11.63 / 14.45 / 14.91	-2.60 / -5.61 / -7.61
11	✔	✗	20.38 / 18.96 / 18.62	19.76 / 18.34 / 17.99	11.92 / 14.41 / 14.88	-1.79 / -3.71 / -5.07
12	✔	✔	20.39 / 19.27 / 19.28	19.78 / 18.68 / 18.68	11.90 / 14.65 / 15.14	-3.35 / -6.79 / -8.89

In summary, under various overlap conditions, the joint model USEF-TP with both IM and SL outperforms versions without these modules or with only a single module across all tasks, demonstrating the effectiveness of IM and SL in enhancing model performance and stability.

6.2.3. Comparison With Previous Models

Figure 6 illustrates the performance of several models on the SD and TSE tasks on the SparseLibriMix dataset across different overlap ratios (0% to 100%). It includes four models: TS-VAD [3], SpEx+ [55], USED-F [20], USEF-TP, as well as various task settings for USEF-TP. The plot (a) shows the DER for the SD task, while the plot (b) displays the SI-SDRi for the TSE task. From (a), we can see that the DER of the models generally increases as the overlap ratio decreases, indicating that as the overlapping segments in the audio decrease, the SD task becomes more challenging. Across different overlap ratios, USEF-TP consistently achieves a lower DER than the Only PVAD model, indicating that the joint model design helps reduce DER and improve SD performance. Except at the 0% overlap ratio, where USEF-TP's DER is slightly higher than that of the USED-F model, USEF-TP outperforms the other models in DER at all other overlap ratios. This advantage is particularly notable at low overlap rates (20% and 40%). The USEF-TP model maintains a low DER across various overlap ratios, demonstrating strong performance and stability.

For the TSE task, as shown in (b), the SI-SDRi of the models generally decreases as the overlap ratio increases, indicating that higher overlap makes the TSE task more challenging. The SI-SDRi of USEF-TP is significantly higher than that of the single-task model (Only TSE), further demonstrating the advantages of the joint model. The USEF-TP model achieves signif-

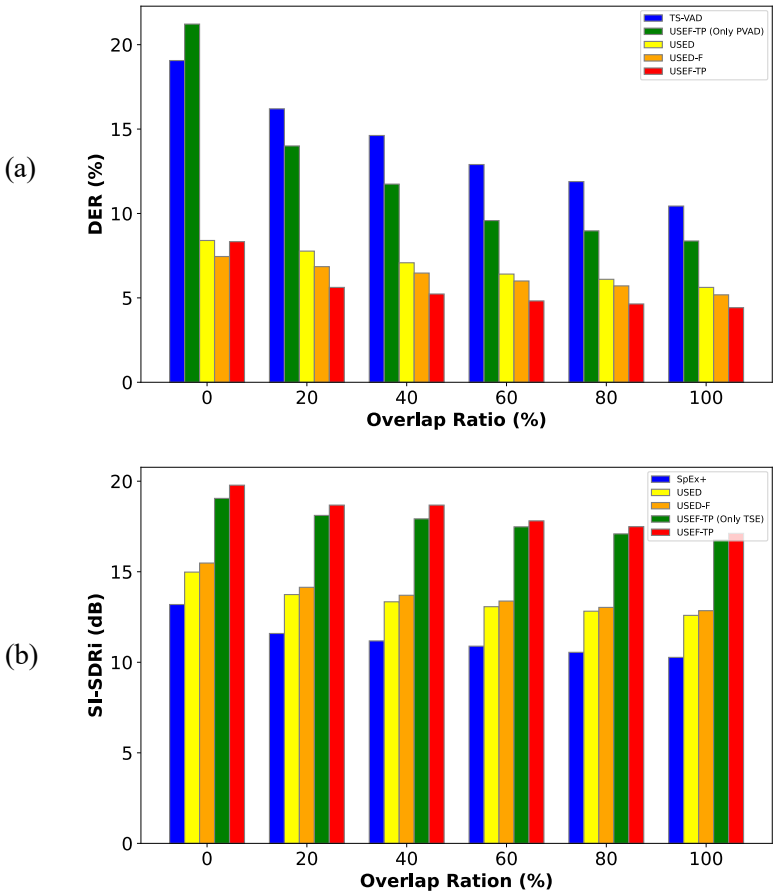

Figure 6: Comparison with previous models on the SparseLibriMix for overlap ratio from 0% to 100%. Due to space constraints, we just present the results of SD and TSE tasks. (a) is the result of the SD task. (b) is the result of the TSE task. The results of the TS-VAD, SpEx+, USED, and USEF-F are given in [20].

icantly higher SI-SDRi than other models across all overlap ratios, with a solid performance at high overlap rates (80% and 100%). It highlights the model's robustness in scenarios with varying overlap ratios.

In summary, the USEF-TP model performs excellently in both SD and TSE tasks, exhibiting superior effectiveness and stability across different overlap conditions. Nonetheless, we observed a phenomenon where the model's DER decreases as the overlap ratio of the test data increases. One possible reason is that, to compare with previous work, we used the same training and test sets as [18, 20], namely LibriMix and SparseLibriMix. The LibriMix dataset mainly contains speech with high overlap ratios, likely leading to the model performing better on high-overlap speech. The data distribution of the SparseLibriMix test set might also be another factor contributing to this phenomenon. In future work, we plan to use LibriSpeech to simulate sparse overlapping speech closer to real-world data to reflect real-world scenarios better.

6.3. Results on the CALLHOME Dataset

Table 7: Comparison of DER (%) with previous models on the CALLHOME Dataset.

Model	Number of Speakers					
	2	**3**	**4**	**5**	**6**	**all**
AHC clustering [2]	15.45	18.01	22.68	31.40	34.27	19.43
Spectral clustering [70]	16.05	18.77	22.05	30.46	36.85	19.83
EEND-EDA [2]	8.50	13.24	21.46	33.16	40.29	15.29
AED-EEND [71]	6.96	12.56	18.26	34.32	44.52	14.22
TS-VAD [3]	12.95	14.57	20.26	27.36	32.66	15.51
DiaPer [66]	7.39	12.08	19.62	30.25	28.84	13.60
USED-F [20]	9.09	12.76	14.67	26.96	26.71	13.16
USED [20]	10.23	12.78	14.68	32.02	25.10	13.51
USEF-TP	9.21	14.43	20.48	32.96	26.27	15.26

Table 7 presents the DER comparison between the proposed USEF-TP model and several state-of-the-art SD systems on the CALLHOME dataset. As shown, USEF-TP achieves an overall DER of 15.26%, which is competitive with models such as EEND-EDA (15.29%) and TS-VAD (15.51%) and outperforms traditional clustering-based methods like AHC and spectral clustering by a significant margin. This demonstrates the effectiveness of USEF-TP in real conversational scenarios.

However, the USEF-TP model still falls short compared to the recently proposed USED [20] (13.51%) and DiaPer [66] (13.60%), particularly in scenarios with five speakers, where the performance gap becomes more evident. A potential reason is that conventional SD or Sspeech separation models (e.g., DiaPer and USED) simultaneously generate outputs for all speakers in the mixed speech, inherently modeling the relationships among all speakers. In contrast, TSE or PVAD models (e.g., USEF-TP) focus only on a target speaker, producing predictions about the target speaker. The ability to capture speaker interactions in SD and speech separation models may be beneficial for improving performance in complex multi-speaker scenarios. In future work, we may leverage this aspect to enhance the model's performance in real conversational scenarios.

7. Conclusion

This paper proposes a Universal Speaker Embedding Free Target speaker extraction and Personal voice activity detection (USEF-TP) model. The USEF-TP model performs target speaker extraction and personal voice activity detection jointly. USEF-TP utilizes an embedding-free framework as the network backbone, and we designed an interaction module to enable PVAD results to inform TSE predictions. Additionally, we introduced an energy loss to enhance the model's performance across mixed speech with varying overlap ratios. Experimental results indicate that our proposed USEF-TP demonstrates superior performance and stability across overlapping conditions in PVAD and TSE tasks.

8. Acknowledgments

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Suzhou City(SYC2022051) and Guangdong Science and Technology Plan (2023A1111120012). Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

References

[1] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with permutation-free objectives, in: Proc. of Interspeech, 2019, pp. 4300–4304.

[2] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, K. Nagamatsu, End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, in: Proc. of Interspeech, 2020, pp. 269–273.

[3] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, et al., Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario, in: Proc. of Interspeech, 2020, pp. 274–278.

[4] M. Cheng, W. Wang, Y. Zhang, X. Qin, M. Li, Target-speaker voice activity detection via sequence-to-sequence prediction, in: Proc. of ICASSP, 2023, pp. 1–5.

[5] Z. Chen, B. Han, S. Wang, Y. Qian, Attention-based encoder-decoder end-to-end neural diarization with embedding enhancer, IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024) 1636–1649.

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

[7] N. Zeghidour, D. Grangier, Wavesplit: End-to-end speech separation by speaker clustering, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 2840–2849.

[8] Y. Luo, Z. Chen, T. Yoshioka, Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation, in: Proc. of ICASSP, 2020, pp. 46–50.

[9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, J. Zhong, Attention is all you need in speech separation, in: Proc. of ICASSP, 2021, pp. 21–25.

[10] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, S. Watanabe, Tf-gridnet: Integrating full-and sub-band modeling for speech separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 3221–3236.

[11] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Q. Yip, D. Ng, B. Ma, Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation, in: Proc. of ICASSP, 2024, pp. 10356–10360.

[12] D. Cai, M. Li, Leveraging asr pretrained conformers for speaker verification through transfer learning and knowledge distillation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024) 3532–3545.

[13] H. Wang, S. Zheng, Y. Chen, L. Cheng, Q. Chen, Cam++: A fast and efficient network for speaker verification using context-aware masking, in: Proc. of Interspeech, 2023, pp. 5301–5305.

[14] X. Qin, N. Li, S. Duan, M. Li, Investigating long-term and short-term time-varying speaker verification, IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2024) 3408–3423.

[15] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, X. Lei, Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit, in: Proc. of Interspeech, 2021, pp. 4054–4058.

[16] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe, End-to-end speech recognition: A survey, IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[17] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe, et al., Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study, in: Proc. of ICASSP, 2024, pp. 11481–11485.

[18] S. Maiti, Y. Ueda, S. Watanabe, C. Zhang, M. Yu, S.-X. Zhang, Y. Xu, Eend-ss: Joint end-to-end neural speaker diarization and speech separation for flexible number of speakers, in: Proc. of SLT, IEEE, 2023, pp. 480–487.

[19] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, J. Le Roux, Ts-sep: Joint diarization and separation conditioned on estimated speaker embeddings, IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[20] J. Ao, M. S. Yldrm, R. Tao, M. Ge, S. Wang, Y. Qian, H. Li, Used: Universal speaker extraction and diarization, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024) 1–14.

[21] B. Zeng, H. Suo, Y. Wan, M. Li, Simultaneous speech extraction for multiple target speakers under meeting scenarios, Journal of Shanghai Jiaotong University (Science) (2024) 1–7.

[22] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, I. Lopez Moreno, Personal VAD: Speaker-Conditioned Voice Activity Detection, in: Proc. of Odyssey, 2020, pp. 433–439.

[23] S. Ding, R. V. Rikhye, Q. Liang, Y. He, Q. Wang, A. Narayanan, T. O'Malley, I. McGraw, Personal vad 2.0: Optimizing personal voice activity detection for on-device speech recognition, in: Proc. of Interspeech, 2022, pp. 3744–3748.

[24] B. Zeng, M. Cheng, Y. Tian, H. Liu, M. Li, Efficient personal voice activity detection with wake word reference speech, in: Proc. of ICASSP, 2024, pp. 12241–12245.

[25] L. Wan, Q. Wang, A. Papir, I. L. Moreno, Generalized end-to-end loss for speaker verification, in: Proc. of ICASSP, 2018, pp. 4879–4883.

[26] X. Qin, D. Cai, M. Li, Robust multi-channel far-field speaker verification under different in-domain data availability scenarios, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022) 71–85.
[27] B. Zeng, M. Li, Usef-tse: Universal speaker embedding free target speaker extraction, arXiv preprint arXiv:2409.02615.
[28] J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, in: Proc. of ICASSP, 2016, pp. 31–35.
[29] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, E. Vincent, Librimix: An open-source dataset for generalizable speech separation, arXiv preprint arXiv:2005.11262.
[30] S. H. Shum, N. Dehak, R. Dehak, J. R. Glass, Unsupervised methods for speaker diarization: An integrated and iterative approach, IEEE Transactions on Audio, Speech, and Language Processing 21 (10) (2013) 2015–2028.
[31] G. Sell, D. Garcia-Romero, Speaker diarization with plda i-vector scoring and unsupervised calibration, in: Proc. of SLT, 2014, pp. 413–417.
[32] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, S. Narayanan, A review of speaker diarization: Recent advances with deep learning, Computer Speech & Language 72 (2022) 101317.
[33] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with self-attention, in: Proc. of ASRU, 2019, pp. 296–303.
[34] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, J. Wu, Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization, in: Proc. of ICASSP, 2023, pp. 1–5.
[35] M. Cheng, M. Li, Multi-input multi-output target-speaker voice activity detection for unified, flexible, and robust audio-visual speaker diarization, arXiv preprint arXiv:2401.08052.
[36] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, S. Watanabe, Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker, in: Proc. of Interspeech, 2021, pp. 2523–2527.
[37] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: Proc. of ICASSP, 2018, pp. 5329–5333.
[38] M. N. Schmidt, R. K. Olsson, Single-channel speech separation using sparse non-negative matrix factorization., in: Proc. of Interspeech, 2006, pp. 2–5.
[39] A. Cichocki, R. Zdunek, S.-i. Amari, New algorithms for non-negative matrix factorization in applications to blind source separation, in: Proc. of ICASSP, 2006, pp. V–V.
[40] R. Lyon, A computational model of binaural localization and separation, in: Proc. of ICASSP, Vol. 8, 1983, pp. 1148–1151.
[41] D. Wang, G. J. Brown, Computational auditory scene analysis: Principles, algorithms, and applications, Wiley-IEEE press, 2006.
[42] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, J. R. Hershey, Single-channel multi-speaker separation using deep clustering, in: Proc. of Interspeech, 2016, pp. 545–549.
[43] Z. Chen, Y. Luo, N. Mesgarani, Deep attractor network for single-microphone speaker separation, in: Proc. of ICASSP, 2017, pp. 246–250.
[44] Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network, IEEE/ACM transactions on Audio, Speech, and Language Processing 26 (4) (2018) 787–796.
[45] D. Yu, M. Kolbæk, Z.-H. Tan, J. Jensen, Permutation invariant training of deep models for speaker-independent multi-talker speech separation, in: Proc. of ICASSP, 2017, pp. 241–245.
[46] Y. Luo, N. Mesgarani, Tasnet: time-domain audio separation network for real-time, single-channel speech separation, in: Proc. of ICASSP, 2018, pp. 696–700.
[47] S. Zhao, B. Ma, Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions, in: Proc. of ICASSP, 2023, pp. 1–5.
[48] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, I. L. Moreno, VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking, in: Proc. of Interspeech, 2019, pp. 2728–2732.
[49] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, J. Černockỳ, Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures, IEEE Journal of Selected Topics in Signal Processing 13 (4) (2019) 800–814.
[50] T. Li, Q. Lin, Y. Bao, M. Li, Atss-Net: Target Speaker Separation via Attention-Based Neural Network, in: Proc. of Interspeech, 2020, pp. 1411–1415.
[51] Z. Zhang, B. He, Z. Zhang, X-tasnet: Robust and accurate time-domain speaker extraction network, in: Proc. of Interspeech, 2020, pp. 1421–1425.
[52] C. Xu, W. Rao, E. S. Chng, H. Li, Spex: Multi-scale time domain speaker extraction network, IEEE/ACM transactions on audio, speech, and language processing 28 (2020) 1370–1384.
[53] K. Liu, Z. Du, X. Wan, H. Zhou, X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion, in: Proc. of ICASSP, 2023, pp. 1–5.
[54] F. Hao, X. Li, C. Zheng, X-tf-gridnet: A time–frequency domain target speaker extraction network with adaptive

speaker embedding fusion, Information Fusion 112 (2024) 102550.
[55] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, H. Li, Spex+: A complete time domain speaker extraction network, in: Proc. of Interspeech, 2020, pp. 1406–1410.
[56] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, H. Li, Multi-stage speaker extraction with utterance and frame-level reference signals, in: Proc. of ICASSP, 2021, pp. 6109–6113.
[57] W. Wang, C. Xu, M. Ge, H. Li, Neural speaker extraction with speaker-speech cross-attention network., in: Proc. of Interspeech, 2021, pp. 3535–3539.
[58] L. Yang, W. Liu, L. Tan, J. Yang, H.-G. Moon, Target speaker extraction with ultra-short reference speech by ve-ve framework, in: Proc. of ICASSP, 2023, pp. 1–5.
[59] B. Zeng, H. Suo, Y. Wan, M. Li, Sef-net: Speaker embedding free target speaker extraction network, in: Proc. of Interspeech, 2023, pp. 3452–3456.
[60] X. Yang, C. Bao, J. Zhou, X. Chen, Target speaker extraction by directly exploiting contextual information in the time-frequency domain, in: Proc. of ICASSP, 2024, pp. 10476–10480.
[61] Z. Pan, M. Ge, H. Li, Usev: Universal speaker extraction with visual cue, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022) 3032–3045.
[62] E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: Visual reasoning with a general conditioning layer, in: Proc. of AAAI, 2018, pp. 3942–3951.
[63] J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey, Sdr–half-baked or well done?, in: Proc. of ICASSP, 2019, pp. 626–630.
[64] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, in: Proc. of ICASSP, 2015, pp. 5206–5210.
[65] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, J. Le Roux, Wham!: Extending speech separation to noisy environments, in: Proc. of Interspeech, 2019, pp. 2019–2821.
[66] F. Landini, T. Stafylakis, L. Burget, et al., Diaper: End-to-end neural diarization with perceiver-based attractors, IEEE/ACM Transactions on Audio, Speech, and Language Processing.
[67] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in: Proc. of Interspeech, 2020, pp. 5036–5040.
[68] Y. Lin, X. Qin, G. Zhao, M. Cheng, N. Jiang, H. Wu, M. Li, Voxblink: A large scale speaker verification dataset on camera, in: Proc. of ICASSP, 2024, pp. 10271–10275.
[69] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proc. of ICLR, 2015.
[70] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, I. L. Moreno, Speaker diarization with lstm, in: Proc. of ICASSP, 2018, pp. 5239–5243.
[71] Z. Chen, B. Han, S. Wang, Y. Qian, Attention-based encoder-decoder network for end-to-end neural speaker diarization with target speaker attractor, in: Proc. of Interspeech, 2023, pp. 3552–3556.