Multi-Input Multi-Output Target-Speaker Voice Activity Detection For Unified, Flexible, and Robust Audio-Visual Speaker Diarization

Ming Cheng, Ming Li, *Senior Member, IEEE*

***Abstract*—Audio-visual learning has demonstrated promising results in many classical speech tasks (e.g., speech separation, automatic speech recognition, wake-word spotting). We believe that introducing visual modality will also benefit speaker diarization. To date, Target-Speaker Voice Activity Detection (TS-VAD) plays an important role in highly accurate speaker diarization. However, previous TS-VAD models take audio features and utilize the speaker's acoustic footprint to distinguish his or her personal speech activities, which is easily affected by overlapped speech in multi-speaker scenarios. Although visual information naturally tolerates overlapped speech, it suffers from spatial occlusion, low resolution, etc. The potential modality-missing problem blocks TS-VAD towards an audio-visual approach. This paper proposes a novel Multi-Input Multi-Output Target-Speaker Voice Activity Detection (MIMO-TSVAD) framework for speaker diarization. The proposed method can take audio-visual input and leverage the speaker's acoustic footprint or lip track to flexibly conduct audio-based, video-based, and audio-visual speaker diarization in a unified sequence-to-sequence framework. Experimental results show that the MIMO-TSVAD framework demonstrates state-of-the-art performance on the VoxConverse, DIHARD-III, and MISP 2022 datasets under corresponding evaluation metrics, obtaining the Diarization Error Rates (DERs) of 4.18%, 10.10%, and 8.15%, respectively. In addition, it can perform robustly in heavy lip-missing scenarios.**

***Index Terms*—Speaker Diarization, Target-Speaker Voice Activity Detection, Audio-Visual Neural Networks**

I. Introduction

Like documenting events in a diary, speaker diarization is the task of automatically detecting multiple speakers' utterance boundaries in conversational data [1]. It aims to split the audio or multi-modal signals into segments with labeled identities, solving the problem of "Who-Spoke-When." As a front-end technique, it is essential in various downstream applications (e.g., speech recognition) [2].

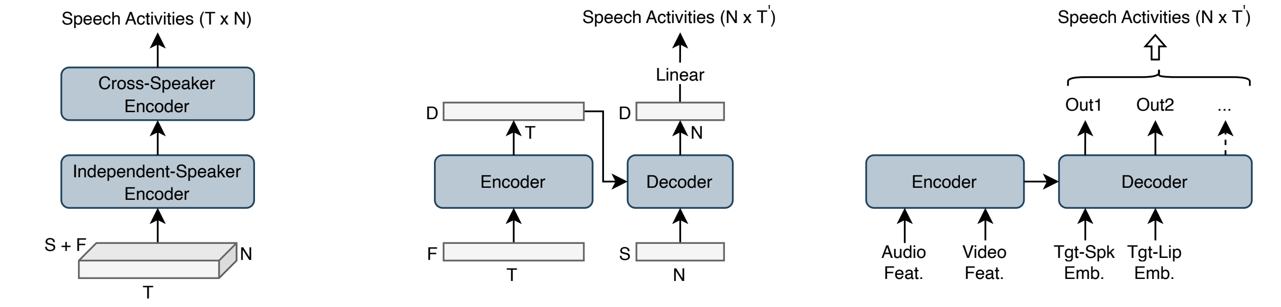
In previous studies, speaker diarization research mainly focuses on audio streams [2]. The conventional method, also known as the modularized method, utilizes cascaded modules to partition the audio signal into short segments and cluster their identities by advanced speaker representation techniques [3]–[5]. These methods cannot handle overlapped speech as each audio segment is supposed to be speaker-homogeneous. Some studies propose additional post-processing techniques (e.g., overlapped speech detection [6], overlap-aware resegmentation [5], [7]) to compromise this effect. Then, End-to-End Neural Diarization (EEND) systems [8]–[11] are proposed to estimate multiple speakers' speech activities as multi-label classification. The end-to-end structure of neural networks leads to ease of optimization and robustness to overlapped speech. Nevertheless, permutation-invariant training in EEND-based methods causes performance degradation when the number of speakers increases in long audios. Although a few studies [12]–[16] have explored the unsupervised clustering to address this problem, their results are still unsatisfactory. Recently, TS-VAD approaches [17]–[20] become attractive, which combine advantages of modularized methods and end-to-end neural networks. As a post-processing method, the TS-VAD framework requires an initial diarization system (e.g., modularized method) to extract each speaker's acoustic footprint as the speaker profile, solving the problem of "Who-Spoke" in advance. Then, a neural network-based model takes speech features and all speaker profiles to predict their corresponding framewise voice activities, aiming to address the "When" problem. This two-stage process has demonstrated excellent performance in popular benchmarks such as DIHARD-III [21] and VoxSRC21-23 [22]–[24].

However, speaker diarization in complex environments (e.g., far-field and highly overlapped speech, a large number of speakers) is still challenging. Using visual information as the complementary modality to improve diarization systems becomes a promising direction. Existing works mainly depend on constructing cross-modal synergy [25]–[28], clustering on audio-visual pairs [29], [30], or end-to-end audio-visual diarization [31], [32], which are basically derived from the previous audio-only methods. Motivated by the highly accurate performance in TS-VAD studies, a question arises if it is feasible to investigate this framework in an audio-visual manner. Although similar works about Active Speaker Detection (ASD) [33]–[35] also estimate the speaker's voice activities using audio-visual signals, they usually work for a single speaker at once and neglect the modality-missing problems. So far, there is a lack of an audio-visual diarization framework that can effectively deal with multi-speaker scenarios where the visual modality often suffers occlusion, off-screen speakers, or unreliable detection.

In this article, we propose a novel MIMO-TSVAD framework for audio-visual speaker diarization in modality-missing scenarios. Let $\mathbf{X}_a$ and $\mathbf{X}_v$ denote the audio and video features, respectively. $\mathbf{E}_{spk}$ represents the target-speaker embeddings in classical TS-VAD systems. We additionally define target-

Ming Cheng and Ming Li are with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems, Digital Innovation Research Center, Duke Kunshan University, Kunshan 215316, China.

Corresponding author: Ming Li, E-mail: ming.li369@dukekunshan.edu.cn



(a) Encoder-Only [17], [18]　(b) Sequence-to-Sequence [19]　(c) Multi-Input Multi-Output

Fig. 1. Overview of different TS-VAD frameworks. In (a) and (b), T and F denote the length and dimension of extracted audio features. N and S denote the number and dimension of speaker embeddings. D indicates the output dimension of the decoder, which can be converted into the length of detected speech activities by a linear layer. In (c), The multi-modal encoder and multi-task decoder take multiple kinds of input (e.g., audio/video features, target-speaker/lip embeddings) and support various output methods. For clarity, front-end extractors to obtain audio-visual features are omitted in the plot.

lip embeddings to indicate speaker identities of corresponding lip tracks, denoted as $\mathbf{E}_{lip}$. Depending on the accessibility of different data during inference, there are four cases , namely $\mathbf{X}_a$ vs. $\mathbf{E}_{spk}$, $\mathbf{X}_v$ vs. $\mathbf{E}_{lip}$, $\mathbf{X}_a + \mathbf{X}_v$ vs. $\mathbf{E}_{lip}$, $\mathbf{X}_a + \mathbf{X}_v$ vs. $\mathbf{E}_{spk} + \mathbf{E}_{lip}$. Our proposed framework can flexibly process arbitrary cases with state-of-the-art performance and strong robustness.

Fig. 1 illustrates the progress of our TS-VAD research. The classical TS-VAD systems [17], [18] can be abstracted into the encoder-only methods, shown in Fig. 1a. $\mathbf{X}_a \in \mathbb{R}^{\mathrm{T}\times\mathrm{F}}$ is extracted from input audio by traditional methods (e.g., MFCC, Fbank) or deep neural networks. $\mathbf{E}_{spk} \in \mathbb{R}^{\mathrm{N}\times\mathrm{S}}$ represents the given speaker embeddings. Each speaker embedding $\in \mathbb{R}^{\mathrm{S}}$ has to be repeated T times and concatenated with $\mathbf{X}_a$, producing a 3-dimensional tensor with the shape of $T \times N \times (F + S)$. Then, encoders (e.g., Bi-LSTM, Transformers) process the input tensor along the time axis (T) and speaker axis (N). As shown in Fig. 1b, we introduce the sequence-to-sequence architecture [19] to factorize the input onto the encoder and decoder separately. This way, the consumption of input memory becomes proportional to $T \times F + N \times S$, enabling the model to process longer audio and more speakers. Moreover, in the previous encoder-only architecture, the output feature length (T') must be equal to the input feature length (T). In the sequence-to-sequence architecture, adjusting the output dimension (T') of the last linear layer can be easily implemented. As the model accepts a fixed-length speech chunk as input when T is determined, using a larger output length (T') to predict the fixed-length voice activities can achieve a higher temporal resolution. In this work, the MIMO-TSVAD framework is designed based on the sequence-to-sequence architecture, demonstrated in Fig. 1c. The multi-modal encoder and multi-task decoder leverage cross-modal and inter-speaker relationships, which can utilize various types of input data to obtain different predictions in a Multi-Input and Multi-Output (MIMO) manner.

This paper extends our previous work that presents the fundamental algorithm of Seq2Seq-TSVAD [19]. Also, a simple version based on lip information has been initially described in our technical report [36] for the Multi-Modal Information based Speech Processing (MISP) 2022 Challenge [37]. The new contributions from this paper are summarized as follows.

- *Unified*: The MIMO-TSVAD framework achieves state-of-the-art performance in the audio-based, video-based, and audio-visual speaker diarization tasks.
- *Flexible*: The proposed framework is jointly designed with an effective multi-stage training strategy. The integrated model can handle various kinds of input features and speaker profiles.
- *Robust*: The proposed framework is jointly designed with an effective multi-stage inference strategy to robustly utilize audio-visual data in modality-missing scenarios.

II. Related Works

A. Modularized Speaker Diarization

The modularized speaker diarization works in a cascaded pipeline. First, a Voice Activity Detection (VAD) [38] module detects active speech in the audio. Next, speech regions are divided into multiple short segments through speech segmentation, such as Speaker Change Detection (SCD) [39] or uniform segmentation [40]. After extracting speaker representations (e.g., i-vectors [41], x-vectors [42]) from those segments, a scoring metric (e.g., cosine distance, probabilistic linear discriminate analysis [43]) measures the pairwise embedding similarities. These segments are finally grouped into different identities by clustering algorithms such as K-Means [3], Agglomerative Hierarchical Clustering (AHC) [40], Spectral Clustering [44], [45], and so on.

The clustering module of the modularized method can flexibly estimate the number of speakers in long audio. However, it typically assumes that each audio segment contains only one speaker. To address this problem, overlapped speech detection [6] and overlap-aware resegmentation [5], [7] can improve the performance by assigning multiple speaker labels to overlapped regions. Speech separation also has been adopted in offline [46]–[48] and online [49] diarization systems.

B. End-to-End Neural Diarization

The EEND framework [8], [9] formulates the diarization problem as a multi-label classification task, relying on the permutation-invariant training to predict all speakers' voice activities simultaneously. The initial EEND models have a fixed number of output speakers limited by the network architecture. Although the Encoder-Decoder based Attractor (EDA) [10], [11] enables EEND-based methods to process audio with a variable number of speakers, the maximum number of speakers is empirically bounded by the training data. To make the number of output speakers flexible and unlimited, EEND-vector clustering (EEND-VC) [12]–[14] integrates end-to-end and clustering approaches, which deploys an EEND model for shortly divided audio blocks and then matches the inter-block speaker labels by clustering on speaker embeddings. In addition, EEND-GLA [15], [16] calculates local attractors from each short block and finds the speaker correspondence based on similarities between inter-block attractors. As its training only requires relative speaker labels within the recording, EEND-GLA is practical for adapting models on in-the-wild datasets without globally unique speaker labels.

Also, several studies promote the EEND-based systems to online inference [16], [50], [51] or improve them in terms of advanced neural network architecture [52]–[55], objective function design [56], [57], unsupervised/semi-supervised learning [58]–[60], and so on.

C. Target-Speaker Voice Activity Detection

The background of TS-VAD can be traced back to the personal VAD [61], which utilizes a given acoustic footprint as the speaker profile to retrieve his or her personalized voice activities. However, the personal VAD only accepts a single speaker at once and ignores the inter-speaker modeling in a conversation. Thus, TS-VAD is designed to process multiple speaker profiles and predict their voice activities simultaneously.

The initial TS-VAD [17] takes speech features (e.g., MFCC) and i-vectors of each speaker as the input, where the output number of speakers is fixed. He et al. [62] modify the model to cope with a flexible number of speakers by determining the maximum number of speakers and outputting null speech activities for zero-padded speaker profiles. Later, LSTM [63] and Transformer [20] modules are implemented along the speaker dimension of models to handle a variable number of speakers. On the other hand, the i-vectors used in TS-VAD are relatively domain-dependent, restricting the system performance on multi-scenario datasets [64]. This finding paves the way for exploring more discriminative speaker embeddings like x-vectors as an alternative. Wang et al. [18] first replace the front-end of TS-VAD with a pre-trained module tailored for extracting frame-level x-vectors. This modification shows superior robustness and generalization than a simple swap of i-vectors for x-vectors in early attempt [17].

Furthermore, the TS-VAD framework has been investigated for multi-channel signal [65], vision-guided system [36], and online inference [66], [67]. Integrating features of both TS-VAD and EEND methods into an entire system has also become a popular trend [20], [55], [68], [69].

D. Audio-Visual Speaker Diarization

As facial attributes and lip movements have been proven to be highly related to speech [70], most early audio-visual speaker diarization methods leverage multi-modal cues by modeling the correspondence between speech signals and talking faces [25], [26]. Also, sound source localization using microphone arrays can establish another cross-modal relationship by mapping the speech direction onto the captured image plane [27], [28]. However, it is sometimes hard to perfectly enroll talking faces or lip movements with related speech segments. Off-screen speakers whose faces are not captured can lead to face detection failure. This "enroll first, diarize later" paradigm may fail in modality-missing scenarios.

Recently, Xu et al. [29] create the AVA Audio-Visual Diarization (AVA-AVD) database containing diverse movie clips and a multi-stage audio-visual speaker diarization system. The proposed method first utilizes VAD module to find speaking segments in the audio, then applies the ASD [34] module to locate a face for each speaking segment. The paired audio-visual inputs are jointly scored to cluster the speaking segments into different identities. Subsequently, Wuerkaixi et al. [30] introduce the lip movement to learn a Dynamic Vision-guided Speaker Embedding (DyViSE) instead of the still facial feature, achieving new state-of-the-art performance on these real-world videos.

In addition, the MISP 2022 [37] database is presented for Chinese Home-TV scenarios. It has over 100 hours of audio-visual signals synchronously captured by different devices, nearly 3.5 times larger than the AVA-AVD [29]. The MISP 2022 Challenge has been successfully held based on this database. Meanwhile, an End-to-End Audio-Visual Speaker Diarization (AVSD) [31] is presented as a competitive baseline method that predicts speech probabilities using audio features and each speaker's lip video. In the AVSD system, available lip videos decide the number and order of output speakers. This way is highly effective when the camera captures all speakers ideally. In contrast, it cannot deal with the out-of-screen speakers.

Apparently, the primary ideas of most existing audio-visual speaker diarization methods originate from the previous audio-based methods. Observing the success of TS-VAD in audio-only speaker diarization and its lack of audio-visual research, this paper further extends the concept of our previous Seq2Seq-TSVAD [19] to build a unified, flexible, and robust framework for audio-visual speaker diarization.

III. Multi-Input Multi-Output Target-Speaker Voice Activity Detection

A. Architecture

Fig. 2 demonstrates our proposed MIMO-TSVAD framework, consisting of three parts: the extractor, the encoder, and the decoder.

1) Audio Extractor: The ResNet-34 [71] is adopted as the audio front-end extractor. The audio signal is first transformed into log Mel-filterbank energies for the model to output a feature map $\in \mathbb{R}^{\mathrm{C}\times\mathrm{T}_1\times\mathrm{H}_1}$, where C, T_1, and H_1 denote the

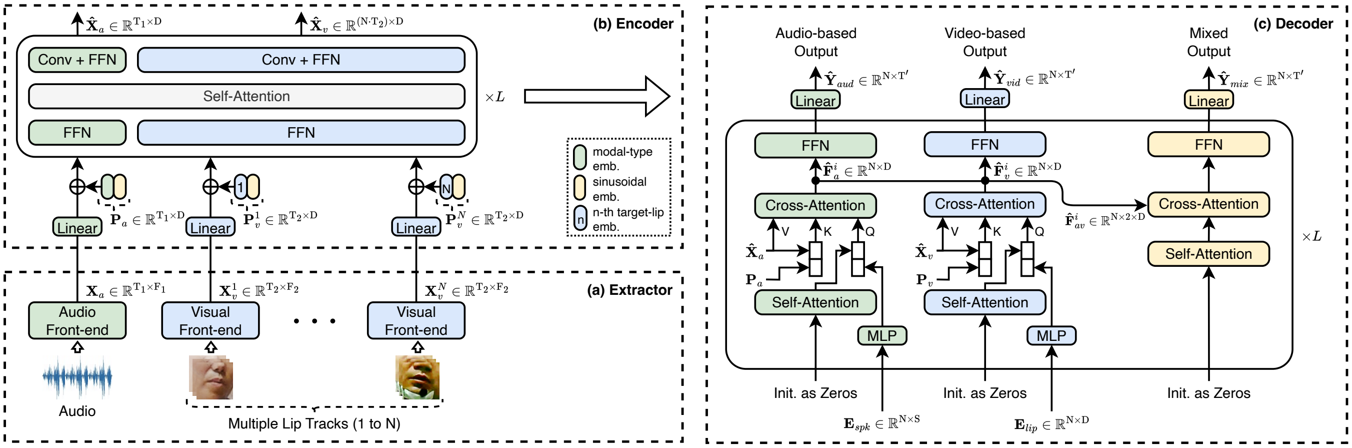

$\hat{\mathbf{X}}_a \in \mathbb{R}^{\mathrm{T}_1 \times \mathrm{D}}$
$\hat{\mathbf{X}}_v \in \mathbb{R}^{(\mathrm{N} \cdot \mathrm{T}_2) \times \mathrm{D}}$
(b) Encoder
Conv + FFN
Conv + FFN
Self-Attention
$\times L$
FFN
FFN
modal-type emb.
sinusoidal emb.
n
n-th target-lip emb.
1
N
Linear
$\mathbf{P}_a \in \mathbb{R}^{\mathrm{T}_1 \times \mathrm{D}}$
Linear
$\mathbf{P}_v^1 \in \mathbb{R}^{\mathrm{T}_2 \times \mathrm{D}}$
Linear
$\mathbf{P}_v^N \in \mathbb{R}^{\mathrm{T}_2 \times \mathrm{D}}$
$\mathbf{X}_a \in \mathbb{R}^{\mathrm{T}_1 \times \mathrm{F}_1}$
$\mathbf{X}_v^1 \in \mathbb{R}^{\mathrm{T}_2 \times \mathrm{F}_2}$
$\mathbf{X}_v^N \in \mathbb{R}^{\mathrm{T}_2 \times \mathrm{F}_2}$
(a) Extractor
Audio Front-end
Visual Front-end
Visual Front-end
Audio
Multiple Lip Tracks (1 to N)
(c) Decoder
Audio-based Output
Video-based Output
Mixed Output
$\hat{\mathbf{Y}}_{aud} \in \mathbb{R}^{\mathrm{N} \times \mathrm{T}'}$
$\hat{\mathbf{Y}}_{vid} \in \mathbb{R}^{\mathrm{N} \times \mathrm{T}'}$
$\hat{\mathbf{Y}}_{mix} \in \mathbb{R}^{\mathrm{N} \times \mathrm{T}'}$
Linear
Linear
Linear
FFN
FFN
FFN
$\hat{\mathbf{F}}_a^i \in \mathbb{R}^{\mathrm{N} \times \mathrm{D}}$
$\hat{\mathbf{F}}_v^i \in \mathbb{R}^{\mathrm{N} \times \mathrm{D}}$
Cross-Attention
Cross-Attention
Cross-Attention
$\hat{\mathbf{F}}_{av}^i \in \mathbb{R}^{\mathrm{N} \times 2 \times \mathrm{D}}$
V
K
Q
V
K
Q
$\hat{\mathbf{X}}_a$
$\mathbf{P}_a$
$\hat{\mathbf{X}}_v$
$\mathbf{P}_v$
Self-Attention
Self-Attention
Self-Attention
$\times L$
MLP
MLP
Init. as Zeros
Init. as Zeros
Init. as Zeros
$\mathbf{E}_{spk} \in \mathbb{R}^{\mathrm{N} \times \mathrm{S}}$
$\mathbf{E}_{lip} \in \mathbb{R}^{\mathrm{N} \times \mathrm{D}}$

Fig. 2. The MIMO-TSVAD framework. (a) Extractor: Front-end modules extract audio features $\mathbf{X}_a$ and video features $\mathbf{X}_v$ from the input data. (b): Encoder: Conformer-based modules leverage multi-modal information to generate the updated audio-visual features $\hat{\mathbf{X}}_a$ and $\hat{\mathbf{X}}_v$. (c) Decoder: Multi-task decoding modules predict voice activities based on different speaker profiles, including target-speaker embeddings $\mathbf{E}_{spk}$, target-lip embeddings $\mathbf{E}_{lip}$, or both. Green, blue, and yellow depict sub-components for processing audio, visual, and audio-visual modalities. For clarity, layer normalization and residual connection of each self-attention, cross-attention, and feedforward layer are omitted in the plot.

number of channels, temporal length, and width. Then, we implement the segmental statistical pooling (SSP) [18] method to aggregate channel-wise features and obtain the audio features $\mathbf{X}_a \in \mathbb{R}^{\mathrm{T}_1 \times \mathrm{F}_1}$, where F_1 is the output dimension of the SSP layer. This process can be viewed as a neural network-based feature extraction that transforms raw audio signals into frame-level representations.

2) Video Extractor: The ResNet18-3D [72] is adopted as the video front-end extractor. A few modifications are employed based on its standard implementation in PytorchVideo[1]. First, the stem layer is set to the convolutional kernel size of 7, stride of 2, and output channels of 32 without the pooling layer. The stride of pooling and convolutional layers in the residual blocks is set to $(1, 2, 2)$ without the temporal downsampling. Finally, a spatial global average pooling layer is placed at the tail. The model transforms a lip video with the length of T_2 and resolution of $H \times W$ into frame-level representations $\mathbf{X}_v \in \mathbb{R}^{\mathrm{T}_2 \times \mathrm{F}_2}$, where F_2 represents the feature dimension.

3) Encoder: We utilize Conformer-based [73] encoder to model long-term and cross-modal relationships between the extracted audio-visual representations. The layout of the designed multi-modal encoder is shown in Fig. 2b, which is mainly stacked by three types of basic modules: the feedforward neural networks, self-attention layers, and convolutional blocks. Inspired by the Mixture of Modality Experts (MoME) in vision-language models [74], the weights of feedforward and convolutional modules are modality-dependent, and shared self-attention layers exchange cross-modal information.

In a TSVAD-like system, the input order of speaker profiles determines the output order of voice activities. Unlike the mixed audio signals, different lip videos are naturally separate tracks to provide features as well as the role of speaker profiles. Hence, we utilize a set of learnable target-lip embeddings to indicate the relative identities of input N lip tracks, denoted as $\mathbf{E}_{lip} \in \mathbb{R}^{\mathrm{N} \times \mathrm{D}}$. Each element $\in \mathbb{R}^{\mathrm{D}}$ in $\mathbf{E}_{lip}$ is repeated to the length of T_2 and added with sinusoidal encodings [75], resulting in the positional embeddings $\mathbf{P}_v^n \in \mathbb{R}^{\mathrm{T}_2 \times \mathrm{D}}$ for the n-th video. The final positional embeddings for all video features can be concatenated as $\mathbf{P}_v \in \mathbb{R}^{(\mathrm{N} \cdot \mathrm{T}_2) \times \mathrm{D}}$. Meanwhile, learnable modality-type embeddings $\mathbf{E}_{aud} \in \mathbb{R}^{\mathrm{D}}$ are initialized to differentiate the audio features from video features. Similarly, $\mathbf{E}_{aud}$ is repeated to the length of T_1 and added with sinusoidal encodings, resulting in the final positional embeddings $\mathbf{P}_a \in \mathbb{R}^{\mathrm{T}_1 \times \mathrm{D}}$ for audio features. Two modality-dependent linear layers map the extracted audio-visual features to the encoder dimension D. Then, the encoder takes the sum of audio-visual features and corresponding positional embeddings as the input. The encoded audio-visual features are denoted as $\hat{\mathbf{X}}_a$ and $\hat{\mathbf{X}}_v$, implying cross-modal information.

4) Decoder: In the Seq2Seq-TSVAD [19], the presented Speaker-Wise Decoder (SW-D) estimates target-speaker voice activities by processing encoded audio features and cross-speaker relationships simultaneously. We further extend its multi-task abilities for inference with different kinds of speaker profiles. As shown in Fig. 2c, the multi-task speaker-wise decoder involves three output branches. The audio-based output (green-colored) performs as same as the previous audio-only TS-VAD, utilizing target-speaker embeddings $\mathbf{E}_{spk}$ as speaker profiles to estimate multiple voice activities from the encoded audio features $\hat{\mathbf{X}}_a$. The video-based output (blue-colored) utilizes the newly introduced target-lip embeddings $\mathbf{E}_{lip}$ to serve as speaker profiles, detecting voice activities from the encoded video features $\hat{\mathbf{X}}_v$. The mixed output (yellow-colored) directly fuses intermediate features of the two modalities by cross-attention mechanism, achieving better usage of their complementary information.

The input of each decoder branch is initialized by zero embeddings $\in \mathbb{R}^{\mathrm{N} \times \mathrm{D}}$ and updated step by step in the subsequent blocks, where N denotes the number of candidate speakers and D represents the decoder dimension. The self-attention layer is adopted to exchange inter-speaker relationships. For the audio-based and video-based branches, $\mathbf{E}_{spk}$ and $\mathbf{E}_{lip}$ go through

[1] https://github.com/facebookresearch/pytorchvideo

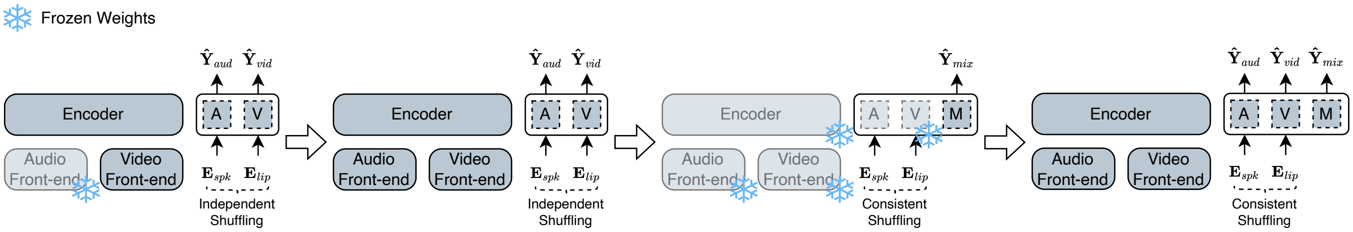


Fig. 3. Multi-stage training process. A, V, and M represent the decoder's audio-based, video-based, and mixed output branches. The predictions from corresponding branches are denoted as $\hat{\mathbf{Y}}_{aud}$, $\hat{\mathbf{Y}}_{vid}$, and $\hat{\mathbf{Y}}_{mix}$, respectively.

multi-layer perception (MLP) modules to align the feature dimension with the decoder dimension D and concatenate with the original queries in respective branch. The adopted MLP module consists of two linear layers with in-between layer normalization and ReLU activation. Meanwhile, we concatenate keys from $\hat{\mathbf{X}}_a$ and $\hat{\mathbf{X}}_v$ with corresponding positional embeddings $\mathbf{P_a}$ and $\mathbf{P_v}$. This way, all key-query calculations in the cross-attention layer can incorporate positional and speaker-related information as the auxiliary feature. Additionally, let $\hat{\mathbf{F}}_a^i \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}}$ and $\hat{\mathbf{F}}_v^i \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}}$ denote the intermediate features of audio-based and video-based branches in the i-th decoder block, respectively. Then, they are concatenated to produce the audio-visual features $\hat{\mathbf{X}}_{av} \in \mathbb{R}^{\mathrm{N}\times 2\times\mathrm{D}}$. For the mixed branch, it takes the second (modality) dimension of $\hat{\mathbf{X}}_{av}$ as the sequence axis to perform the cross-attention operation. This way, two modalities are dynamically fused into the mixed branch. Finally, a linear projection layer with sigmoid activation transforms the decoder embeddings into posterior probabilities of the estimated voice activities. The output dimension T' of the linear projection controls the temporal resolution of system prediction. For example, if the chunk length of input audio-visual data is fixed at L_{chunk} seconds. The prediction for this duration will be evenly divided into T' frames. Each frame-level output indicates whether the target speaker is speaking during the corresponding time interval. In this case, the temporal resolution can be calculated as L_{chunk}/T', representing the unit duration of each frame-level prediction. By adjusting the linear projection layer, T' can be easily set as several multiples of T_1 or T_2. Given N speakers, the predictions from the audio-based, video-based, and mixed output are represented as $\hat{\mathbf{Y}}_{aud}$, $\hat{\mathbf{Y}}_{vid}$, and $\hat{\mathbf{Y}}_{mix} \in \mathbb{R}^{\mathrm{N}\times\mathrm{T}'}$, respectively.

It is worth noting that the original version of SW-D in Seq2Seq-TSVAD [19] also introduces target-speaker and positional embeddings into the self-attention layers. This operation is found to be unnecessary in this work as the decoder still works well after removing it.

B. Multi-Stage Training

To process multiple input and output types in a unified framework, we design a multi-stage training strategy to optimize the model progressively, as shown in Fig. 3. Meanwhile, we introduce two categories of modality masking techniques during training, which enable the model to process audio-only, video-only, or audio-visual signals flexibly. The training process can be described as follows.

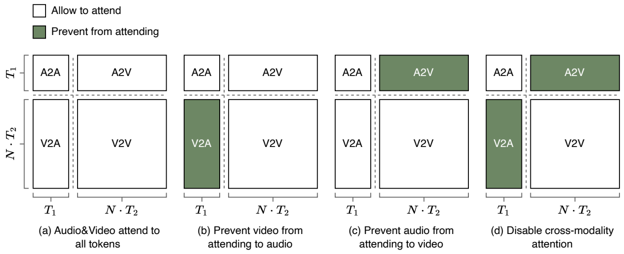

Allow to attend
Prevent from attending
T_1
$N \cdot T_2$
A2A
A2V
V2A
V2V
T_1
$N \cdot T_2$
(a) Audio&Video attend to all tokens
(b) Prevent video from attending to audio
(c) Prevent audio from attending to video
(d) Disable cross-modality attention

Fig. 4. Different attention masks in the encoder. The first T_1 tokens of the input feature sequence are the audio features. The rest $N \cdot T_2$ tokens are video features extracted from N lip tracks.

- *Stage 1*: We copy and freeze the parameters of the pre-trained speaker embedding model to initialize the audio front-end extractor. Only audio and video-based output branches are adopted to train on fully simulated data.
- *Stage 2*: The audio front-end model is unfrozen. Training data for model adaption on the specific dataset is added at a given ratio.
- *Stage 3*: All pre-trained parameters are frozen. The mixed output branch is initialized and trained separately.
- *Stage 4*: Finally, all parameters are unfrozen to be fine-tuned jointly.

1) Model-Level Modality Masking: In training *stages 1-2*, the mixed output branch in the decoder is not initialized. Cross-modal information is only exchanged by self-attention layers in the encoder. Hence, we implement different attention masks to adapt the model to different input features. Fig. 4 demonstrates four possible cases, reflecting that the information can be bidirectional, one-way, or prohibited from flowing between two modalities. During training, one of the masks is randomly selected. During inference, the attention mask is generated according to the existence of each modality. In other words, tokens should only attend to the existing modalities. To decouple two output branches to avoid learning shortcuts from matched target-speaker and target-lip embeddings, the speaker order of $\mathbf{E}_{lip}$ should be shuffled independently from the $\mathbf{E}_{spk}$. As $\mathbf{E}_{lip}$ represents the relative identities, shuffling $\mathbf{E}_{lip}$ can be done equivalently by shuffling the input order of lip tracks in practice. Ground truth labels for different output branches must be re-assigned based on their shuffled results.

Given an audio-visual recording with N existing speakers, except target-lip embeddings $\mathbf{E}_{lip}$ as learnable parameters built in the model, the other inputs involve audio signal $\mathbf{A}$, target-speaker embeddings $\{\mathbf{e}_n \mid 1 \le n \le N\}$, and lip tracks

$\{\mathbf{l}_n \mid 1 \leq n \leq N\}$. Each $\mathbf{e}_n$ and $\mathbf{l}_n$ represents the n-th speaker embedding and lip track, respectively. The ground truth labels for voice activities can be denoted as a binary matrix $\mathbf{Y} \in (0,1)^{\mathrm{N}\times\mathrm{T}'}$, where $\mathbf{Y}(n,t)$ represents the speaking existence of the n-th speaker at time t. The audio-based output $\hat{\mathbf{Y}}_{aud}$ and video-based output $\hat{\mathbf{Y}}_{vid}$ are modeled with our proposed MIMO-TSVAD as follows:

$$\mathbf{E} = sh_1\left(\{\mathbf{e}_n \mid 1 \leq n \leq N\}\right), \tag{1}$$

$$\mathbf{L} = sh_2\left(\{\mathbf{l}_n \mid 1 \leq n \leq N\}\right), \tag{2}$$

$$\hat{\mathbf{Y}}_{aud}, \hat{\mathbf{Y}}_{vid} = \text{MIMO\_TSVAD}\left(\mathbf{A}, \mathbf{E}, \mathbf{L}\right), \tag{3}$$

where $sh_1(\cdot)$ and $sh_2(\cdot)$ represent two independent operations for shuffling speaker-orders (n) in $\mathbf{E}$ and $\mathbf{L}$, respectively. The audio-based output loss $\mathcal{L}_{aud}$ and video-based output loss $\mathcal{L}_{vid}$ are described as:

$$\mathcal{L}_{aud} = BCE\left(sh_1\left(\mathbf{Y}\right), \hat{\mathbf{Y}}_{aud}\right), \tag{4}$$

$$\mathcal{L}_{vid} = BCE\left(sh_2\left(\mathbf{Y}\right), \hat{\mathbf{Y}}_{vid}\right), \tag{5}$$

where ground truth labels $\mathbf{Y}$ are re-assigned based on the same shuffling operations. $BCE\left(y, \hat{y}\right)$ measures the binary cross-entropy between the target y and predicted $\hat{y}$. This way, the trained encoder can handle different input features.

2) Data-Level Modality Masking: In training *stages 3-4*, the mixed output branch in the decoder is additionally introduced to combine the audio-based and video-based output, which requires that speaker orders of $\mathbf{E}_{spk}$ and $\mathbf{E}_{lip}$ are consistent. However, in realistic scenarios, it is difficult for each speaker to obtain perfectly paired target-speaker embedding and lip movement all the time. Hence, we implement different data masks to adapt the model to uncertain speaker profiles. During training, each speaker has a probability of 0.5 to conduct data masking. Once the n-th speaker is selected, either the speaker embedding $\mathbf{e}_n$ or lip track $\mathbf{l}_n$ will be masked by zeros. Let $\mathbf{M}_{spk}^{in} = \{a_n \mid a_n \in \{0,1\}, 1 \leq n \leq N\}$ and $\mathbf{M}_{spk}^{out} \in (0,1)^{\mathrm{N}\times\mathrm{T}'}$ record the masking states of target-speaker embeddings. If the n-th speaker embedding is masked, $a_n \in \mathbf{M}_{spk}^{in}$ and $\mathbf{M}_{spk}^{out}(n,:)$ will be set to zeros. Similarly, the masking states of lip tracks are denoted as $\mathbf{M}_{lip}^{in} = \{b_n \mid b_n \in \{0,1\}, 1 \leq n \leq N\}$ and $\mathbf{M}_{lip}^{out} \in (0,1)^{\mathrm{N}\times\mathrm{T}'}$. If the n-th lip track is masked, $b_n \in \mathbf{M}_{lip}^{in}$ and $\mathbf{M}_{lip}^{out}(n,:)$ will be set to zeros. In this situation, the proposed model output can be obtained as follows:

$$\mathbf{E}' = sh\left(\{a_n \times \mathbf{e}_n \mid 1 \leq n \leq N\}\right), \tag{6}$$

$$\mathbf{L}' = sh\left(\{b_n \times \mathbf{l}_n \mid 1 \leq n \leq N\}\right), \tag{7}$$

$$\hat{\mathbf{Y}}'_{aud}, \hat{\mathbf{Y}}'_{vid}, \hat{\mathbf{Y}}'_{mix} = \text{MIMO\_TSVAD}\left(\mathbf{A}, \mathbf{E}', \mathbf{L}'\right), \tag{8}$$

where $sh(\cdot)$ represents the consistent speaker-order shuffling operation. Each input $\mathbf{e}_n$ and $\mathbf{l}_n$ will be retained or zeroed by multiplying by the $a_n \in \{0,1\}$ and $b_n \in \{0,1\}$. The audio-based and video-based output losses are modified as follows:

$$\mathcal{L}'_{aud} = BCE\left(sh(\mathbf{M}_{spk}^{out} \wedge \mathbf{Y}), \hat{\mathbf{Y}}'_{aud}\right), \tag{9}$$

$$\mathcal{L}'_{vid} = BCE\left(sh(\mathbf{M}_{lip}^{out} \wedge \mathbf{Y}), \hat{\mathbf{Y}}'_{vid}\right), \tag{10}$$

where $\wedge$ denotes the element-wise logical AND operation. Ground truth labels $\mathbf{Y}$ are not only re-assigned by speaker-order shuffling but also filtered by the masking states. Then, the mixed output loss $\mathcal{L}'_{mix}$ is written by:

$$\mathcal{L}'_{mix} = BCE\left(sh(\left(\mathbf{M}_{spk}^{out} \vee \mathbf{M}_{lip}^{out}\right) \wedge \mathbf{Y}), \hat{\mathbf{Y}}'_{mix}\right), \tag{11}$$

where $\vee$ denotes the element-wise logical OR operation.

Since the masking probability is independent for all speakers, speaker embeddings or lip tracks of multiple speakers may be masked at a time. Because for the same speaker, we can only randomly mask either the speaker embedding or lip track simultaneously. If the n-th speaker is masked among all N inputs, the model still generates an output of shape $N \times T'$. As long as one of the $\mathbf{e}_n$ or $\mathbf{l}_n$ is available, the mixed output branch of the proposed model can work properly for the n-th speaker. This way, the trained decoder can be compatible with uncertain speaker profiles.

To summarize, the first two stages utilize the model-level modality masking to address different input features (e.g., audio-only, video-only, or audio-visual), which are optimized by the total diarization loss of $\mathcal{L}_{aud} + \mathcal{L}_{vid}$. Meanwhile, it prevents the model from receiving many zero-masked features caused by the data-level modality masking in early training. The last two stages utilize the data-level modality masking to solve uncertain inference conditions with modality-mismatched speaker profiles (e.g., target-speaker embeddings, target-lip embeddings, or mixed), which are optimized by the total diarization loss of $\mathcal{L}'_{aud}+\mathcal{L}'_{vid}+\mathcal{L}'_{mix}$. Finally, the trained model can flexibly deal with varying accessibilities of audio-visual data.

C. Multi-Stage Inference

Given the synchronized audio-visual data, the algorithm aims to find all speaker identities and localize their speaking regions automatically. To adequately utilize audio-visual data in modality-missing scenarios, we design a multi-stage inference strategy to predict voice activities iteratively.

1) Prior Steps: Following the commonly used paradigm in previous TS-VAD methods [17]–[20], a modularized diarization system is necessary to obtain an initial result. Each detected speaker's non-overlapped speech segments are aggregated to extract the target embedding, initializing the MIMO-TSVAD model to conduct audio-based speaker diarization.

For the video signals, we extract each speaker's lip track as the same as our previous works [36], [76]. The RetinaFace [77] detector is deployed to localize face images with five-point facial landmarks in each video frame. As a talking face is assumed not to move dramatically in a short time window, the K-Means algorithm in Scikit-Learn toolkit [78] utilizes coordinates of detected faces to cluster the same person in adjacent frames. After obtaining each speaker's face track, we crop the lip region of interest (Lip-RoI) based on an empirical setting reported in the CAS-VSR-W1k [79] database for large-scale lip reading tasks. Let $\mathbf{p_1}$, $\mathbf{p_2}$, and $\mathbf{p_3}$ represent

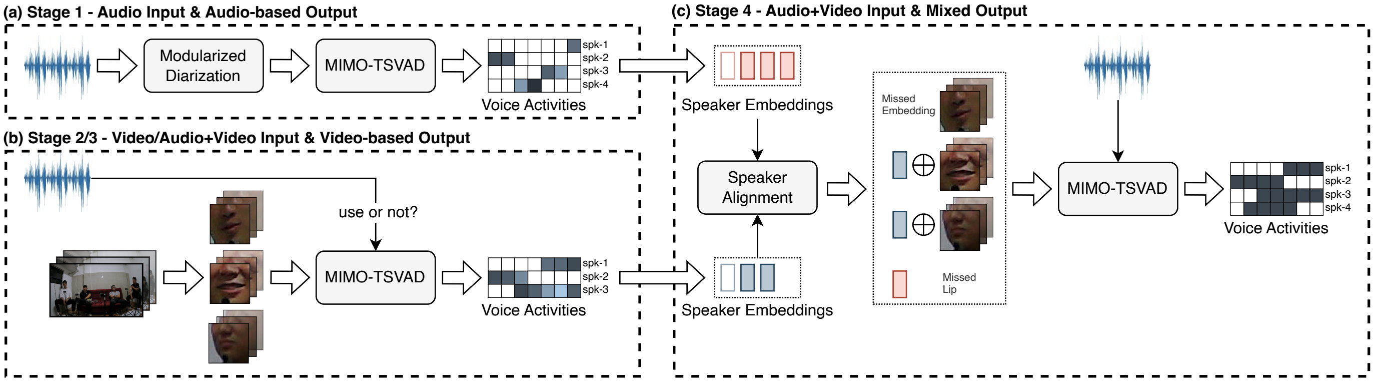


Fig. 5. Multi-stage inference process. The MIMO-TSVAD model is weight-shared in all stages. (a) The MIMO-TSVAD model is a typical audio-based system in inference *stage 1*. (b) The MIMO-TSVAD decoder is based on the video-based output. When the encoder takes video-only input, it operates in inference *stage 2*. Otherwise, it operates in inference *stage 3* if the audio input is also available. (c) The speaker alignment module finds off-screen speaker embedding to build the new set of mixed speaker profiles. Then, the MIMO-TSVAD model operates in inference *stage 4* to take audio-visual input and mixed output. In this example, the first speaker embedding is not successfully extracted because his non-overlapped speech duration is too short. Also, the last speaker's lip track is undetected, but the speaker alignment module replenishes his speaker embedding (pink-colored).

coordinates of the nose tip, left mouth corner, and right mouth corner. The Lip-RoI bounding box is defined as:

$$x_{center}, y_{center} = \frac{\mathbf{p_2} + \mathbf{p_3}}{2}, \tag{12}$$

$$width = min\left\{3.2 \times d_{MN}, 2 \times max\left\{d_{MN}, d_{p2p3}\right\}\right\}, \tag{13}$$

where d_{MN} denotes the Euclidean distance between the nose tip and the center of the mouth, d_{p2p3} denotes the Euclidean distance between $\mathbf{p_2}$ and $\mathbf{p_3}$. Furthermore, undetected frames of each lip track will be padded by zeros, guaranteeing audio-visual synchronization. As this work mainly focuses on the TS-VAD research, we adopt this simple detection and tracking method to extract multiple speakers' lip tracks.

2) *Diarization:* The MIMO-TSVAD framework requires at least two input parts: features and speaker profiles. The features (e.g., audio, video, or audio-visual data) imply rich voice activity information in conversations. The speaker profiles (e.g., target-speaker embeddings, target-lip embeddings) are the reference to separate individual speakers' corresponding voice activities from the provided features. As mentioned in Section I, the MIMO-TSVAD framework supports four cases according to the accessibility of different input data, which can be described as a multi-stage inference process.

- *Stage 1*: The model takes audio features $\mathbf{X}_a$ and target-speaker embeddings $\mathbf{E}_{spk}$, which is an audio-only system.
- *Stage 2*: The model takes video features $\mathbf{X}_v$ and target-lip embeddings $\mathbf{E}_{lip}$, which is a video-only system.
- *Stage 3*: The model takes audio-visual features $\mathbf{X}_a + \mathbf{X}_v$ and target-lip embeddings $\mathbf{E}_{lip}$. The only difference with the *stage 2* is using auxiliary audio features.
- *Stage 4*: The model takes audio-visual features $\mathbf{X}_a + \mathbf{X}_v$ and the mix of target-speaker and target-lip embeddings $\mathbf{E}_{spk} + \mathbf{E}_{lip}$. Each speaker's target speech activities can be extracted if at least one kind of speaker profile exists.

Fig. 5 illustrates the overall inference process for an audio-visual recording. First, the audio signal undergoes the modularized diarization system to obtain initial speaker profiles. Then, the inference *stage 1* of MIMO-TSVAD predicts voice activities based on the audio-based output. This method ensures an audio-based solution is still available even if all visual information is lost. For lip tracks extracted from the video signal, the inference *stage 2* of MIMO-TSVAD predicts voice activities based on the video-based output. If the audio signal is fed as an auxiliary feature, this stage becomes the inference *stage 3*. The inference *stages 2-3* of MIMO-TSVAD directly utilizes the input lip videos to serve as speaker profiles, which does not require an additional module like the modularized method in inference *stage 1*. However, a critical problem is that they can not detect the existence of completely off-screen speakers. To solve the problem, a Speaker Alignment (SA) module is deployed to find the undetected off-screen speaker embeddings. We first extract speaker embeddings based on the predicted voice activities from the inference *stage 1* and *stage 3*, respectively. By pairwise similarity measurement between two embedding sets, the Hungarian algorithm [80] is employed to obtain matching relationships with the highest cosine scores. The off-screen speaker embeddings from the audio-based output can be discovered if they are not successfully matched with a pairwise similarity score higher than the pre-set speaker verification threshold. Then, the matched speaker embeddings are averaged, and the unmatched off-screen speaker embeddings are replenished. Finally, each available speaker profile can be unimodal or bimodal. The inference *stage 4* of MIMO-TSVAD predicts voice activities based on the mixed output, tackling complex modality-missing situations.

During the whole inference, the pre-trained MIMO-TSVAD model is able to adopt shared weights for conducting different inference stages flexibly. All available data can be fully exploited to boost the final diarization performance, whether viewed from the aspects of audio-visual features or complementary speaker profiles.

IV. Experimental Settings

A. Datasets

The proposed MIMO-TSVAD framework is a multi-modal extension of our previous work [19], which supports audio-based, video-based, and audio-visual speaker diarization. We first verify its advantages on the VoxConverse [81] and DIHARD-III [82] datasets for audio-based speaker diarization. The VoxConverse dataset is an in-the-wild dataset with 20.29 hours of development set and 43.53 hours of test set collected from YouTube. We select the first 172 recordings (80%) of the original development set for training MIMO-TSVAD models. The last 44 recordings (20%) remain for validation. The DIHARD-III dataset is a multi-domain dataset with 34.15 hours of development set and 33.01 hours of evaluation set, including 11 complex scenarios (e.g., interview, clinical, restaurant). We select the first 203 recordings (80%) of the original development set for training MIMO-TSVAD models. The last 51 recordings (20%) remain for validation.

In addition, the MISP 2022 [37] dataset is adopted for audio-visual speaker diarization. The MISP 2022 dataset targets the Chinese home-TV scenario with 106.09 hours of training set, 2.6 hours of development set, and 3.12 hours of evaluation set. Audio-visual signals are synchronously captured by different devices. The single-channel microphone records the near-field ($< 0.5m$) audio. The 2-channel microphone array and high-definition camera capture the middle-field ($1 - 1.5m$) data. The 6-channel microphone array and wide-angle camera capture the far-field ($3 - 5m$) data. Also, we utilize the NARA-WPE [2] and SETK [3] toolkits to do dereverberation and MVDR-based beamforming for the multi-channel audio signals, augmenting the diversity of training data. Under the MISP challenge rules, all fields can be used in model training and validation. Only far-field data is allowed for evaluation.

We also create the simulated audio-visual dataset to benefit neural networks from large-scale training data. The VoxCeleb2 [83] dataset is adopted as the source audio corpus, which covers thousands of unique speaker identities. As the video contents of the VoxCeleb2 dataset are not available for downloading nowadays, we utilize the MISP 2022 [37] dataset as a supplementary video corpus. Although the audio and video signals are not semantically consistent regarding the dialogue content, the combination is still feasible because speaker diarization primarily focuses on recognizing speaker identities and voice activities rather than the speech content. During training, each simulated data is generated in an on-the-fly manner as follows.

- *Step 1*: A speaker label is randomly selected from the audio corpus. The single-speaker utterance is created by alternately concatenating his or her source speech and silent (zero-padded) segments, where each segment length is sampled from a uniform distribution of 0-4 seconds.
- *Step 2*: We cut the video corpus into active (speech) and inactive (silent) lip tracks. An active lip track is randomly selected and cropped into the same duration for each speech segment created in the first step. An inactive lip track is randomly selected and cropped into the same duration for each silent segment. Then, the extracted lip tracks are concatenated in the same order as the audio, creating the corresponding single-speaker lip track.
- *Step 3*: For each simulation, audio-visual data of 1-4 speakers can be obtained by repeating the above steps. All single-speaker utterances are averaged as the mixed audio. No additional operations are required for the lip tracks, as they are already separate. The average overlap ratio of simulated data is estimated to be around 30%.

B. Network Configuration

1) Speaker Embedding Extraction: The ResNet-34 is utilized as the pattern extractor, where residual blocks have widths (number of channels) of $\{64, 128, 256, 512\}$ with a total downsampling factor of 8. After adding the statistical pooling [42] layer, a linear projection layer outputs the 256-dim speaker embedding. The ArcFace ($s = 32, m = 0.2$) [84] is used as the classifier. The model trained on the VoxCeleb2 [83] dataset achieves an equal error rate (EER) of 0.814% on the Vox-O [85] trial. Other training details are the same as [86]. The pre-trained model is used for speaker embedding extraction in the subsequent MIMO-TSVAD system. Also, the Speaker Alignment (SA) module employs the pre-trained speaker verification threshold of 0.3479 during inference.

2) MIMO-TSVAD: The audio front-end ResNet-34 is initialized by the pre-trained speaker embedding model. We replace the original statistical pooling (SP) with SSP [18] layer to achieve frame-level feature extraction. The video front-end ResNet18-3D has residual blocks with widths of $\{32, 64, 128, 256\}$ to output 256-dim frame-level features, which are randomly initialized. The back-end encoder-decoder modules have 6 blocks sharing the same settings: 512-dim attentions with 8 heads and 1024-dim feed-forward layers with a dropout rate of 0.1. The kernel size of convolutions in Conformer blocks is 15, and the other implementation details are the same as [73].

C. Training and Inference Details

All training data is split into fixed-length chunks by sliding window and normalized with a mean of 0 and a standard deviation of 1. The chunk length can be set to different values (e.g., 8 seconds, 16 seconds, 32 seconds). As the model only accepts a fixed-length chunk as input, the chunk lengths used for training and inference should always be consistent, but the chunk shift can vary flexibly. The subsequent experimental results investigate the impacts of different chunk lengths and shifts. Then, the input acoustic features are 80-dim log Mel-filterbank energies with a frame length of 25 ms and a frameshift of 10 ms. The input lip videos are transformed into grayscale with a resolution of 88×88 and frames per second (FPS) of 25. Assume that the maximum number of speaker profiles (speaker capacity) is set to C. When speakers in a recording cannot reach C, empty speaker profiles are padded

[2] https://github.com/fgnt/nara_wpe

[3] https://github.com/funcwj/setk

TABLE I
PERFORMANCE OF DIFFERENT MIMO-TSVAD MODELS ON THE VOXCONVERSE TEST SET (COLLAR = 250 MS). THE DIARIZATION ERROR RATE (DER) IS THE SUM OF MISS (MI), FALSE ALARM (FA), AND SPEAKER CONFUSION (SC) RATES.

ID	**VAD Resolution**	**SPK Capacity**	**Chunk Length/Shift**	**1-10 SPKs (%)**				**10+ SPKs (%)**				**Total (%)**			
				MI	**FA**	**SC**	**DER**	**MI**	**FA**	**SC**	**DER**	**MI**	**FA**	**SC**	**DER**
S1	80ms	10	8s / 8s	2.43	1.48	1.10	5.01	2.80	10.24	3.44	16.48	2.51	3.27	1.58	7.36
S2		20		1.99	1.35	1.05	4.39	2.10	2.52	3.57	8.19	2.01	1.59	1.56	5.16
S3		30		2.05	1.29	1.00	4.34	2.20	1.74	3.56	7.50	2.08	1.38	1.52	**4.98**
S4	10ms	10	8s / 8s	2.27	1.48	1.00	4.75	2.73	10.08	3.36	16.17	2.36	3.24	1.49	7.09
S5		20		1.93	1.32	1.02	4.27	2.12	2.41	3.53	8.06	1.97	1.54	1.54	5.05
S6		30		2.02	1.34	1.01	4.37	2.17	1.72	3.60	7.49	2.05	1.42	1.55	**5.02**
S7	10ms	30	16s / 16s	2.05	1.20	0.96	4.21	2.26	1.82	3.48	7.56	2.09	1.33	1.48	4.90
S8			32s / 32s	2.04	1.16	0.91	4.11	2.35	1.67	3.34	7.36	2.10	1.26	1.40	4.76
S9			64s / 64s	1.87	1.15	0.86	3.88	2.26	1.66	3.15	7.07	1.94	1.26	1.33	**4.53**
S10	10ms	30	16s / 2s	1.95	1.02	0.85	3.82	2.22	1.39	3.28	6.89	2.01	1.10	1.35	4.46
S11			32s / 2s	1.95	0.99	0.79	3.73	2.28	1.26	3.15	6.69	2.01	1.05	1.27	4.33
S12			64s / 2s	1.85	1.02	0.74	3.61	2.21	1.30	2.96	6.47	1.92	1.07	1.19	**4.18**

by zeros or speakers not appearing in the current chunk. This padding method aligns the dimension of batched training data and forces the model to distinguish valid and invalid speaker profiles. Meanwhile, all the input speaker profiles should be shuffled to make the model invariant to speaker order.

We implement the BCE loss and Adam optimizer to train the neural network. As described in Section III-B, the first training stage starts with a linear learning rate warm-up from 0 to *1e-4* in 2,000 iterations. From the second training stage, real data from the VoxConverse [81] and DIHARD-III [82] datasets is added to the simulated data at a ratio of 0.2. For experiments on the MISP 2022 [37] dataset, this ratio is adjusted to 0.5, which follows our previous challenge setting [36] to mitigate the large domain gap between simulated data (English) and test set (Chinese). In the last finetuning stage, the learning rate is decayed to *1e-5*. Additive noise from Musan [87] and reverberation from RIRs [88] are applied for audio augmentation. For video augmentation, input lip videos undergo each item of the following procedures with a probability of 0.5: rotation with an angle range $[5, 20]$; horizontal flipping; cropping with the scale range $[0.8, 1]$; transformation of contrast, brightness, and saturation in the range $[-25, 25]$. The training process takes around 200k iterations with a batch size of 32 on 8 × NVIDIA RTX-3090 GPUs.

The inference process follows the Section III-C. The prior steps extract lip tracks and target embeddings of speakers with non-overlapped speech longer than 2 seconds. If the speakers are less than speaker capacity C, empty speaker profiles are padded by zeros. Otherwise, the excess speaker profiles are inferred in the next group. We split test data into chunks to feed each MIMO-TSVAD inference stage. The predictions are stitched chunk by chunk. For the far-field recordings in the MISP 2022 dataset, we directly average the predictions from all dereverberated channels. Lastly, reference VAD can revise the diarization results if the specific evaluation metric allows. The timestamps marked as active speech will assign a positive label to the speaker with the highest predicted score. The predictions at non-speech timestamps will be zeroed. All experiments are repeated three times to report the mean values.

TABLE II
COMPARISONS OF MIMO-TSVAD MODELS WITH OTHERS ON THE VOXCONVERSE TEST SET (COLLAR = 250 MS).

Method	**DER (%)**
ByteDance [89] †	5.17
DKU-DukeECE [23] ‡	4.94
Microsoft [20]	4.57
PET-TSVAD [90]	4.35
pyannote.audio [91] ≀	4.00
LSTM-SC [18]	6.33
VBx [5]	5.62
AHC [23]	5.35
+ MIMO-TSVAD (S12 in Table I) *	**4.18**

† VoxSRC-21 2nd-ranked result with 3-system fusion. The 1st-ranked team dose does not report the DER for this set.
‡ VoxSRC-22 1st-ranked result with 4-system fusion.
≀ VoxSRC-23 2nd-ranked result with WavLM [92] model.
* participates in the VoxSRC-23 1st-ranked winning (fusion) system.

V. RESULTS AND DISCUSSIONS

A. Evaluation of Audio-based Diarization

When only loading audio-related modules (green-colored in Fig. 2), the inference *stage 1* of MIMO-TSVAD is equivalent to the basic Seq2Seq-TSVAD [19]. Using the pre-trained speaker embedding model, we implement the AHC [23], LSTM-SC [18], and VBx [5] as modularized diarization methods to extract initial speaker profiles. The related hyper-parameters are tuned on the development sets of VoxConverse [81] and DIHARD-III [82] datasets, respectively.

For the VoxConverse dataset, the AHC obtains a Diarization Error Rate (DER) of 5.35% on the test set with a tolerance collar of 250 ms, better than the VBx (5.62%) and LSTM-SC (6.33%). Thus, Table I illustrates the performance of different MIMO-TSVAD models initialized by the AHC result. Two types of VAD resolutions (duration per frame-level prediction) are provided as coarse (80 ms) and precise (10 ms) options, which can be implemented easily by adjusting the dimension of the model output. As the dataset has up to 21 speakers in a single recording, we choose the speaker capacity (maximum speaker embeddings) of 10, 20, and 30 to cover insufficient,

TABLE III
PERFORMANCE OF DIFFERENT MIMO-TSVAD MODELS ON THE DIHARD-III EVALUATION SET (REFERENCE VAD). THE DIARIZATION ERROR RATE (DER) IS THE SUM OF MISS (MI), FALSE ALARM (FA), AND SPEAKER CONFUSION (SC) RATES.

ID	VAD Resolution	SPK Capacity	Chunk Length/Shift	1-5 SPKs (%)				5+ SPKs (%)				Total (%)			
				MI	FA	SC	DER	MI	FA	SC	DER	MI	FA	SC	DER
S1	80ms	5	8s / 8s	4.58	3.17	3.21	10.96	10.92	2.80	9.56	23.28	5.47	3.11	4.10	12.68
S2		10		4.61	3.08	3.17	10.86	10.58	2.58	9.23	22.39	5.45	3.01	4.03	**12.49**
S3		20		4.60	3.19	3.14	10.93	10.70	2.47	9.04	22.21	5.45	3.09	3.97	12.51
S4	10ms	5	8s / 8s	3.80	2.39	3.30	9.49	10.17	2.74	9.87	22.78	4.69	2.44	4.22	11.35
S5		10		3.76	2.39	3.22	9.37	10.29	2.02	9.40	21.71	4.68	2.34	4.09	11.11
S6		20		3.94	2.23	3.19	9.36	10.49	1.83	9.19	21.51	4.86	2.18	4.04	**11.08**
S7	10ms	20	16s / 16s	3.64	2.42	3.07	9.13	10.21	1.88	8.83	20.92	4.56	2.35	3.88	10.79
S8			32s / 32s	3.55	2.39	2.87	8.81	9.99	1.88	8.86	20.73	4.45	2.31	3.71	**10.47**
S9			64s / 64s	3.49	2.63	2.89	9.01	9.57	2.25	8.91	20.73	4.34	2.58	3.73	10.65
S10	10ms	20	16s / 2s	3.60	2.23	2.88	8.71	10.32	1.59	8.35	20.26	4.54	2.14	3.65	10.33
S11			32s / 2s	3.51	2.23	2.72	8.46	10.15	1.56	8.43	20.14	4.44	2.14	3.52	**10.10**
S12			64s / 2s	3.47	2.44	2.74	8.65	9.76	1.76	8.38	19.90	4.35	2.35	3.53	10.23

suitable, and sufficient capacities, respectively. Systems S1-3 reveal that total DERs reduce obviously from 7.36% to 4.98% when the speaker capacity increases from 10 to 30. The gain mainly comes from the subset of recordings with over 10 speakers. Systems S4-6 show that the precise resolution (10 ms) slightly improves over a coarse resolution (80 ms), except System S6. It is speculated that the tolerance collar makes the evaluation insensitive to utterance boundaries. Based on the VAD resolution of 10 ms and speaker capacity of 30, Systems S7-9 increase the chunk length of model training to decrease total DERs from 4.90% to 4.53%. Systems S10-12 utilize a 2-second chunk shift for inference. Beyond simply stitching chunk-wise predictions, overlapped regions are averaged as score-level fusion. It can be seen that dense inference can further decrease the lowest total DER to 4.18%. Table II compares our proposed method with the current state-of-the-art results. Our best performance significantly outperforms previous ones, especially for some multi-system fusion results in early VoxSRC challenges. Notably, System S12 participates in the latest VoxSRC-23 winning system [24]. Although the pyannote.audio [91] team achieves a DER of 4.00% based on large-scale data and unsupervised WavLM models, it does not beat our winning system on the VoxSRC-23 challenge set.

TABLE IV
COMPARISONS OF MIMO-TSVAD MODELS WITH OTHERS ON THE DIHARD-III EVALUATION SET (REFERENCE VAD).

Method	DER (%)
Hitachi-JHU [93] †	11.58
USTC-NELSLIP [21] ‡	11.30
Wang et al. [64]	11.30
ANSD-MA-MSE [94]	11.12
AHC [23]	16.77
VBx [5]	16.58
LSTM-SC [18]	15.40
+ MIMO-TSVAD (S11 in Table III)	**10.10**

† DIHARD-III 2nd-ranked result with 5-system fusion.
‡ DIHARD-III 1st-ranked result with 5-system fusion.

For the DIHARD-III dataset, reference VAD is provided according to the challenge's track 1 rules. The LSTM-SC obtains the 15.40% DER on the evaluation set, better than the VBx (16.58%) and AHC (16.77%). Thus, Table III illustrates the performance of different MIMO-TSVAD models initialized by the LSTM-SC result. Since the number of speakers in each recording is up to 9, we explore the speaker capacity of 5, 10, and 20. The experimental results show similar conclusions to Table I, which means the precise VAD resolution, larger speaker capacity, longer chunk length, and dense inference can usually enhance the diarization performance. Nevertheless, two phenomena should be mentioned. First, the benefits of precise VAD resolution are more significant than the maximum speaker capacity here. As the DIHARD-III dataset is annotated at 10 ms and evaluated without the tolerance collar, its results are more likely affected by the temporal precision of estimated speech activities. Second, the diarization performance saturates when the chunk length grows to 32 seconds in System S8 and S11, which reveals that the performance cannot be further improved easily. Finally, System 11 achieves the lowest total DER of 10.10%. Table IV compares our proposed method with the current state-of-the-art results. Our best performance demonstrates superiority over existing approaches.

Notably, the DER results of MIMO-TSVAD models in Tables I and III differ slightly from those reported in the original Seq2Seq-TSVAD paper [19]. A clarification of the inference configurations is provided as follows.

- First, for the VoxConverse dataset, this work does not apply any post-processing based on estimated VAD information, whereas [19] revises TS-VAD predictions using results from a separate diarization system. In this case, we observed that a finer-grained inference strategy, achieved by adjusting chunk lengths and chunk shifts, already leads to significant DER improvements. As a result, the best-performing MIMO-TSVAD model achieves competitive accuracy without requiring additional VAD-based refinement.
- Second, for the DIHARD-III dataset, [19] directly reports the best DER under a specific configuration (chunk length = 16 s, VAD resolution = 10 ms, speaker capacity =

TABLE V
PERFORMANCE OF DIFFERENT MIMO-TSVAD MODELS ON THE MISP 2022 EVALUATION SET (REFERENCE VAD). THE DIARIZATION ERROR RATE (DER) IS THE SUM OF MISS (MI), FALSE ALARM (FA), AND SPEAKER CONFUSION (SC) RATES.

ID	**Inference**	**Input**			**Output**			**MI (%)**	**FA (%)**	**SC (%)**	**DER (%)**
		Audio	**Video**	**Audio-visual**	**Audio-based**	**Video-based**	**Mixed**				
S1	Stage 1	✓			✓			7.85	2.01	13.49	23.35
S2	Stage 2		✓			✓		7.36	3.31	4.34	15.01
S3	Stage 3			✓		✓		4.77	2.10	3.19	**10.06**
S4 *	Stage 1	✓			✓			8.25	2.85	14.75	25.85
	Stage 2		✓			✓		9.89	1.53	5.16	16.58
	Stage 3			✓		✓		4.60	2.48	3.31	**10.39**
	Stage 4 w/o SA ▽			✓			✓	3.84	2.69	2.25	8.78
	Stage 4 w/ SA ▽			✓			✓	3.84	2.68	2.22	**8.74**
S5	Stage 1	✓			✓			9.05	1.98	12.95	23.98
	Stage 2		✓			✓		7.45	2.74	4.22	14.41
	Stage 3			✓		✓		4.65	2.43	3.08	**10.16**
	Stage 4 w/o SA ▽			✓			✓	3.97	2.55	1.68	8.20
	Stage 4 w/ SA ▽			✓			✓	3.93	2.56	1.66	**8.15**

* indicates the model trained without the proposed multi-stage training strategy.
▽ denotes the abbreviation of speaker alignment.

20). In contrast, this work systematically explores a wide range of inference setups to gain a deeper understanding of their impact on performance.
- Third, the reported DERs in this work are averaged over three independent runs to mitigate randomness, whereas [19] reports single-run outcomes. These differences may account for the observed variations in DER.

TABLE VI
COMPARISONS OF MIMO-TSVAD MODELS WITH OTHERS ON THE MISP 2022 EVALUATION SET (REFERENCE VAD).

Method	**DER (%)**
E2E-AVSD (Official Baseline) [31]	13.88
NPU-FlySpeech [32] [a]	10.90
SJTU [95] [b]	10.82
WHU-Alibaba [36] [c]	8.82
VBx [5]	33.37
AHC [23]	30.02
LSTM-SC [18]	29.30
+ MIMO-TSVAD (S5 in Table V)	**8.15**

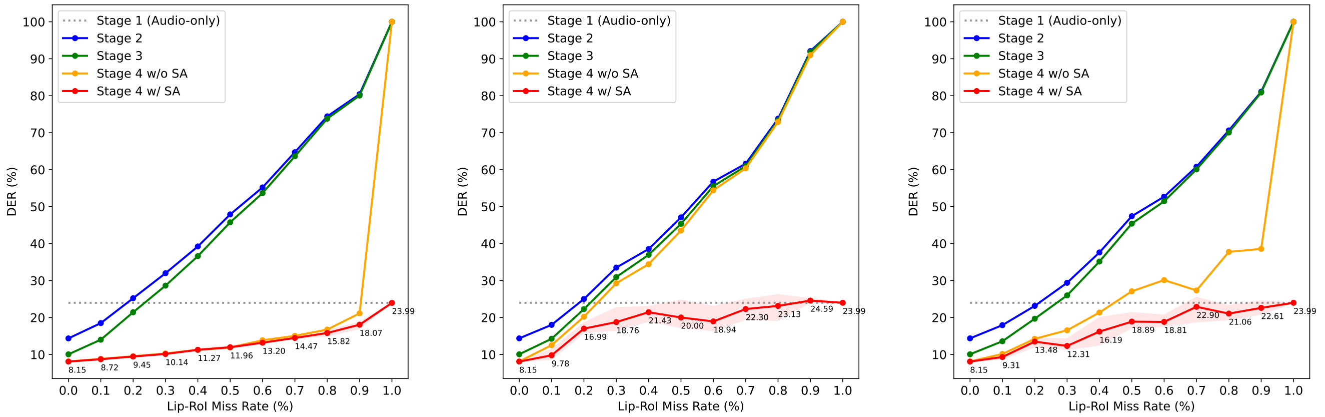
[a, b, c] denote the 3rd-, 2nd-, and 1st-ranked submissions to the audio-visual diarization track of MISP 2022 Challenge, respectively.

B. Evaluation of Video-based and Audio-Visual Diarization

When loading all modules in Fig. 2, the MIMO-TSVAD can conduct various inference stages flexibly. Limited by the large GPU memory consumption of multi-modal data, we only train the models under the VAD resolution of 10 ms, speaker capacity of 6, chunk length of 8 seconds, and chunk shift of 2 seconds. Also, we utilize the first channel of far-field audio in the MISP 2022 training set to tune the AHC [23], LSTM-SC [18], and VBx [5], respectively.

For the MISP 2022 evaluation set, reference VAD is provided according to the challenge's track 1 rules. The LSTM-SC obtains the 29.30% DER, better than the AHC (30.02%) and VBx (33.37%). Thus, Table V illustrates the performance of different MIMO-TSVAD models initialized by the LSTM-SC result. To investigate the potential of audio-based, video-based, and mixed output methods, Systems S1-3 are individually trained for each inference stage. Experimental results show that the video-based System 2 obtains 15.01% DER, surpassing the audio-based System 1 with 23.35% DER. The improvement mostly comes from the speaker confusion (SC) error decreasing from 13.49% to 4.34%. Visual modality demonstrates a significant advantage in determining speaker identities as it does not suffer from the issue of overlapping speakers. Then, System 3 decreases the DER to 10.06% by introducing auxiliary audio information. About 3/4 improvement (3.8% of 4.95% DER) is contributed by miss (MI) and false alarm (FA) rates. Compared with video signals of a limited 25 FPS, it can be seen that the audio modality can bring higher temporal precision for predicted utterance boundaries.

However, the independent training of Systems S1-3 is not cost-effective. Accordingly, we propose a multi-stage training strategy to obtain an integrated model for all inference stages. Once trained, the model supports all the functionalities of Systems S1-3 and unlocks the new mixed output method. We first train System S4 starting from the last stage described in Section III-B. Without the multi-stage design, experimental results show that although the inference *stages 1-3* of System 4 degrade considerably compared with the counterpart Systems 1-3, the inference *stage 4* still reduces the DER to 8.74% by adopting the advanced mixed output method. Then, we train System S5 with the presented multi-stage training strategy as shown in Fig. 3. It can be seen that the performance degradation of inference *stages 1-3* is alleviated obviously. Meanwhile, its inference *stage 4* reaches the lowest DER of 8.15%. As no dramatic modality-missing problem exists in the MISP 2022 dataset, the benefits of adopting the Speaker Alignment (SA) module are not noteworthy here but are thoroughly investigated in the next section. Furthermore, Table VI



(a) Partially Off-screen (b) Completely Off-screen (c) Hybrid

Fig. 6. DERs (%) of MIMO-TSVAD models on the simulated MISP 2022 evaluation sets with different Lip-RoI miss rates. Each value is the average performance on three copies of the independent simulation, where the red-shaded bands indicate the maximum-minimum intervals.

compares our proposed method with the current state-of-the-art results. The MIMO-TSVAD method updates our previous system [36] that has won the audio-visual diarization track of the MISP 2022 Challenge.

C. *Evaluatioin of Robustness to Lip-Missing Scenarios*

A speaker's lip track may not always be available in real scenarios. To explore the impact of lip-missing problems on the MIMO-TSVAD framework, we newly simulate test data based on the original MISP 2022 evaluation set. Three lip-missing scenarios are created using zeros to randomly mask the Lip-RoI for simulating off-screen data, described as follows.

- *Partially Off-screen*: Each speaker's Lip-RoIs are removed during a period. The total number of speakers in the video remains unchanged.
- *Completely Off-screen*: The Lip-RoIs of some selected speakers are entirely removed. Fewer speakers exist in the video.
- *Hybrid*: Both situations above may happen.

Fig. 6 demonstrates how the performance of the MIMO-TSVAD system (S5 in Table V) changes under different lip-missing scenarios and degrees. Since the inference *stages 2-3* entirely rely on the video-based output, their DERs increase almost linearly as the Lip-RoI miss rates rise in all scenarios. Notably, the inference *stage 4* without speaker alignment performs differently in different lip-missing scenarios. In Fig. 6a, it utilizes the mixed output method to keep considerable robustness as long as each speaker's few Lip-RoI segments can successfully extract the paired speaker embedding. Once the Lip-RoI miss rate reaches 100%, the system directly fails. In Fig. 6b, its DERs also increase rapidly with the Lip-RoI miss rate increases because all the missed Lip-RoI comes from completely off-screen speakers whose speaker embeddings cannot be enrolled by the video-based output of the previous stage. In Fig. 6c, its DER curve combines the characteristics in Fig. 6a and Fig. 6b. The inference *stage 4* without speaker alignment may recall partially off-screen speakers' voice activities but is useless for completely off-screen speakers. Lastly, using the speaker alignment module dramatically improves the robustness of inference *stage 4*. With the help of undetected speaker embeddings, the mixed output method can work stably in different cases.

TABLE VII
COMPUTING EFFICIENCY OF MIMO-TSVAD MODELS FOR EACH 8-SECOND INPUT DATA IN DIFFERENT INFERENCE STAGES, REGARDING THE NUMBER OF PARAMETERS, FLOATING-POINT OPERATIONS (FLOPS), GPU MEMORY, AND INFERENCE TIME.

Stage	Params (M)	FLOPs (G)	Memory (MB)	Time (s)
1	76.56	151.80	494.49	0.0279
2	64.11	515.60	943.60	0.0499
3	85.78	667.01	1040.50	0.0617
4	153.72	669.16	1313.22	0.0797

In general, our proposed MIMO-TSVAD framework exhibits strong robustness to complex lip-missing scenarios. With the Lip-RoI missing rate varying from 0 to 100%, it transits from an audio-visual to an audio-only system, maintaining the DER within a satisfactory range.

D. *Computing Efficiency*

Table VII illustrates the computing efficiency of MIMO-TSVAD models trained in Table V. For each 8-second input data, several metrics for the different inference stages are presented as follows. First, the number of parameters for each stage is an essential indicator. Second, Floating-Point Operations (FLOPs) are used to measure the computational complexity. Then, we count the required GPU memory and the average time for inferring each input chunk, which is tested on the NVIDIA RTX-3090 GPU. Starting from the second stage, the use of visual modality significantly increases the computational load. The more multi-modal information the model uses, the greater the computational load is required. Improving the computing efficiency of multi-modal speaker diarization systems is still challenging.

VI. Conclusions

This paper proposes a novel MIMO-TSVAD framework to tackle speaker diarization under complicated audio-visual data accessibilities. The model with jointly designed multi-stage training and inference strategies is compatible with different scenarios in a unified framework. Experimental results show that the MIMO-TSVAD framework performs well for audio-based, video-based, and audio-visual speaker diarization. It obtains new state-of-the-art DERs of 4.18% on the Vox-Converse [81] test set, 10.10% on the DIHARD-III [82] evaluation set, and 8.15% on the MISP 2022 [37] evaluation set, respectively. Furthermore, the MIMO-TSVAD framework demonstrates strong robustness against lip-missing problems. In simulated scenarios with varying lip-missing degrees, it guarantees that the DERs of the audio-visual system are always no worse than the audio-only system. In the future, we will further improve the current approach regarding advanced lip extraction and clustering methods, better use of multi-channel audio, etc.

Acknowledgments

This research is funded in part by the National Natural Science Foundation of China (62171207), Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100), 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] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, no. C, 2022.

[3] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. ICASSP*, 2018, pp. 5239–5243.

[4] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2020.

[5] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101254, 2022.

[6] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, "But system for the second DIHARD speech diarization challenge," in *Proc. ICASSP*, 2020, pp. 6529–6533.

[7] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. INTERSPEECH*, 2021, pp. 3111–3115.

[8] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. INTERSPEECH*, 2019, pp. 4300–4304.

[9] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. ASRU*, 2019, pp. 296–303.

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

[11] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[12] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *Proc. ICASSP*, 2021, pp. 7198–7202.

[13] ——, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Proc. INTERSPEECH*, 2021, pp. 3565–3569.

[14] K. Kinoshita, M. Delcroix, and T. Iwata, "Tight integration of neural- and clustering-based diarization through deep unfolding of infinite Gaussian mixture model," in *Proc. ICASSP*, 2022, pp. 8382–8386.

[15] S. Horiguchi, S. Watanabe, P. García, Y. Xue, Y. Takashima, and Y. Kawaguchi, "Towards neural diarization for unlimited numbers of speakers using global and local attractors," in *Proc. ASRU*, 2021, pp. 98–105.

[16] S. Horiguchi, S. Watanabe, P. García, Y. Takashima, and Y. Kawaguchi, "Online neural diarization of unlimited numbers of speakers using global and local attractors," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 31, pp. 706—-720, dec 2022.

[17] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. INTERSPEECH*, 2020, pp. 274–278.

[18] W. Wang, Q. Lin, D. Cai, and M. Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2645–2658, 2022.

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

[20] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," in *Proc. ICASSP*, 2023, pp. 1–5.

[21] Y. Wang, M. He, S. Niu, L. Sun, T. Gao, X. Fang, J. Pan, J. Du, and C.-H. Lee, "USTC-NELSLIP system description for DIHARD-III challenge," *arXiv:2103.10661*, 2021.

[22] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and M. Li, "The DKU-DukeECE-Lenovo system for the diarization task of the 2021 VoxCeleb speaker recognition challenge," *arXiv:2109.02002*, 2021.

[23] W. Wang, X. Qin, M. Cheng, Y. Zhang, K. Wang, and M. Li, "The DKU-DukeECE diarization system for the VoxCeleb speaker recognition challenge 2022," *arXiv:2210.01677*, 2022.

[24] M. Cheng, W. Wang, X. Qin, Y. Lin, N. Jiang, G. Zhao, and M. Li, "The DKU-MSXF diarization system for the VoxCeleb speaker recognition challenge 2023," in *Man-Machine Speech Communication*. Singapore: Springer Nature Singapore, 2024, pp. 330–337.

[25] E. El Khoury, C. Sénac, and P. Joly, "Audiovisual diarization of people in video content," *Multimedia Tools Appl.*, vol. 68, no. 3, pp. 747—-775, feb 2014.

[26] I. Kapsouras, A. Tefas, N. Nikolaidis, G. Peeters, L. Benaroya, and I. Pitas, "Multimodal speaker clustering in full length movies," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2223—-2242, 2017.

[27] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2018.

[28] J. S. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," in *Proc. INTERSPEECH*, 2019, pp. 371–375.

[29] E. Z. Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "Ava-avd: Audio-visual speaker diarization in the wild," in *Proc. MM*. Association for Computing Machinery, 2022, pp. 3838—3847.

[30] A. Wuerkaixi, K. Yan, Y. Zhang, Z. Duan, and C. Zhang, "Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization," in *Proc. MMSP*, 2022, pp. 1–6.

[31] M.-K. He, J. Du, and C.-H. Lee, "End-to-end audio-visual neural speaker diarization," in *Proc. INTERSPEECH*, 2022, pp. 1461–1465.

[32] L. Zhang, H. Zhao, Y. Li, B. Pang, Y. Wang, H. Wang, W. Rao, Q. Wang, and L. Xie, "The FlySpeech audio-visual speaker diarization system for MISP challenge 2022," *arXiv:2307.15400*, 2023.

[33] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "Ava active speaker: An audio-visual dataset for active speaker detection," in *Proc. ICASSP*, 2020, pp. 4492–4496.

[34] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proc. MM*. Association for Computing Machinery, 2021, p. 3927–3935.

[35] Y. Jiang, R. Tao, Z. Pan, and H. Li, "Target active speaker detection with audio-visual cues," in *Proc. INTERSPEECH*, 2023, pp. 3152–3156.

[36] M. Cheng, H. Wang, Z. Wang, Q. Fu, and M. Li, "The WHU-Alibaba audio-visual speaker diarization system for the MISP 2022 challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[37] Z. Wang, S. Wu, H. Chen, M.-K. He, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. Siniscalchi, O. Scharenborg, D. Liu, B. Yin, J. Pan, J. Gao, and C. Liu, "The multimodal information based speech processing (MISP) 2022 challenge: Audio-visual diarization and recognition," in *Proc. ICASSP*, 2023, pp. 1–5.

[38] S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. ICASSP*, 2018, pp. 5549–5553.

[39] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *Proc. ICASSP*, 2017, pp. 4945–4949.

[40] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. INTERSPEECH*, 2018, pp. 2808–2812.

[41] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[42] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[43] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proc. SLT*, 2014, pp. 413–417.

[44] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM based similarity measurement with spectral clustering for speaker diarization," in *Proc. INTERSPEECH*, 2019, pp. 366–370.

[45] Q. Lin, Y. Hou, and M. Li, "Self-attentive similarity measurement strategies in speaker diarization," in *Proc. INTERSPEECH*, 2020, pp. 284–288.

[46] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*, 2018, pp. 1561–1565.

[47] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME*, 2020, pp. 1–7.

[48] X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, S. Chen, Y. Zhao, G. Liu, Y. Wu, J. Wu, S. Liu, J. Li, and Y. Gong, "Microsoft speaker diarization system for the VoxCeleb speaker recognition challenge 2020," in *Proc. ICASSP*, 2021, pp. 5824–5828.

[49] G. Morrone, S. Cornell, D. Raj, L. Serafini, E. Zovato, A. Brutti, and S. Squartini, "Low-latency speech separation guided diarization for telephone conversations," in *Proc. SLT*, 2023, pp. 641–646.

[50] E. Han, C. Lee, and A. Stolcke, "Bw-eda-eend: streaming end-to-end neural speaker diarization for a variable number of speakers," in *Proc. ICASSP*, 2021, pp. 7193–7197.

[51] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, P. García, and K. Nagamatsu, "Online end-to-end neural diarization with speaker-tracing buffer," in *Proc. SLT*, 2021, pp. 841–848.

[52] Y. C. Liu, E. Han, C. Lee, and A. Stolcke, "End-to-end neural diarization: From transformer to conformer," in *Proc. INTERSPEECH*, 2021, pp. 3081–3085.

[53] M. Rybicka, J. Villalba, N. Dehak, and K. Kowalczyk, "End-to-end neural speaker diarization with an iterative refinement of non-autoregressive attention-based attractors," in *Proc. INTERSPEECH*, 2022, pp. 5090–5094.

[54] Y. Fujita, T. Komatsu, R. Scheibler, Y. Kida, and T. Ogawa, "Neural diarization with non-autoregressive intermediate attractors," in *Proc. ICASSP*, 2023, pp. 1–5.

[55] Z. Chen, B. Han, S. Wang, and Y. Qian, "Attention-based encoder-decoder network for end-to-end neural speaker diarization with target speaker attractor," in *Proc. INTERSPEECH*, 2023, pp. 3552–3556.

[56] T.-Y. Leung and L. Samarakoon, "Robust end-to-end speaker diarization with conformer and additive margin penalty," in *Proc. INTERSPEECH*, 2021, pp. 3575–3579.

[57] Y.-R. Jeoung, J.-Y. Yang, J.-H. Choi, and J.-H. Chang, "Improving transformer-based end-to-end speaker diarization by assigning auxiliary losses to attention heads," in *Proc. ICASSP*, 2023, pp. 1–5.

[58] Y. Ding, Y. Xu, S.-X. Zhang, Y. Cong, and L. Wang, "Self-supervised learning for audio-visual speaker diarization," in *Proc. ICASSP*, 2020, pp. 4367–4371.

[59] Y. Dissen, F. Kreuk, and J. Keshet, "Self-supervised speaker diarization," in *Proc. INTERSPEECH*, 2022, pp. 4013–4017.

[60] Y. Takashima, Y. Fujita, S. Horiguchi, S. Watanabe, L. P. G. Perera, and K. Nagamatsu, "Semi-supervised training with pseudo-labeling for end-to-end neural diarization," in *Proc. INTERSPEECH*, 2021, pp. 3096–3100.

[61] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Proc. Odyssey*, 2020, pp. 433–439.

[62] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker," in *Proc. INTERSPEECH*, 2021, pp. 3555–3559.

[63] C.-Y. Cheng, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Multi-target extractor and detector for unknown-number speaker diarization," *IEEE Signal Processing Letters*, vol. 30, pp. 638–642, 2023.

[64] Y.-X. Wang, J. Du, M. He, S.-T. Niu, L. Sun, and C.-H. Lee, "Scenario-dependent speaker diarization for DIHARD-III challenge," in *Proc. INTERSPEECH*, 2021, pp. 3106–3110.

[65] W. Wang, X. Qin, and M. Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the M2MeT challenge," in *Proc. ICASSP*, 2022, pp. 9171–9175.

[66] W. Wang, Q. Lin, and M. Li, "Online target speaker voice activity detection for speaker diarization," in *Proc. INTERSPEECH*, 2022, pp. 1441–1445.

[67] W. Wang and M. Li, "End-to-end online speaker diarization with target speaker tracking," *arXiv:2310.08696*, 2023.

[68] ——, "Incorporating end-to-end framework into target-speaker voice activity detection," in *Proc. ICASSP*, 2022, pp. 8362–8366.

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

[70] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[72] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018, pp. 6450–6459.

[73] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[74] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," in *Proc. NeurIPS*, vol. 35. Curran Associates, Inc., 2022, pp. 32 897–32 912.

[75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30. Curran Associates, Inc., 2017.

[76] M. Cheng, H. Wang, Y. Wang, and M. Li, "The dku audio-visual wake word spotting system for the 2021 misp challenge," in *Proc. ICASSP*, 2022, pp. 9256–9260.

[77] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc. CVPR*, 2020, pp. 5202–5211.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[79] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. FG*, 2019, pp. 1–8.

[80] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[81] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: Speaker diarisation in the wild," in *Proc. INTERSPEECH*, 2020, pp. 299–303.

[82] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third DIHARD diarization challenge," in *Proc. INTERSPEECH*, 2021, pp. 3570–3574.

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

[84] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.

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

[86] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc. ICASSP*, 2022, pp. 6722–6726.

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

[88] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[89] K. Wang, X. Mao, H. Wu, C. Ding, C. Shang, R. Xia, and Y. Wang, "The ByteDance speaker diarization system for the VoxCeleb speaker recognition challenge 2021," *arXiv:2109.02047*, 2021.

[90] D. Wang, X. Xiao, N. Kanda, M. Yousefi, T. Yoshioka, and J. Wu, "Profile-error-tolerant target-speaker voice activity detection," in *Proc. ICASSP*, 2024, pp. 11 906–11 910.

[91] S. Baroudi, H. Bredin, A. Plaquet, and T. Pellegrini, "pyannote.audio speaker diarization pipeline at voxsrc 2023," *The VoxCeleb Speaker Recognition Challenge*, vol. 2023, 2023.

[92] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[93] S. Horiguchi, N. Yalta, P. Garcia, Y. Takashima, Y. Xue, D. Raj, Z. Huang, Y. Fujita, S. Watanabe, and S. Khudanpur, "The Hitachi-JHU DIHARD III system: Competitive end-to-end neural diarization and x-vector clustering systems combined by dover-lap," *arXiv:2102.01363*, 2021.

[94] M.-K. He, J. Du, Q.-F. Liu, and C.-H. Lee, "Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1561–1573, 2023.

[95] T. Liu, Z. Chen, Y. Qian, and K. Yu, "Multi-speaker end-to-end multi-modal speaker diarization system for the MISP 2022 challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

Ming Cheng is currently a Ph.D. candidate in Computer Science at Wuhan University. He received his Master's degree in Electrical and Electronic Engineering from The University of Hong Kong and Bachelor's degree in Measuring and Control Technologies and Instruments from China Jiliang University. His research interests include speech signal processing and multimodal behavior analysis.

Ming Li (Senior Member, IEEE) received his Ph.D. in Electrical Engineering from University of Southern California in 2013. He is currently a Professor of Electronical and Computer Engineering at Duke Kunshan University. He is also an Adjunct Professor at School of Computer Science in Wuhan University. His research interests are in the areas of audio, speech and language processing as well as multimodal behavior signal processing. He has published more than 200 papers and served as the member of IEEE speech and language technical committee, APSIPA speech and language processing technical committee, the editorial board member of the IEEE/ACM Transactions on Audio, Speech, and Language Processing and Computer Speech & Language. He is an area chair at Interspeech 2016, 2018, 2020, 2024 and 2025 as well as the technical program co-chair of Odyssey 2022 and ASRU 2023. Works co-authored with his colleagues have won first prize awards at Interspeech Computational Paralinguistic Challenges 2011, 2012 and 2019, ASRU 2019 MGB-5 ADI Challenge, Interspeech 2020 and 2021 Fearless Steps Challenges, VoxSRC 2021, 2022 and 2023 Challenges, ICASSP 2022 M2MeT Challenge, ICASSP 2023 MISP challenge, IJCAI 2023 ADD challenge, ICME 2024 ChatCLR challenge and Interspeech 2024 AVSE challenge. He received the IBM faculty award in 2016, the ISCA Computer Speech and Language 5-years best journal paper award in 2018 and the youth achievement award of outstanding scientific research achievements of Chinese higher education in 2020.