DyViSE: Dynamic Vision-Guided Speaker Embedding for Audio-Visual Speaker Diarization

Abudukelimu Wuerkaixi[†], Kunda Yan[†], You Zhang[‡], Zhiyao Duan[‡], Changshui Zhang[†]
[†]Institute for Artificial Intelligence, Tsinghua University (THUAI),
State Key Lab of Intelligent Technologies and Systems,
Beijing National Research Center for Information Science and Technology (BNRist),
Department of Automation, Tsinghua University, Beijing, P.R.China
Emails: {wekxabdk21, ykd22}@mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn
[‡]Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA
Emails: {you.zhang, zhiyao.duan}@rochester.edu

***Abstract*—Speaker diarization aims to determine "who spoke when" in multi-speaker scenarios. Audio-visual speaker diarization leverages visual information in addition to audio signals and has shown improved performance. Existing audio-visual methods extract speaker embeddings for each video clip using audio and facial features, and then perform clustering according to their similarity. However, this approach would not work well for noisy or overlapped speech where audio features are corrupted, nor for off-screen speakers where visual features are missing. In this work, we propose dynamic vision-guided speaker embedding (DyViSE), a novel method for leveraging visual information to extract speaker embeddings in a multi-stage system. DyViSE uses dynamic lip movement information to *denoise* audio in a latent space and integrates facial features to obtain an identity-discriminative embedding for each speaking segment. DyViSE is trained with a deep clustering loss along with an exemplary loss. DyViSE demonstrates remarkable performance on both real-world videos and artificially assembled videos. Our code is available at https://github.com/urkax/DyViSE.**

***Index Terms*—speaker diarization, audio-visual, speech overlap, deep clustering**

I. Introduction

Speaker diarization is a challenging and fundamental task in speech processing which aims to solve the problem of "who spoke when" [1], [2]. It can be approached by segmenting a multi-speaker video into speaking segments and then clustering such segments according to speaker identity. There have been numerous works that address speaker diarization using only the audio modality [3], [4]. However, the performance is still far from satisfactory in the real world, due to background noise and simultaneous speaking from multiple people [2].

Researchers have explored the possibility of utilizing the visual modality for speaker diarization [5]–[9]. It has been proven that using facial features improves the diarization performance [10]. A prominent method for audio-visual speaker diarization is a multi-stage system [10]. The system firstly detects speaking segments in the audio with Voice Activity Detection (VAD), then applies Active Speaker Detection (ASD) to locate a speaking face for each speaking segment. The facial features and acoustic features are then jointly used to cluster the speaking segments into different speakers.

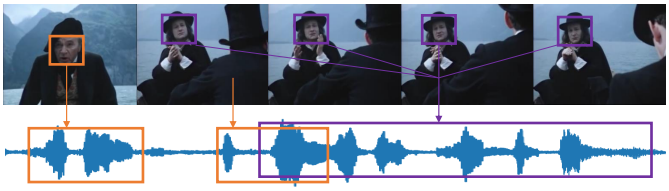

Fig. 1. Key frames and audio waveform of a video segment with two persons speaking back and forth with their voices overlapped. In the latter part of the segment, the face of one speaker is on screen, while the other is off-screen.

However, current audio-visual speaker diarization studies have yet to plumb the depths of the audio overlapping issue. Besides, in real-world applications, the face of a speaker is often out of sight or not clearly seen. Figure 1 is an example. Current audio-visual speaker diarization methods (e.g. [10]) cluster speaking segments according to the facial features of the located face images, hence the diarization performance may be limited in missing face scenarios. In the AVA-AVD speaker diarization dataset [10], 60.7% of the speaking segments are from off-screen speakers by our experiments.

Inspired by lip-guided audio separation [11], we propose a Dynamic Vision-Guided Speaker Embedding (DyViSE) to handle the noisy audio and overlapping voice problems. Taking an audio segment and its corresponding face frames as input, our neural network leverages the lip movement information to extract a "denoised" audio-visual embedding, reducing the effect of noise and overlapped speech. The DyViSE embedding network is trained with the deep clustering loss [12] to encourage more discriminative embeddings for speaking segments from different speakers. We also propose an exemplary loss where a pre-trained speaker recognition network serves as a "teacher" network to guide the training of the DyViSE model in order to make it converge faster.

We then propose a multi-stage system to tackle speaker

This work is partially funded by the Natural Science Fundation of China (NSFC 62176132), the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 62061136001/DFG TRR-169, and the National Science Foundation (NSF) grant 1741472. You Zhang also would like to thank the synergistic activities provided by the NRT program on AR/VR funded by NSF grant DGE-1922591.

978-1-6654-7189-3/22/$31.00 © 2022 IEEE

diarization with DyViSE. Speaking segments and their corresponding face tracks are first detected with overlap-aware VAD and ASD methods, then their DyViSE embeddings are extracted. Finally, such embeddings are clustered using agglomerative clustering to achieve speaker diarization. As it is the vision-guided speaker embedding instead of the visual feature that is used in the clustering process, the proposed system is more robust to missing faces. Experiments on real-world speaker diarization datasets demonstrate that our proposed method outperforms a state-of-the-art method [10] and our own designed baselines. To further examine the effectiveness of DyViSE in overlapped speech cases, we created synthetic audio-visual speaker diarization datasets by artificially assembling individual speaking videos with a higher overlapping rate. Our proposed system achieves a decrease in Diarization Error Rate (DER) by a large margin compared to the baseline method AVR-Net [10] on our synthesized datasets.

The contributions of this work are threefold. First, we propose a novel dynamic vision-guided speaker embedding extraction method for audio-visual speaker diarization, which uses dynamic vision information to address the overlapped speech problem. Second, we propose a multi-stage system that uses this embedding in clustering speaking segments and it relieves the missing face problem from which existing audio-visual speaker diarization methods suffer. Third, we are the first to systematically investigate the abovementioned problems in audio-visual diarization through experiments.

II. Related Work

A. Speaker Diarization

Speaker diarization can be categorized into audio-only and audio-visual in terms of the input modality. Audio-only speaker diarization has received extensive attention for many years. The latest advancements in audio-only methods mostly adopt a multi-stage framework [3], [4]. Several end-to-end diarization systems have also been proposed to address the speech overlapping issue [13], [14]. However, end-to-end frameworks require a high computation complexity and the length of segments that can be processed is limited.

Audio-visual speaker diarization utilizes additional information provided by the visual modality. Speaker diarization can benefit from visual signals, especially when audio is not reliable. It is proved that diarization systems that associate facial and audio features outperform unimodal systems [8], [15], [16]. Moreover, audio cues for each speaker are also present in lip movements. Researchers leverage the synchronization between lip movements and speech signals for active speaker detection, a prior step for diarization [7], [17], [18]. However, current speaker diarization methods disregard the ability of lip movements to transcribe speech, which is useful for handling overlapped speech as discussed in Section I.

B. Vision-Guided Speech Embedding

The concept of extracting audio cues from lip movements has been applied in many research areas, among which lip reading [19] and audio-visual speech recognition (AVSR) [20] are two representative ones. They attempt to recognize or separate speech in silent or audible videos using lip movement features, which demonstrate the effectiveness of leveraging lip movements for audio cues. Recent audio-visual speech separation methods have tried a variety of deep learning networks [11], [21], [22] to extract the target speech from a mixture of voices. Some of them are audio-visual synchronization networks fusing lip movements and audio features for the separation process [21], [22], while VisualVoice [11] introduces face appearance as an additional signal and concatenates the embeddings extracted from lip motion and face appearance to explicitly guide speech separation. Benefiting from a variety of complementary information, VisualVoice [11] yields remarkable performance and generalization ability. Inspired by these works, we incorporate lip movements to alleviate the issue of speech overlapping in audio-visual speaker diarization. Instead of adopting audio-visual speech separation as a preprocessing step, we borrow its idea but train an embedding network that extracts dynamic vision-guided speaker embedding from speaking segments. This embedding can be viewed as an implicitly "denoised" audio embedding plus visual features, and is then used for speaker diarization. The rationale for this design instead of the "separation-diarization" approach is that reconstructing the separated speech is itself a challenging problem and often introduces artifacts to the separated audio. Nevertheless, we take the "separation-diarization" approach as a baseline for comparison in our experiments in Section V-A.

III. Method

A. Overview

In this work, we utilize a multi-stage framework following commonly used audio-visual speaker diarization systems [7], [10] and propose a novel speaker embedding named DyViSE to handle noisy audio and overlapped speech problems.

Prior to the proposed embedding network, our multi-stage system adopts three pre-processing steps to obtain audio-visual speaking segments for later stages to extract embeddings and diarize speakers. The first step is to use a pre-trained U-Net-based network [23] for overlap-aware VAD on the audio modality, and to use a face detection method [24] to obtain face tracks on the speaking segments. We assume that overlaps only happen between two but not more speakers. As the second step, TalkNet [25] is employed for ASD to iterate through all the corresponding face tracks and discriminate whether one is speaking for the active voice. Thus, we divide the original video into a set of audio segments $s_n(t)$ and their associated face tracks $f_n(t)$, where $n \in \{1, 2, \cdots, N\}$ denotes the speech segment index and t denotes the frame index within the segment. As the last step, we adopt a face alignment network [26] and crop the mouth regions using detected facial landmarks as in [11] to obtain the lip movement tracks $l_n(t)$. We discuss prior steps in more detail in the supplementary material.

The audio-visual segments are cut or padded into 2.55s in duration. $p_n(t) = (s_n(t), l_n(t), f_n(t_i))$ indicates an audio-visual pack combining the audio segment, the lip movement

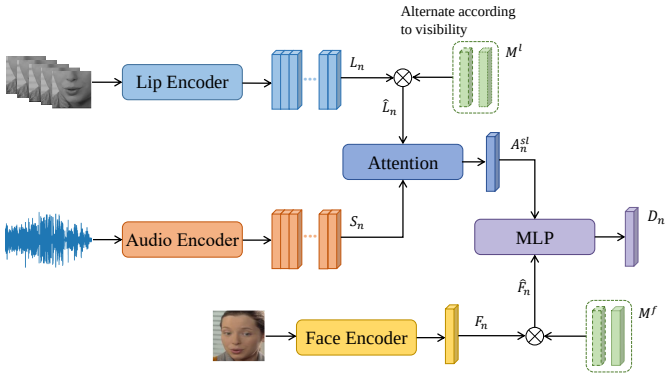

Fig. 2. Architecture of the proposed DyViSE network.

track and a face image randomly selected from the face track, where t_i denotes one of the face frames. When the face of a segment is off-screen, $f_n(t_i)$ and $l_n(t)$ are padded with zero. Our network takes each $p_n(t)$ as input to produce an identity-discriminative embedding D_n. After embeddings are extracted for all segments, the cosine similarity is calculated for all pairs of segments. Finally, agglomerative clustering is used to obtain speaker diarization results.

B. Network Architecture for DyViSE

The architecture of our network is shown in Figure 2. Given an audio-visual pack, our network firstly encodes lip movement features $\hat{L}_n$ and audio features S_n respectively and then combines them to extract cleaner audio embedding A_n^{sl}. Facial embedding $\hat{F}_n$ is also extracted as complementary information. We fuse the audio and facial embedding using an MLP network and lastly acquire identity-discriminative embedding D_n, named DyViSE.

Speaker diarization systems typically leverage a pre-trained Speaker Recognition (SR) model to encode audio features. As typical SR models are trained only on clean audio datasets without overlapped speech, features extracted with these models may encounter problems in overlapped speech scenarios. Recently, the pre-trained audio models based on Transformers exhibit excellent performance on downstream tasks in the field of speech processing. Among them, WavLM [27] performs pre-training on simulated noisy and overlapped speech data. Thus, we adopt and adapt the pre-trained SR model [27] that uses WavLM as the frontend and the state-of-the-art backend model ECAPA-TDNN. Specifically, we remove the last temporal pooling layer of the SR model and add a trainable 1D convolutional layer on its top. We use it to extract a d-dimensional audio feature vector $S_n(t)$ for each frame t of a 2.55s audio segment, resulting in an audio feature matrix $S_n \in \mathbb{R}^{T\times d}$, where T denotes the number of frames.

The lip motion encoder aims to encode dynamic lip motion features. We adopt the trained lip motion analysis network [11], which consists of a 3D convolutional layer, ShuffleNet V2 network, and a temporal convolutional network. We also append a trainable linear layer to form the lip motion encoder. The encoder takes lip movement track frames $l_n(t)$ as input and outputs lip movement features $L_n \in \mathbb{R}^{T\times d}$, where a vector with the same dimension as $S_n(t)$ is extracted for each frame. Considering the missing face problem, we employ learnable masks $M^l = [M_1^l, M_2^l]$ inspired by [10] to distinguish between face-visible and face-invisible scenarios, where $M_v^l \in \mathbb{R}^d, v \in \{1, 2\}$. For each of the two cases, we multiply the output of the lip motion encoder by a mask vector: $\hat{L}_n(k) = L_n(k) \odot M_v^l$, where $k \in \{1, 2, \cdots, T\}$ is the frame index, and v indicates the visibility condition.

We integrate speech features S_n and lip movement features $\hat{L}_n$ with three attention modules, following the multi-modal fusion method [25], [28]. The attention is calculated among features of frames. S_n and $\hat{L}_n$ firstly interact with each other through two attention modules. One of them takes S_n to generate key and value while takes $\hat{L}_n$ to generate query, the other performs vice versa. The outputs of the two modules are concatenated along the temporal direction. Then a self-attention module is applied and the output features of all frames are averaged to get a "denoised" audio embedding $A_n^{sl} \in \mathbb{R}^{2d}$.

The face encoder is a pre-trained face recognition model [29]. We also use learnable mask vectors M^f to distinguish face-visible and face-invisible scenarios. The facial embedding $\hat{F}_n$ and "denoised" audio embedding A_n^{sl} are concatenated and passed through a multilayer perceptron (MLP), and lastly merged into the final DyViSE embedding D_n.

C. Deep Clustering Loss and Exemplary Loss

Deep clustering loss is employed to train identity-discriminative embeddings. For a sequence of audio-visual segment embeddings D_n, $n \in \{1, 2, \cdots, N\}$, from the same video, we obtain a similarity matrix, donated as $[a_{ij}]_{N\times N}$, $a_{ij} = cos\langle D_i, D_j\rangle = \frac{D_i \cdot D_j}{\|D_i\|\|D_j\|}$. Then using identity label t_n, $n \in \{1, 2, \cdots, N\}$, we generate ground-truth matrix $[r_{ij}]_{N\times N}$, where $r_{ij} = \mathbb{1}(t_i = t_j)$ and $\mathbb{1}$ is the indicator function. Deep clustering loss is formulated as

$$\mathcal{L}_d = \frac{1}{N \times N} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{(a_{ij} - r_{ij})^2}{\sqrt{d_i d_j}}, \tag{1}$$

where $d_i = |\{k : t_k = t_i\}|$ denotes the number of segments with the same label as t_i [12],

For clustering problems, the embeddings of one speaker might spread into a large area of the feature space, which may lead to the instability of convergence, especially for small-scale datasets. Thus we additionally propose an exemplary loss to maximize the cosine similarity between our output embedding with a large-scale pre-trained speaker embedding G_n. The pre-trained SR network takes the non-overlap speech of the corresponding speaker as input. The exemplary loss is computed as

$$\mathcal{L}_e = -\frac{1}{N} \sum_{n=1}^{N} cos\langle D_n, G_n\rangle . \tag{2}$$

Finally, we minimize the weighted sum of the two losses as the final loss: $\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_e$.

IV. Experimental Setup

A. AVA-AVD Dataset

AVA-AVD [10] is annotated on 117 movies with diverse outdoor scenarios. A 15-min long segment starting from the 15 minutes into the movie is selected and cut into three 5-minute clips. This results in 29 hours of clips in total. Each clip has at least two speakers and 7.7 speakers on average. Speaker identities of speaking activities are annotated. Bounding boxes of all on-screen faces are also annotated, regardless of speaking activity. Among all speaking segments, 60.7% of them do not show any face of speakers.

B. Synthesized Datasets with Higher Overlap Ratio

Current large-scale audio-visual speaker diarization datasets contain overlapped speech cases [10], but the amount may not be sufficient for reliable analyses. To demonstrate the effectiveness of our proposed DyViSE on addressing the overlapped speech problems, we synthesize speaker diarization datasets with higher overlap ratios using audio-visual face tracks from the AVA-AVD dataset and the Voxceleb2 dataset [30]. Voxceleb2 contains over 1 million utterances from 6,112 speakers, and it is originally designed for audio-visual speaker recognition. To be more specific, we define *overlap ratio* as the ratio between the length of a clip when multiple speakers are speaking and the total length of speaking segments [31], then average over all clips.

The dataset synthesis algorithm follows the diarization-style audio-only mixture simulation in [31]. The extension to our audio-visual scenario is straightforward. The algorithm details are in the supplementary material. The algorithm takes a parameter β to control the overlap ratio. With a smaller β, the overlap ratio of the synthesized dataset is higher. We set the upper limit of the number of simultaneous speakers to two in this synthesis process. For dataset clips synthesis, face tracks go with the corresponding speaking utterances and we do not generate the whole scene.

We synthesize clips based on AVA-AVD with speakers of the original AVA-AVD clips and their speaking segments as utterances. Therefore the simulated clips contain the same number of speakers and speaking segments as the AVA-AVD clips but with different overlaps. As for Voxceleb2, we randomly choose two speakers for each synthesized clip. For each speaker, we sample a minimum of 20 and a maximum of all the utterances. Besides, 50% of face tracks in Voxceleb2 synthesis are discarded to simulate out of sight voices. By adjusting β, we can synthesize clips with different overlap ratios. We name them *AVA-AVD synthesis ($\beta = x$)* or *Voxceleb2 synthesis ($\beta = x$)* where x is a constant. The statistics of AVA-AVD and three synthesized datasets are listed in Table I. We present a demo figure on AVA-AVD synthesis ($\beta = 3$) in the supplementary material.

The training, validation and test set split of AVA-AVD follows the protocol in the original publication [10], while the *Voxceleb2 synthesis* dataset is split by 3:1:1.

TABLE I
STATISTICS OF AVA-AVD AND THE THREE SYNTHESIZED DATASETS, INCLUDING THE AMOUNT OF CLIPS, TOTAL DURATION, THE MINIMUM/AVERAGE/MAXIMUM NUMBER OF SPEAKERS AND THEIR OVERLAP RATIO.

	#Clips	Duration	#Speakers	Overlap ratio
AVA-AVD	351	29 hours	2/7.7/24	4.28%
AVA-AVD synthesis (β=10)	351	52 hours	2/7.7/24	10.79%
AVA-AVD synthesis (β=3)	351	26 hours	2/7.7/24	20.60%
Voxceleb2 synthesis (β=3)	2000	1833 hours	2/2/2	20.10%

C. Implementation Details

All videos are converted to 25 Frames-Per-Second (FPS). The number of frames T is 64 after conversion. The speaker recognition (SR) model is pre-trained on Voxceleb1 and Voxceleb2 datasets as in [27] and the face recognition (FR) model from [29] is pre-trained on the MS1MV3 dataset. The dimensionality d of frame-wise lip movement feature L_n^k and audio feature S_n^k is set to 256. The MLP used to fuse the audio embedding A_n^{sl} and facial feature $\hat{F}_n$ is two linear layers connected by ReLU activation, and their hidden layer dimension is 512. The loss weight λ is simply set to 1.

D. Evaluation Metric

We use diarization error rate (DER) to evaluate the performance, defined as the fraction of speaking segments that is not correctly attributed to speakers or to non-speech [32]. Overlapped speech is also taken into account. DER is composed of three parts: Speaker error rate (SPKE) is the percentage of speaking time where the wrong speaker identity is assigned; False alarm rate (FA) is the percentage of non-speaking time that is classified as speaking; Missing speech rate (MS) is the percentage of speaking time that is not detected as speaking. For all of these metrics, the lower the better. We adopt a 0.25s collar as a tolerance around speaker boundaries.

E. Comparison Methods

Besides our proposed DyViSE, we design two baseline methods that utilize the same pre-trained models, but without dynamic vision guidance. The baseline method *SR+FR* refers to the late fusion of the two modalities. The audio embeddings and facial embeddings of two speaking segments are extracted by these models, then the cosine similarities between two speaking segments are calculated on the audio and facial embeddings respectively before they are averaged to obtain the overall similarity. *AVSS+SR+FR* refers to a serial system that first explicitly separates overlapped audio with the audio-visual speech separation (AVSS) model VisualVoice [11], and then applies the *SR+FR* method on the separated audio. Some state-of-the-art audio-visual speaker diarization [7], [10] and person verification [33] methods are also included in the comparisons.

V. Results and Discussions

A. Performance on Datasets with Overlap Speech

We first compare the diarization performance on the AVA-AVD dataset. The results are shown in Table II. We include two settings for our proposed multi-stage system as introduced

TABLE II
PERFORMANCE COMPARISON ON THE AVA-AVD DATASET. FOR VAD PRE-PROCESSING, GROUND-TRUTH ANNOTATIONS (GT) AND AN OVERLAP-AWARE VAD APPROACH (V) ARE USED.

VAD	Method	MS	FA	SPKE	DER
GT	SR+FR	2.14	0.0	21.68	23.82
	AVSS+SR+FR	2.33	0.0	22.18	24.51
	WST [7]	3.11	0.0	37.72	40.83
	AVR-Net [10]	2.45	0.0	25.38	27.83
	DyViSE	1.98	0.0	20.86	**23.46**
V	WST [7]	9.45	31.80	42.18	83.43
	AVR-Net [10]	13.37	37.29	29.36	80.02
	DyViSE	11.08	24.19	35.93	**71.20**

TABLE III
DER COMPARISON ON AVA-AVD DATASET AND THREE SYNTHESIZED DATASETS. "-" INDICATES THAT THE TRAINING FAILS TO CONVERGE.

	AVA-AVD	AVA-AVD synthesis (β=10)	AVA-AVD synthesis (β=3)	Voxceleb2 synthesis (β=3)
SR+FR	23.82	27.71	31.62	10.97
AVSS+SR+FR	24.51	26.80	30.02	9.88
GMU [33]	29.22	-	-	-
AVR-Net [10]	27.83	29.47	32.21	9.61
DyViSE	**23.46**	**25.50**	**28.61**	**6.72**

in Section III-A. One is using oracle VAD (GT) to assume the speaking segments extraction is perfect, and the other is detecting speaking segments with a trained overlap-aware VAD network (V). DyViSE exceeds previous methods by far in both settings. This suggests the effectiveness of our proposed DyViSE and the superiority is robust to the performance of speaking segments detection. It is undeniable that DyViSE benefits from the pre-trained models through exemplary loss, but DyViSE also outperforms SR+FR, a simple fusion of pre-trained models. This shows the effectiveness of our proposed dynamic vision guidance with deep clustering loss. Comparing results in the GT and V setting, the DERs are much higher when using VAD model to detect speaking segments. It indicates the challenge of VAD on the AVA-AVD dataset. Nevertheless, DyViSE also outperforms other methods in the V setting, proving the validity of it in the scenario that the VAD might not be perfect.

To focus on the evaluation of the identity discrimination ability of our method, the following experiments are conducted with oracle speaking segments and active speaker detection annotation from the datasets. The DER results on AVA-AVD and the three synthesized datasets are presented in Table III. Each method are respectively trained and tested on each dataset. With smaller β, the overlap ratio of a dataset is higher, and the performances of the models degrade. The superiority of DyViSE is more obvious on datasets with a higher overlap ratio and a larger scale. For example, on the Voxceleb2 synthesis dataset whose overlap ratio is 20.10%, DyViSE outperforms previous SOTA methods by 30.0%. This demonstrates the superiority of our proposed DyViSE on audio-visual speaking clips with overlap.

To further understand the contribution of solving the overlapped speech problems, we perform speaker diarization on overlap-only speaking segments in each video. As shown in Table IV, the absolute DER value get reduced by 2.99% and 4.75% compared to the best baseline system on the two overlap-only datasets respectively. This suggests that DyViSE has significantly better ability to handle overlapped speech and hence improve the overall performance.

TABLE IV
DER COMPARISON ON OVERLAP-ONLY SEGMENTS.

	AVA-AVD synthesis (β=3) overlap-only	Voxceleb2 synthesis (β=3) overlap-only
SR+FR	34.46	14.08
AVSS+SR+FR	33.20	12.72
AVR-Net [10]	36.82	13.58
DyViSE	**30.21**	**7.97**

TABLE V
AVERAGE PRECISION (AP) OF METHODS ON THE CLASSIFICATION OF PAIRS OF SPEAKING SEGMENTS.

	Both-faces-on-screen	One-face-on-screen	No-face-on-screen
SR+FR	85.48%	75.33%	69.31%
AVR-Net [10]	85.71%	74.95%	68.07%
DyViSE	**87.35%**	**81.47%**	**71.28%**

B. Overcoming Challenges of Missing Faces

Previous audio-visual methods use pairwise similarity between facial features of two speaking segments to assist diarization of speaker identities [10]. Visual information is not fully utilized and is only helpful when both faces are on-screen to distinguish two segments' identities. Different from previous works, DyViSE implicitly "denoised" the audio embedding in the latent space with dynamic visual features so that visual information still takes effect when one face is on-screen and the other is off-screen.

To prove this, we design a classification task where models predict whether a pair of audio-visual speaking segments belong to the same person. We randomly select 5000 pairs of speaking segments from the test set clips of AVA-AVD synthesis ($\beta = 3$). Half of the pairs are two segments of the same speaker, while the other half are two segments of different speakers but from the same clip. We experiment with three settings: both face frames are on-screen, only one of the face frames are on-screen, and both face frames are off-screen. We take the cosine similarity of two segments' embeddings and linearly map it to [0,1] as the same identity prediction score. The average precision (AP) is taken as the evaluation metric, which is the higher the better. As shown in Table V, DyViSE outperforms the baseline methods in all settings. And the superiority is more significant under the one-face-on-screen setting, conforming to our design that the visual information of the on-screen speaker can guide the discrimination. In the no face scenario, DyViSE will purely rely on the audio embedding. But it still surpasses other methods, which may be due to the audio separation ability it learns during training.

C. Ablation Study

DyViSE utilizes lip movement features and facial features. We conduct an ablation study for these two modules. We also

TABLE VI
ABLATION STUDY ON THE AVA-AVD SYNTHESIS (β=3) DATASET.

	DER
DyViSE	**28.61**
w/o lip movement features	32.73
w/o facial features	31.54
w/o exemplary loss	29.82

study the role of exemplary loss for training. The results shown in Table VI demonstrate the importance of these modules. The overlap ratio of AVA-AVD synthesis (β=3) dataset is high. DyViSE without lip movement features performs poorly on this dataset, indicating the importance of dynamic visual features for handling overlapped data. When DyViSE is trained only with deep clustering loss without exemplary loss, it achieves satisfactory performance, which demonstrates the effectiveness of deep clustering loss, but the exemplary loss also plays the role of guidance and boosts the performance.

VI. CONCLUSION

Overlapped and noisy audio has always been a problem of speaker diarization both for audio-only and audio-visual systems. In this work, we propose DyViSE which utilizes dynamic visual information to improve the quality of audio features in a latent space. Besides, previous audio-visual speaker diarization models suffer from problems of faces being out of sight. DyViSE alleviates the issue as it uses a speaking segment's visual information to "denoise" the audio embedding instead of using the raw visual features in the clustering process. To extensively evaluate the performance of DyViSE, we not only conducted experiments on the AVA-AVD dataset but also synthesized datasets with higher overlap ratios. Experimental results showed that our proposed method outperforms previous methods and baselines under multiple settings. For future work, our method can be explored in people diarization [34], which not only diarizes the identity for speaking activities, but also for visual appearances.

REFERENCES

[1] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. IEEE Spoken Language Technology Workshop*, 2021.

[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, p. 101317, 2022.

[3] 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.

[4] Y. Kwon, J. Jung, H. Heo, Y. J. Kim, B. Lee, and J. S. Chung, "Adapting speaker embeddings for speaker diarisation," in *Proc. Interspeech*, 2021.

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

[6] I. D. Gebru, S. O. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, 2018.

[7] J. S. Chung, B. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," in *Proc. Interspeech*, 2019.

[8] 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.

[9] X. Bost, G. Linarès, and S. Gueye, "Audiovisual speaker diarization of TV series," in *Proc. ICASSP*, 2015, pp. 4799–4803.

[10] E. Z. Xu, Z. Song, C. Feng, M. Ye, and M. Z. Shou, "AVA-AVD: Audio-visual speaker diarization in the wild," *CoRR*, vol. abs/2111.14448, 2021.

[11] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *Proc. CVPR*, 2021, pp. 15 495–15 505.

[12] R. Lu, Z. Duan, and C. Zhang, "Audio-visual deep clustering for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 11, pp. 1697–1712, 2019.

[13] A. Zhang, Q. Wang, Z. Zhu, J. W. Paisley, and C. Wang, "Fully supervised speaker diarization," in *Proc. ICASSP*, 2019, pp. 6301–6305.

[14] 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.

[15] R. Sharma and S. Narayanan, "Using active speaker faces for diarization in TV shows," *CoRR*, vol. abs/2203.15961, 2022.

[16] F. Vallet, S. Essid, and J. Carrive, "A multimodal approach to speaker diarization on TV talk-shows," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 509–520, 2012.

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

[18] N. Sarafianos, T. Giannakopoulos, and S. Petridis, "Audio-visual speaker diarization using fisher linear semi-discriminant analysis," *Multim. Tools Appl.*, vol. 75, no. 1, pp. 115–130, 2016.

[19] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," *CoRR*, vol. abs/2011.07557, 2020.

[20] R. Shashidhar, S. Patilkulkarni, and S. Puneeth, "Combining audio and visual speech recognition using LSTM and deep convolutional neural network," *International Journal of Information Technology*, 2022.

[21] C. Li and Y. Qian, "Deep audio-visual speech separation with attention mechanism," in *Proc. ICASSP*, 2020, pp. 7314–7318.

[22] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "DNN driven speaker independent audio-visual mask estimation for speech separation," in *Proc. Interspeech*, 2018, pp. 2723–2727.

[23] J. Tian, X. Hu, and X. Xu, "Royalflush speaker diarization system for ICASSP 2022 multi-channel multi-party meeting transcription challenge," *CoRR*, vol. abs/2202.04814, 2022.

[24] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3fd: Single shot scale-invariant face detector," in *Proc. ICCV*, 2017, pp. 192–201.

[25] 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. ACM Multimedia*, 2021, pp. 3927–3935.

[26] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks)," in *Proc. ICCV*, 2017, pp. 1021–1030.

[27] 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, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *CoRR*, vol. abs/2110.13900, 2021.

[28] A. Wuerkaixi, Y. Zhang, Z. Duan, and C. Zhang, "Rethinking audio-visual synchronization for active speaker detection," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022.

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

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[31] 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.

[32] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J. Bonastre, "NIST RT'05S evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings," in *International Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[33] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1079–1092, 2021.

[34] E. el Khoury, C. Sénac, and P. Joly, "Audiovisual diarization of people in video content," *Multim. Tools Appl.*, vol. 68, no. 3, pp. 747–775, 2014.