MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

LATE AUDIO-VISUAL FUSION FOR IN-THE-WILD SPEAKER DIARIZATION

Pan, Zexu; Wichern, Gordon; Germain, François G; Subramanian, Aswin; Le Roux, Jonathan

TR2024-029 March 19, 2024

Abstract

Speaker diarization has been well studied for constrained scenarios but little explored for in-the-wild videos, which have more speakers, shorter utterances, and inconsistent on-screen speakers. We address this gap by proposing an audio-visual diarization model which combines audio-only and visual-centric sub-systems via late fusion. For audio, we improve the attractor-based end-to-end system EEND-EDA with an attention mechanism and a speaker recognition loss to handle the larger speaker number and retain the speaker identity across recordings. The visual-centric sub-system leverages facial attributes and lip-audio synchrony for identity and speech activity estimation of on-screen speakers. Both sub-systems surpass the state of the art (SOTA) by a wide margin, with the fused audio- visual system achieving a new SOTA on the AVA-AVD benchmark.

Hands-free Speech Communication and Microphone Arrays (HSCMA) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

LATE AUDIO-VISUAL FUSION FOR IN-THE-WILD SPEAKER DIARIZATION

Zexu Pan, Gordon Wichern, François G. Germain, Aswin Subramanian, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

Speaker diarization has been well studied for constrained scenarios but little explored for in-the-wild videos, which have more speakers, shorter utterances, and inconsistent on-screen speakers. We address this gap by proposing an audio-visual diarization model which combines audio-only and visual-centric sub-systems via late fusion. For audio, we improve the attractor-based end-to-end system EEND-EDA with an attention mechanism and a speaker recognition loss to handle the larger speaker number and retain the speaker identity across recordings. The visual-centric sub-system leverages facial attributes and lip-audio synchrony for identity and speech activity estimation of on-screen speakers. Both sub-systems surpass the state of the art (SOTA) by a wide margin, with the fused audio-visual system achieving a new SOTA on the AVA-AVD benchmark.

***Index Terms*—** Speaker diarization, EEND-EDA, attention attractors, speaker recognition, audio-visual

1. INTRODUCTION

There is a long history of research looking to extract the rich information contained in speech signals, e.g., speaker localization, speech recognition, and emotion recognition [1–3]. However, these algorithms have often been optimized only for the case of isolated speech, making an effective preprocessing algorithm to find "who spoke when," i.e., speaker diarization [4], highly desirable. Classical audio-only diarization algorithms [4, 5] have typically followed a multi-stage cascaded approach with voice activity detection (VAD), frame segmentation, speaker embedding extraction, and clustering, with each stage optimized independently. With such an approach, errors tend to accumulate, resulting in sub-optimal performance. Recently, end-to-end (E2E) algorithms, such as the E2E neural diarization (EEND) with an encoder-decoder based attractor (EDA), named EEND-EDA [6, 7], have gained increasing attention due to their ability to flexibly handle an unknown number of speakers. Permutation invariant training (PIT) [8, 9] is used to address the speaker order ambiguity between network outputs and ground-truth labels.

As does human speech perception often rely on various sensory stimuli, such as observing lip movements [10] or body gestures [11], audio-visual diarization algorithms leverage such synergies between speech signals and visual features. The WST model [12] enrolls speakers based on audio-visual correspondence in a cascaded diarization system. E2E methods [13, 14] fuse audio and visual representations, training similarly to audio-only EEND [15].

Both audio-only and audio-visual diarization models have demonstrated remarkable performance in constrained meetings or conversation scenarios [17], where there are rarely more than 10 people and no off-screen speakers. However, they are largely untested for in-the-wild recordings such as those of the recent AVA-AVD dataset [16], where a 5-minute movie clip can contain up to 20 speakers, with much shorter typical utterances (cf. the statistics in Fig. 1). Such conditions especially test the robustness of PIT with its factorial complexity with the number of speakers, and the capacity of the EDA module to decode representative attractors. Recordings also contain various background noises, sound effects, and music, off-screen utterances/speakers, and irrelevant speakers.

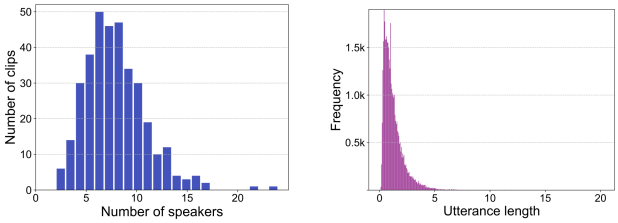

Fig. 1: Histograms of the number of speakers (left) and the utterance length in seconds (right) of the audio recordings in the AVA-AVD dataset (plots reproduced from [16]).

The recent AVR-Net [16] and DyViSE [18] target in-the-wild videos by enhancing the speaker embedding extraction stage through audio-visual early fusion in the cascaded diarization framework. Alternatively, we advocate for late fusion, allowing the visual signals to explicitly play at least two important roles that could help diarization: the synchronization between the lip movements and the speech signals provides strong hints about the speech activity [19], and the facial attributes provide robust evidence about speaker identity [20].

In this work, we build an audio-visual speaker diarization system targeting in-the-wild videos. We propose a model named AV-EEND-EDA++, depicted in Fig. 2. It comprises an audio-only sub-system, EEND-EDA++, shown in Fig. 3, and a visual-centric sub-system, V-AHC, shown in the top part of Fig. 2. We build EEND-EDA++ upon EEND-EDA, proposing an attention-based EDA module to enhance the network's capacity when decoding a large number of speaker attractors, and to jointly train the speaker attractors on speaker recognition to enhance the discriminative power of the attractor representation. V-AHC explicitly leverages visual signals to perform on-screen face tracks diarization. It uses an active-speaker detection technique [19] to enhance the on-screen speaker activity and speaker change detection, followed by agglomerative hierarchical clustering (AHC) based on face recognition results [21] for speaker clustering. The two sub-systems are combined with a permutation-invariant late-fusion technique based on speaker activity probability.

Experimental results show that EEND-EDA++ better preserves speaker identity across recordings thanks to the speaker recognition loss. Both our proposed EEND-EDA++ and V-AHC outperform existing audio-visual diarization algorithms on the AVA-AVD benchmark [16], with our AV-EEND-EDA++ achieving the new SOTA in terms of diarization and Jaccard error rate.

2. PROPOSED AV-EEND-EDA++

2.1. Audio-only EEND-EDA++

Related work: EEND-EDA [6] is an E2E diarization model which encodes audio features into audio embeddings $e_t \in \mathbb{R}^D$, $t \in [1, \dots, T]$, using Transformer encoders (SA-EEND) without

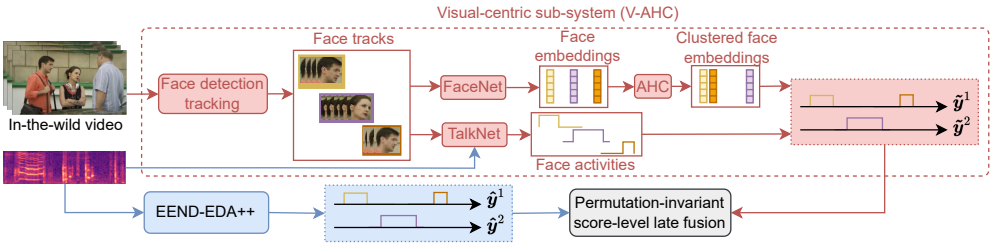

Fig. 2: Our proposed audio-visual speaker diarization model named AV-EEND-EDA++. The blue path is the audio-only sub-system named EEND-EDA++, while the red path is the visual-centric clustering-based sub-system named V-AHC. The diarization results of the EEND-EDA++ and V-AHC sub-systems are fused with a permutation-invariant score-level late fusion. The figure is best viewed in color.

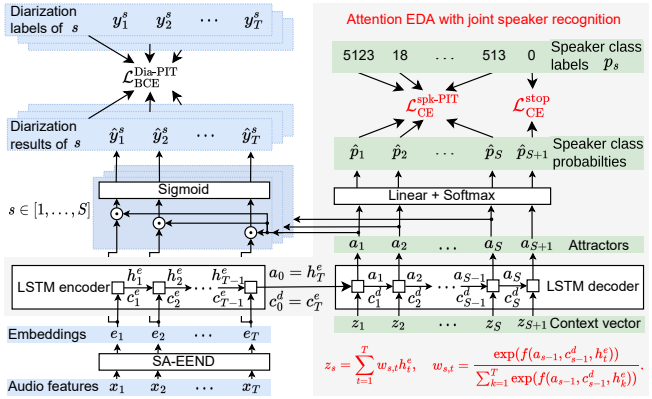

Fig. 3: Our proposed EEND-EDA++. We introduce an attention mechanism in EDA and train the attractors on speaker recognition. The symbol $\odot$ is the inner product. Novelties are marked in red.

positional encoding, where D is the embedding dimension and T the number of audio frames. In the EDA module, an LSTM encoder encodes the time-shuffled e_t, and from its last hidden and cell states, an LSTM decoder estimates a flexible number of speaker attractors $a_s \in \mathbb{R}^D$, $s \in [1, \dots, S]$, controlled by a stop flag. Speaker activity at time t for each speaker is obtained by taking the inner product of each attractor with the audio embedding e_t and applying a sigmoid.

Attention EDA: We propose using an attention LSTM decoder, such that the estimation of attractor a_s is conditioned on a distinct context vector z_s instead of the zero vector in EEND-EDA, with z_s computed as a weighted sum of the EDA encoder outputs h_t^e:

$$z_s = \sum_{t=1}^{T} w_{s,t} h_t^e, \tag{1}$$

$$w_{s,t} = \frac{\exp(f(a_{s-1}, c_{s-1}^d, h_t^e))}{\sum_{\tau=1}^{T} \exp(f(a_{s-1}, c_{s-1}^d, h_\tau^e))}, \tag{2}$$

where $f(\cdot)$ is a linear layer with hyperbolic tangent (tanh) activation. We denote this method EEND-EDA+Att. We use SA-EEND with positional encoding, and do not shuffle e_t before passing them to the EDA LSTM encoder, contrary to what EEND-EDA did.

Objective functions for speaker recognition: The EEND-EDA attractors are trained via binary classification, in which the attractors iteratively predict whether there are remaining speakers or not. The attractors are thus not trained to encode speaker identity that remains valid across recordings. Since each speaker's activity is conditioned on its attractor, it is expected that an attractor better correlated with speaker identity will benefit the diarization task. Inspired by speaker extraction approaches in which the extraction is conditioned on an attractor that is jointly trained to recognize the speaker [22–24], we also train the attractors using a speaker recognition loss, such that the attractors explicitly represent speaker information. We refer to this approach as EEND-EDA+Spk.

We employ a softmax layer to transform each attractor a_s into a probability distribution $\hat{p}_s$ over the speakers in the dataset. During training, we define the ground-truth speaker class labels p_s, for $s \in [1, \dots, S]$, using the actual number of speakers S in a recording, where $p_s(j)$, for $j \in [1, \dots, J]$, is a binary label indicating if the s-th speaker in the recording is the j-th speaker in the speaker dataset, and J denotes the total number of speakers in the training set. We introduce an additional class representing "Not a speaker," corresponding to $j = 0$, and use it as a stop flag, such that the network learns to stop decoding attractors at inference time whenever an attractor falls into this class. The speaker classification objective function for the first S attractors is defined as:

$$\mathcal{L}_{\text{CE}}^{\text{spk-PIT}} = \arg\min_{\pi \in \mathcal{P}_S} -\sum_{s=1}^{S} \sum_{j=0}^{J} p_{\pi(s)}(j) \log \hat{p}_s(j), \tag{3}$$

where $\mathcal{P}_S$ denotes the set of permutations over $\{1, \dots, S\}$. We use PIT to find the optimum permutation order between the estimated attractors and the speaker labels, relying on Sinkhorn's algorithm (SinkPIT) [25] to avoid the factorial complexity in the number of speakers. At the same time, the S+1-th attractor is trained to fall into the "Not a speaker" class with:

$$\mathcal{L}_{\text{CE}}^{\text{stop}} = -\sum_{j=0}^{J} p_{S+1}(j) \log \hat{p}_{S+1}(j) = -\log \hat{p}_{S+1}(0). \tag{4}$$

Overall objective function: Our final audio-only sub-system EEND-EDA++ is shown in Fig. 3 and combines EEND-EDA+Att and EEND-EDA+Spk. The overall objective function is:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{BCE}}^{\text{Dia-PIT}} + \beta(\mathcal{L}_{\text{CE}}^{\text{spk-PIT}} + \alpha \mathcal{L}_{\text{CE}}^{\text{stop}}), \tag{5}$$

where

$$\mathcal{L}_{\text{BCE}}^{\text{Dia-PIT}} = \arg\min_{\pi \in \mathcal{P}_S} -\sum_{s=1}^{S} \sum_{t=1}^{T} (\gamma y_t^{\pi(s)} \log(\hat{y}_t) + (1 - y_t^{\pi(s)}) \log(1 - \hat{y}_t)) \tag{6}$$

is the objective of the diarization task, with y_t and $\hat{y}_t$ respectively denoting the ground-truth and estimated speaker activity probabilities at frame t. We again use the SinkPIT algorithm here. Different loss terms are balanced using scalar weights α and β. We also impose a scalar weight γ on the positive class (speaker active) to account for class imbalance when there are many speakers involved.

2.2. Proposed visual-centric V-AHC

Diarization in in-the-wild videos has unique challenges including speakers who are partially or completely off-screen and irrelevant on-screen speakers. However, visual signals such as face recordings are robust to acoustic noises, providing a strong cue about speaker identity and the places of articulation that discriminate between speech and non-speech signals. We aim to leverage the available visual signals, specifically the detected face tracks, to per-

form an on-screen speaker diarization. Our proposed visual-centric clustering-based model named V-AHC is illustrated in the upper half of Fig. 2, in the red dotted box.

Speech activity extraction: For each face track, we perform audio-visual active speaker detection using a pre-trained TalkNet model [19], that leverages the synchronization between the lip movements and the audio signals to determine the frame-level speech activity of the face. The use of audio signals is vital as there could be a talking face without audible speech sometimes.

Speaker identity extraction: To determine the identity of the speaker for a face track, we utilize deep face recognition models that have demonstrated remarkable performance in practical applications. Specifically, we randomly sample up to 50 images from the face track and average their embeddings extracted from a pre-trained FaceNet model [21].

Agglomerative hierarchical clustering: To identify speaker clusters, we use agglomerative hierarchical clustering (AHC) [26] on all face tracks. The distances between face tracks are computed as the negative cosine similarity of their face embeddings averaged over 50 random frames. The diarization result for each speaker cluster is determined by combining the speech activities obtained with TalkNet for the respective face tracks. In cases where no face is detected for some segment, we set the corresponding speech activity to zero.

2.3. Permutation-invariant late fusion

The audio-only sub-system performs diarization on the entire recording, but is negatively impacted by acoustic noise. In contrast, the visual-centric sub-system is resilient to acoustic noise but will miss speech activities from off-screen speakers. To exploit the benefits of both models, we propose a score-level late-fusion strategy combining them by comparing their speech activity probabilities, as detailed next. As such, we aim to capitalize on their complementary strengths and enhance the overall diarization performance.

To determine the one-to-one correspondence between the audio and visual results, we first silence-pad the less-speaker modality to have the same speaker number as the other one, and then compare every audio diarization result $\hat{\boldsymbol{y}}^s$ with every visual diarization result $\tilde{\boldsymbol{y}}^{s'}$ and calculate the matching score to be the summation of the audio scores $\hat{y}_t^s$ for all t where visual scores $\tilde{y}_t^{s'}$ show active speech. The best correspondence is determined by the highest matching score over all permutations:

$$\underset{\pi \in \mathcal{P}_{\max(S,S')}}{\arg\max} \sum_{s=1}^{S} \sum_{t=1}^{T} \hat{y}_t^s \mathbb{1}[\tilde{y}_t^{\pi(s)} = 1], \tag{7}$$

where we consider $\tilde{y}_t^{s'} = 0$ if $s' > S'$. For the best pairs, we replace the audio score $\hat{y}_t^s$ with the visual score $\tilde{y}_t^{\pi(s)}$ when the latter shows active speech, because we found the visual score to be more reliable. In some cases where we know the speaker overlapping ratio is small in the training data distribution, like AVA-AVD, we employ a post-processing technique that mutes the other speakers at those time frames where the visual score shows one speaker is active.

3. EXPERIMENTAL SETUP

3.1. Datasets

In-the-wild dataset AVA-AVD: The AVA-AVD dataset [16] is one of the few publicly available in-the-wild audio-visual diarization datasets. It was built upon the AVA-Active Speaker dataset [27], which consists of multilingual movies depicting diverse daily activities, in order to foster the development of diarization methods for challenging conditions. The train, validation, and test sets consist of 243, 54, and 54 videos respectively, 5 minutes each.

Simulated proxy dataset VoxCeleb2-AVD: Since AVA-AVD is small, prior works [16] pre-train the models on a large simulated dataset. We use the VoxCeleb2 dataset [28] to simulate a pre-training dataset for the audio-only models, which we name VoxCeleb2-AVD. We simulate 2×10^5, 500, and 500 recordings for the train, validation, and test sets respectively, and each is 5 minutes long. The test set has distinct speakers from train and validation sets. We simulate VoxCeleb2-AVD close to AVA-AVD, following the distributions shown in Fig. 1. We also randomly sample music and noise clips from the MUSAN dataset [29] and the Freesound Dataset 50k (FSD50K) [30], and add them to each audio recording. We follow the *cocktail fork* [31] protocol in setting the energy levels between speech, noise, and music in VoxCeleb2-AVD.

3.2. Implementation details

We train the audio-only sub-system on 5-minute recordings, and the processing of input audio features and the audio-only sub-system model settings follows EEND-EDA [6]. We set the scalar weights as $\alpha = 0.01$ and $\beta = 0.1 \times 0.92^{Epoch}$, where *Epoch* is the epoch number, as we empirically found that the speaker recognition loss is important during initial training iterations, but less beneficial as the model converges for diarization. We set $\gamma = 5$. The model is trained using the Adam optimizer with the learning rate schedule as in [6] and $10\,000$ warm-up steps. The training of EEND-EDA++ requires the speaker identity labels in the dataset, and while VoxCeleb2 has such labels, AVA-AVD does not. Since the VoxCeleb2 training set has 5994 speakers, we can hope to find speakers with similar voice characteristics as speakers in AVA-AVD. We thus map every speaker of AVA into its closest speaker in the VoxCeleb2 training set, based on the L_2 distance between their speaker embedding representations extracted using RawNet3 [32], as proxy for the speaker label.

3.3. Baselines

We present the results of 4 speaker diarization baselines on AVA-AVD. WST [12] is an audio-visual system that uses audio-visual correlation to help first enroll the speakers and then diarize. VBx [4] is a recent audio-only cascaded system using Bayesian clustering, with 2 variants, VBx-ResNet34 and VBx-ResNet101. AVR-Net [16] is an audio-visual cascaded system built upon VBx-ResNet34 and TalkNet. DyViSE [18] is an audio-visual cascaded system, which denoises audio with visual information in a latent space and integrates facial features to obtain identity discriminative embeddings.

4. RESULTS

4.1. Audio-only models

Results on VoxCeleb2-AVD: In Table 1, we present the results of the baseline EEND-EDA and our proposed EEND-EDA++ trained and evaluated on the simulated VoxCeleb2-AVD dataset. EEND-EDA++ achieves the best diarization error rate (DER) and Jaccard error rate (JER). We also present two ablation studies of EEND-EDA++ on that same VoxCeleb2-AVD. EEND-EDA+Spk, trained with our speaker recognition loss but without the attention mechanism, performs badly in terms of DER and JER due to higher MS. This is probably because the vanilla EDA has limited capacity in producing representative speaker attractors, thus the speaker loss adversely affects the model training. EEND-EDA+Att, which has the attention mechanism in EDA but is not trained with our speaker recognition loss, outperforms EEND-EDA in DER, but is not better than EEND-EDA++ except for the MS submetric.

Results on AVA-AVD: In Table 2, we compare EEND-EDA++ with baselines on the AVA-AVD benchmark. The results of systems 1-4 are taken from [16]. The previous SOTA is system 3, which reports a DER of 70.9%. Without pre-training, EEND-EDA in system 6

Table 1: Results on the VoxCeleb2-AVD dataset. We report DER [%], which is the sum of missed speech (MS), false alarm (FA), and speaker error (SE). We also report JER [%]. The lower the better for all metrics. All systems (Sys.) in this paper use a collar of 0.25 s [33].

Sys.	Model	MS	FA	SE	DER	JER
7	EEND-EDA [6]	17.4	9.1	18.9	45.4	66.7
9	EEND-EDA++	15.8	6.7	18.2	40.8	62.8
15	EEND-EDA+Spk	25.1	6.4	17.6	49.1	71.2
16	EEND-EDA+Att	14.6	8.6	20.8	43.9	66.6

Table 2: Results on the AVA-AVD benchmark. M indicates the modality used, either audio (A) or audio-visual (AV). PT indicates if the model is pre-trained, either on a VoxCeleb2-based dataset (Sys. 1-5) as in [16], on our VoxCeleb2-AVD (Sys. 7-11 and 13-14), or on the AVA and some face recognition datasets (Sys. 12). FT indicates if the model is fine-tuned on AVA-AVD. *System 11 uses the ground-truth speaker number at inference.

Sys.	Model	M	PT	FT	MS	FA	SE	DER	JER
1	WST [12]	AV			11.6	40.6	36.1	88.4	-
2	VBx-ResNet34 [4]	A			8.7	44.6	35.3	88.5	-
3	VBx-ResNet101 [4]	A	✓	✓	8.7	44.6	17.6	70.9	-
4	AVR-Net [16]	AV			8.7	44.6	20.1	73.3	-
5	DyViSE [18]	AV			11.1	24.2	35.9	71.2	-
6	EEND-EDA [6]		✗	✓	46.6	27.3	22.6	96.4	94.7
7	EEND-EDA	A	✓	✗	24.2	2.8	24.4	51.4	83.5
8	EEND-EDA		✓	✓	28.5	5.3	15.1	48.9	78.5
9	EEND-EDA++			✗	22.8	6.8	20.8	50.4	80.5
10	EEND-EDA++	A	✓	✓	25.0	5.6	17.0	47.6	76.4
11	EEND-EDA++*			✓	23.2	7.8	16.7	47.7	74.3
12	V-AHC				56.2	0.9	9.2	66.3	78.4
13	AV-EEND-EDA++	AV	✓	✓	18.5	7.2	20.4	46.1	68.8
14	AV-EEND-EDA++†				20.0	4.7	20.4	45.1	76.0

performs badly with a DER of 96.4%. Systems 7-10 are pre-trained on our VoxCeleb2-AVD, and they all outperform the baselines by a wide margin in terms of DER, with our EEND-EDA++ achieving the best DER of 47.6% and JER of 76.4%. It is worth mentioning that EEND-EDA++ is an audio-only sub-system, but still outperforms the audio-visual baselines by a wide margin.

Results on AVA-AVD with oracle speaker counting: For EEND-EDA++, systems 11 and 10 are the same, except that the former uses the ground-truth number of speakers at inference. Both get similar DER, but system 10 is lagging behind system 11 by 2% in terms of JER. We observe informally that system 10 typically underestimates the speaker number. Nevertheless, system 10 still achieves the best DER and JER among non-oracle systems 1-10.

Visualization: In Figs. 4 and 5, we show the t-SNE plots of embeddings e_t within an example video in the VoxCeleb2-AVD and the AVA-AVD datasets. Yellow represents the non-speech region, while each of the other colors represents a speaker. For both datasets, the embedding clusters of our system 9 are separated further apart than the baseline system 7, showing that our speaker-recognition objective pushes the embeddings from different speakers away from each other. In Fig. 6, we show the t-SNE plot of attractors across different videos from the VoxCeleb2-AVD dataset. We randomly selected 3 speakers not seen during training, and if these speakers appear in a video, we match the attractors to the speaker labels by computing the best permutation between the estimated diarization streams and ground-truth diarization streams. We then plot the attractors that were matched to the three speakers, with one color per speaker. We see that our system 9 has more obvious clusters for the attractors than baseline system 7, which means that the speaker identities are better matched across videos. Note we cannot show attractor plots for AVA-AVD as it does not have ground-truth speaker labels.

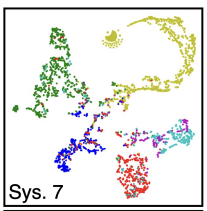

Fig. 4: e_t t-SNE plot of a VoxCeleb2-AVD video.

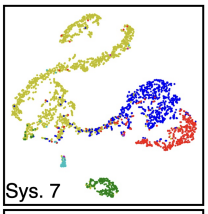

Fig. 5: e_t t-SNE plot of an AVA-AVD video.

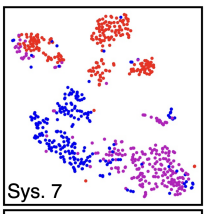

Fig. 6: a_s t-SNE plot of VoxCeleb2-AVD videos.

4.2. Visual-centric model

In Table 2, our proposed visual-centric on-screen speaker diarization sub-system V-AHC achieves a DER of 66.3%, which surpasses the baseline systems 1 to 5. V-AHC has a very high MS of 56.2%, which shows that the off-screen speakers are indeed a frequent feature in the AVA-AVD dataset. However, V-AHC has very low FA of 0.9%, showing the strength of the visual signals when present.

4.3. Audio-visual models

We fuse the visual-centric sub-system V-AHC with the audio-only sub-system 10, resulting in the audio-visual system 13. System 13 outperforms system 10 with a reduction of 1.5% in DER and 7.6% in JER. Compared to its audio sub-system counterpart, the MS decreases significantly, but the FA and SE increase. This may be because the final number of speakers is set as the maximum of the audio and visual sub-systems. The number of output streams thus increases, which decreases MS but introduces more FA and SE.

We also present system 14, in which instead of fusing the full-length recording-level visual diarization results to the audio diarization results as described in Section 2.3, we fuse each face activities output from TalkNet (face-track-level) to one of the audio diarization streams, with the same score-level decisions. System 14 outperforms system 10 with an absolute reduction of 2.5% in DER and 0.4% in JER. Compared to its audio sub-system counterpart, MS and FA decrease, but SE increases. This is because the final number of output speakers is the same as the audio sub-system, so the visual signal generally helps the MS and FA. However, the fusion without considering visual speaker identity causes SE to increase. Although the improvements in DER are here better compared to recording-level audio-visual fusion, the JER here only improves marginally. Overall recording-level fusion is preferred as the improvement on JER is large which indicates per-speaker diarization evaluation improves a lot, showing the importance of utilizing the identity information from the visual signals using FaceNet and the AHC algorithm.

5. CONCLUSION

We studied the speaker diarization problem for in-the-wild videos. We proposed a late audio-visual fusion model, AV-EEND-EDA++, that comprises an audio-only sub-system, EEND-EDA++, and a visual-centric sub-system, V-AHC. For audio-only sub-systems, we show that the speaker identity is better preserved in our EEND-EDA++ with our speaker recognition loss. Our proposed sub-systems and the fused audio-visual model outperform SOTA on the AVA-AVD benchmark.

6. REFERENCES

[1] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DoA estimation with an audio-visual fusion mechanism," in *Proc. ICASSP*, 2021.

[2] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," in *Proc. Interspeech*, 2022.

[3] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," in *Proc. Interspeech*, 2020.

[4] 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," *Comput. Speech Lang.*, vol. 71, pp. 101254, 2022.

[5] 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, pp. 101317, 2022.

[6] 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.

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

[8] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016.

[9] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017.

[10] Z. Pan, M. Ge, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3032–3045, 2022.

[11] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Process. Lett.*, vol. 29, pp. 1467–1471, 2022.

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

[13] Q. Qiu, T. Xu, and E. Chen, "Visual-enhanced end-to-end neural diarization," in *Proc. Int. Conf. Image Video Signal Process.*, 2022.

[14] M.-K. He, J. Du, and C.-H. Lee, "End-to-end audio-visual neural speaker diarization," in *Proc. Interspeech*, 2022.

[15] 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.

[16] E. Z. Xu, Z. Song, C. Feng, M. Ye, and M. Z. Shou, "AVA-AVD: Audio-visual speaker diarization in the wild," in *Proc. ACM Multimedia*, 2022.

[17] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[18] 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.

[19] 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.

[20] S. I. Serengil and A. Ozpinar, "LightFace: A hybrid deep face recognition framework," in *Proc. Innov. Intell. Syst. Appl. Conf.*, 2020.

[21] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015.

[22] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *Proc. ASRU*, 2017.

[23] Z. Pan, R. Tao, C. Xu, and H. Li, "MuSE: Multi-modal target speaker extraction with visual cues," in *Proc. ICASSP*, 2021.

[24] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[25] H. Tachibana, "Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using Sinkhorn's algorithm," in *Proc. ICASSP*, 2021.

[26] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classif.*, vol. 1, pp. 7–24, 1984.

[27] 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.

[28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

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

[30] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2021.

[31] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, "The cocktail fork problem: Three-stem audio separation for real-world soundtracks," in *Proc. ICASSP*, 2022.

[32] J.-W. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022.

[33] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J. F. Bonastre, "NIST RT'05S evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2005.