



End-to-End Audio-Visual Neural Speaker Diarization

Mao-kui He¹, Jun Du^{1,*}, Chin-Hui Lee²

¹University of Science and Technology of China, HeFei, China

²Georgia Institute of Technology, Atlanta, GA, USA

✉jundu@ustc.edu.cn

Abstract

In this paper, we propose a novel end-to-end neural-network-based audio-visual speaker diarization method. Unlike most existing audio-visual methods, our audio-visual model takes audio features (e.g., FBANKs), multi-speaker lip regions of interest (ROIs), and multi-speaker i-vector embeddings as multi-modal inputs. And a set of binary classification output layers produces activities of each speaker. With the finely designed end-to-end structure, the proposed method can explicitly handle the overlapping speech and distinguish between speech and non-speech accurately with multi-modal information. I-vectors are the key point to solve the alignment problem caused by visual modality error (e.g., occlusions, off-screen speakers or unreliable detection). Besides, our audio-visual model is robust to the absence of visual modality, where the diarization performance degrades significantly using the visual-only model. Evaluated on the datasets of the first multi-model information based speech processing (MISP) challenge, the proposed method achieved diarization error rates (DERs) of 10.1%/9.5% on development/eval set with reference voice activity detection (VAD) information, while audio-only and video-only system yielded DERs of 27.9%/29.0% and 14.6%/13.1% respectively.

Index Terms: audio-visual speaker diarization, end-to-end, multimodal information, visual VAD, MISP challenge

1. Introduction

Speaker diarization is the process of partitioning speech segments according to the speaker identity. Its applications lay in broadcast news, meetings, telephone conversations, etc [1]. It also helps automatic speech recognition (ASR) performance in multi-speaker conversation scenarios in meetings (ICSI [2], AMI [3]) and home environments (CHiME-6 [4]). Speaker diarization is very challenging when unimodal data is available. Audio-based diarization suffers from inherently ambiguous acoustic data because they contain mixed speech signals emitted by several speakers, corrupted by reverberations, interference sounds, and background noises. For visual-based data, speakers may not face the camera, move in a multi-party interaction way, or be occluded by other speakers. Hence, lip or face information is not always reliable throughout the conversation.

Conventional audio-only systems, which mainly include voice activity detection (VAD), speech segmentation, speaker feature extraction, and speaker clustering, are widely used in speaker diarization [5]. However, this framework inherently makes an assumption that every segment can only be assigned with a single speaker label. Re-segmentation is considered to handle overlapping segments, but overlap detection is still a challenging task. End-to-end neural speaker diarization (EEND) [6] and target-speaker voice activity detection (TS-VAD) [7] were proposed to deal with the overlap problem by

directly predicting each speaker's activeness for each frame, which can fundamentally handle the speakers overlap regions in the recordings. However, those audio-based methods still struggle to diarize the low-quality, high-overlapping speech [7, 8]. Vision-only methods can consist of face or lip tracking [9, 10] for multiple persons and visual voice activity detection [11, 12] for each individual person. Although vision-based systems are easier to process overlapping segments, multi-speaker visual tracking is also a challenging task. Furthermore, visual features are often missing due to occlusion, off-screen speakers, or unreliable detection. Therefore, visual-only diarization performances are still not satisfactory for many applications.

Facial attributes and lip motion are highly related to speech [13]. Therefore, several methods are proposed to seek the synergy between utterances and lip movements in the last decade by adopting techniques such as mutual information [14], canonical correlation analysis [15], and deep learning [16, 17, 18]. In recent works, audio-visual correspondence is also used for associating talking faces and voice tracks [19, 20]. Different modalities are fused by linear combination [17, 19], temporal alignment [21, 22], and Bayesian method [23]. Moreover, sound-source localization using a circular or linear microphone arrays provides horizontal (azimuth) speech directions. Another cross-modal relationship can be established by mapping this sound direction onto the image plane [23, 17].

In this paper, we mainly focus on exploring the effects of lip motion and speech on speaker diarization using high-definition lip ROI and single-channel audios. By manually removing fragments of lip ROI, we can easily compare the impact of different degrees of lip miss on speaker diarization. Different from deep audio-visual synchronisation network based methods [16, 17, 18] and clustering on audio-visual pairs scored by audio-visual relation network [24], we propose an end-to-end audio-visual neural network that directly predicts speech probabilities for all speakers simultaneously inspired by audio-only end-to-end frameworks. In addition to simply stitching the extracted frame-level audio embeddings with synchronized frame-level multi-speaker lip embeddings, we take i-vectors of all speakers as additional inputs to address the alignment problem when lip is missing in inputs caused by occlusions, speakers out of cameras or unreliable lip detection. To handle flexible number of speakers in the audio-visual datasets, we utilize the visual embeddings and i-vectors as inputs in our audio-visual framework when the number of speakers is not equal to the designed input nodes. By evaluating on MISP mid-field audio and mid-field video with cameras for each speaker, our proposed method achieves the best results compared to audio-only and video-only methods. Meanwhile, the experimental results under different degrees of lip missing also show that the proposed method is more robust than the video-only systems.

*corresponding author

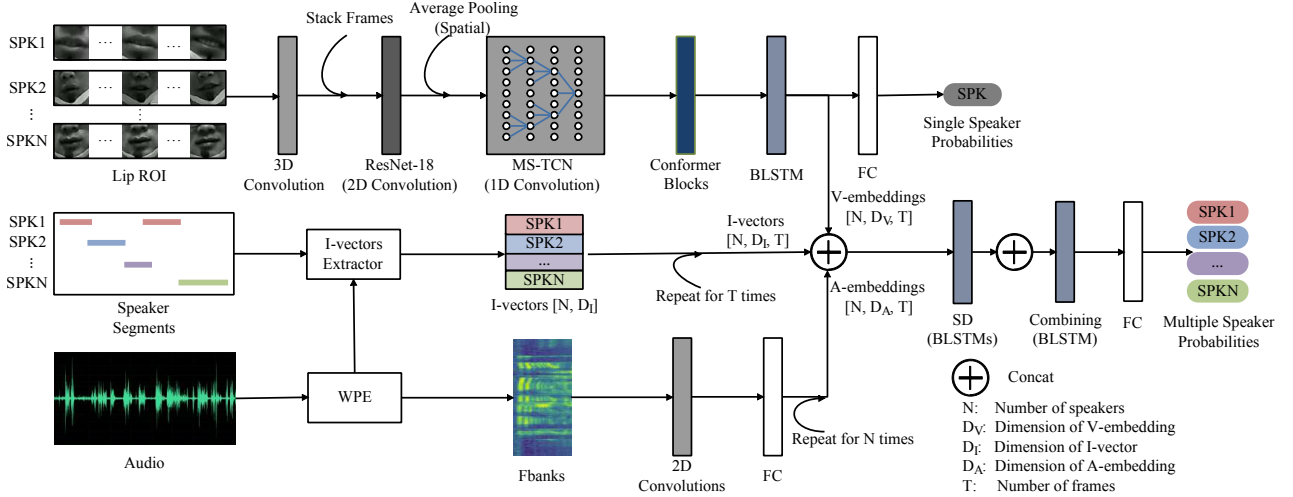


Figure 1: The illustration of network structure

2. Audio-Visual Diarization Datasets

Currently, various audio-visual diarization datasets have been released. AVDIAR [23] is released to enable audio-visual scene analysis of unstructured informal meetings and gatherings, but it does not provide the training set. VoxConverse [19] is a challenging dataset where the diverse speaker diarization data from ‘in the wild’ videos are included. The AMI corpus [3] is a multi-modal dataset consisting of 100-hour meeting recordings. The close-talking and far-field audio, individual and room-view video are all available. AVA-AVD [24] is recorded recently to cover the diverse scenes and complicated acoustic conditions, which contains completely off-screen speakers.

Recently, the newly released multi-modal information based speech processing (MISP) dataset [25] provides more than 100-hour audio and video recordings of several people in a living room watching and chatting while interacting with a smart speaker/TV. These sessions are usually accompanied by high overlap ratios in multi-talker conversations and real domestic noise backgrounds such as TV, air conditioning and movements. In this study, we evaluate the proposed method using MISP mid-field audios and mid-field videos for the following reasons. First, the MISP dataset provides a large amount of multi-modal data for supervised training and testing. Second, we mainly focus on exploring the effects of lip motion and speech on speaker diarization, rather than the relationship between the sound source location and the position of the person in the image plane. Finally, the lip ROI can be clearly extracted from mid-field video of the speaker, which ensures that we can evaluate the performance of the proposed method in partially missing visual modality by manually setting different degrees of lip missing ratios.

3. Proposed Method

3.1. Probabilistic Model

To handle all possible cases in our model, let N be the maximum number of speakers in the whole dataset. Hence at time t we have N lip ROIs $\mathbf{X}_t = (X_{t,1}, \dots, X_{t,n}, \dots, X_{t,\hat{N}}, \dots, X_{t,N}) \in \mathbb{R}^{W \times H \times N}$, where the \hat{N} is the number of speakers in the current session. The observed random variable $X_{t,n} \in \mathbb{R}^{W \times H}$ is the lip figure with width W and height H of person n at

time t . We randomly select a fake lip from the silent lips to complement $n > \hat{N}$ and the case where the existing human lips are missing. For the audio-based data, without loss of generality, we use $\mathbf{Y}_t = (Y_{t,1}, \dots, Y_{t,f}, \dots, Y_{t,F}) \in \mathbb{R}^F$ to denote the F -dimensional FBANKs of single-channel audio signals. The time series $\mathbf{X}_{1:T} = \{\mathbf{X}_1, \dots, \mathbf{X}_t, \dots, \mathbf{X}_T\}$ and $\mathbf{Y}_{1:T} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_t, \dots, \mathbf{Y}_T\}$ represent the visual and audio observations, respectively.

The objective of speaker diarization is to assign speech signal to persons. For this purpose, we introduce a time series of discrete variables $\mathbf{S}_{1:T} = \{\mathbf{S}_1, \dots, \mathbf{S}_t, \dots, \mathbf{S}_T\} \in \{0, 1\}^{N \times T}$ where the vector $\mathbf{S}_t = (S_{t,1}, \dots, S_{t,n}, \dots, S_{t,N}) \in \{0, 1\}^N$ has binary-valued elements so that $S_{t,n} = 1$ if speaker n speaks during the time-step t , and $S_{t,n} = 0$ if speaker n is silent when $n \leq \hat{N}$. The $S_{t,n}$ is 0 from 1 to T when n is a fake person where $n > \hat{N}$. For brevity, we ignore the subscript $1 : T$ hereinafter. To avoid alignment problem where model is confused about which speaker to assign for current speech where all lips show silence caused by lip missing problem, we introduce i-vector [26] $\mathbf{I} = \{\mathbf{I}_1, \dots, \mathbf{I}_n, \dots, \mathbf{I}_{\hat{N}}, \dots, \mathbf{I}_N\} \in \mathbb{R}^{D_I \times N}$ as speaker related observations where $\mathbf{I}_n \in \mathbb{R}^{D_I}$ is the i-vector of speaker n . When $n > \hat{N}$, we randomly select fake speaker i-vector that is not belong to this session. Then the temporal speaker diarization problem can be formulated as finding the most probable time series of state $\hat{\mathbf{S}}$ among all possible speaker label sequences \mathcal{S} with the observed variables $(\mathbf{X}, \mathbf{Y}, \mathbf{I})$, as follows:

$$\hat{\mathbf{S}} = \arg \max_{\mathbf{S} \in \mathcal{S}} P(\mathbf{S} | \mathbf{X}, \mathbf{Y}, \mathbf{I}) \quad (1)$$

where $P(\mathbf{S} | \mathbf{X}, \mathbf{Y}, \mathbf{I})$ can be factorized using conditional independence assumption as follows:

$$P(\mathbf{S} | \mathbf{X}, \mathbf{Y}, \mathbf{I}) = P(\mathbf{S} | \mathbf{E}_V, \mathbf{E}_A, \mathbf{I}) P(\mathbf{E}_V | \mathbf{X}) P(\mathbf{E}_A | \mathbf{Y}) \quad (2)$$

where $\mathbf{E}_V = \{\mathbf{E}_{V_1}, \dots, \mathbf{E}_{V_n}, \dots, \mathbf{E}_{V_N}\} \in \mathbb{R}^{T \times D_V \times N}$ is the frame-level D_V -dimensional visual latent variables for N speakers, namely V-embedding. $\mathbf{E}_A \in \mathbb{R}^{T \times D_A}$ is the frame-level D_A -dimensional audio latent variables, namely A-embedding. Here, we assume the V-embedding and A-embedding are conditioned independently on the visual and audio observations respectively.

3.2. Visual Embedding

We build a visual network to compute V-embeddings of n -th person $\mathbf{E}_{V_n} \in \mathbb{R}^{T \times D_V}$ by directly adding several conformer blocks [27] and a BLSTM [28] layer to the original temporal convolutional networks for lipreading [29]. By projecting the V-embeddings \mathbf{E}_{V_n} to the frame-level speech/non-speech probabilities $\hat{S}_n^V = (\hat{S}_{1,n}^V, \dots, \hat{S}_{i,n}^V, \dots, \hat{S}_{T,n}^V) \in (0, 1)^T$ for n -th person through full connect (FC) layers, the whole network can be regarded as visual voice activity detection (V-VAD) module. After pre-training the visual network as a V-VAD task, V-embeddings are equipped with capability that represent the states of speaking or silent. Meanwhile, the visual-only speaker diarization results can be directly generated with V-VAD by combining all of persons' VAD in one session. The visual network is illustrated in the top row of Figure 1.

3.3. Audio Embedding

Before computing FBANKs, we firstly perform dereverberation on the original audio signals with NARA-WPE [30] as in [31]. The FBANKs are taken as the inputs of audio-based network with CNNs composed of 2D Convolution, BatchNorm, and Relu for each layer. Then, a fully connected layer projects high dimensional CNNs output to low dimensional A-embeddings \mathbf{E}_A . Unlike visual network, we don't pre-train audio network on any other tasks and directly optimize it with audio-visual decoding network. The audio network is illustrated in the bottom row of Figure 1.

3.4. Speaker Embedding

There are two kinds of assigning problems without speaker embeddings. When someone is speaking but his/her lip is missing, either none or random one would be labeled with speech in the output. When someone is speaking and the other one is not speaking but both of their lips are moving, both of two persons or a random person would be labeled with speech in the output. Those two errors occur when occlusions, off-screen speakers, error detection and the moving lips although someone is not speaking. In audio-only speaker diarization approaches [7], multi-speaker voice activities are estimated simultaneously by taking both audio feature and their i-vectors as inputs. Therefore, we introduce i-vectors to deal with the unreliable V-embedding problem in audio-visual speaker diarization mentioned above. In the training stage, we compute i-vectors with non-overlapping segments of each speaker in oracle labels. In the inference stage, we estimate i-vectors with visual-only speaker diarization results. The audio is also dereverbed with NARA-WPE and the whole process is illustrated on the left side of the middle row in Figure 1.

3.5. Embedding Combination

Given V-embedding \mathbf{E}_V , A-embedding \mathbf{E}_A and i-vectors \mathbf{I} , the goal of embedding combining block is to predict the probabilities of all speakers simultaneously. First, V-embeddings are repeated K times for concatenation due to the different frame shift between videos and audios. A-embeddings are repeated N times and i-vectors are repeated T times respectively. Then, the audio-visual speaker embeddings $\mathbf{E}_S = \{\mathbf{E}_{S_1}, \dots, \mathbf{E}_{S_n}, \dots, \mathbf{E}_{S_N}\} \in \mathbb{R}^{T \times D_S \times N}$ are obtained by splicing the V-embedding, A-embedding, and i-vectors. $\mathbf{E}_{S_n} \in \mathbb{R}^{T \times D_S}$ is the original frame-level audio-visual speaker embeddings where $D_S = D_V + D_A + D_I$. A shared speaker detection

(SD) component comprising 2-layer bidirectional LSTM with projection (BLSTMP) is used to further extract audio-visual features. Then, we adopt a 1-layer BLSTMP followed by FC layer to produce N outputs $\hat{\mathbf{S}}^{AV} = (\hat{S}_1^{AV}, \dots, \hat{S}_n^{AV}, \dots, \hat{S}_N^{AV}) \in (0, 1)^{T \times N}$ corresponding to the speech/non-speech probabilities for each of the N speakers respectively. The embeddings combining block is illustrated on the right side of the middle row in Figure 1.

3.6. Optimization

The audio-visual speaker diarization network is optimized in the following three steps: First, we copy the parameters of the pre-trained lipreading model to the visual network and train the V-VAD model with a learning rate of 10^{-4} . The loss function of V-VAD is written as follows:

$$J_{V_n} = \frac{1}{T} \sum_{t=1}^T BCE(S_{t,n}, \hat{S}_{t,n}^V) \quad (3)$$

where $BCE(\cdot, \cdot)$ is the binary cross entropy function between the labels and the outputs. Second, we freeze the visual network parameters and train the audio network and audio-visual decoding block on synchronized middle-field audio and video with a learning rate of 10^{-4} . The loss function is written as follows:

$$J_{AV} = \frac{1}{NT} \sum_{n=1}^N \sum_{t=1}^T BCE(S_{t,n}, \hat{S}_{t,n}^{AV}) \quad (4)$$

Finally, we unfreeze the visual network parameters and train the whole network jointly on synchronized middle-field audio and video with a learning rate of 10^{-5} . The loss function is written as follows:

$$J_{Joint} = \lambda \cdot \frac{1}{N} \sum_{n=1}^N J_{V_n} + J_{AV} \quad (5)$$

where $\lambda = 0.1$ in our experiments.

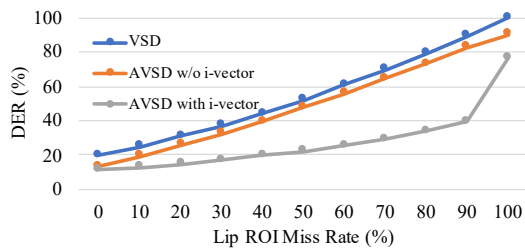
4. Experiments

4.1. Experimental Setup

We extracted 40-dimensional FBANKs (i.e., $F = 40$) with 25 ms frame length and 10 ms frame shift. The video is recorded 25 frames per second (40 ms frame shift) and the tracked lip ROI size is 96×96 ($W \times H$). The 256-dimensional (D_V) V-embeddings need to be repeated $4K$ times for each frame. The 100-dimensional (D_I) i-vector extractor was trained on CN-CELEB [32]. The dimension of A-embeddings (D_A) was set to 256. We used three conformer blocks that were equipped with 256 encoder dims, 4 attention heads, 32 conv kernel size and a 256-cell BLSTM in the V-VAD network. A-embeddings are extracted with 4 layers 2D CNN. In audio-visual decoding block, all of BLSTMP layers contained 896 cells. A threshold is used to obtain the decision of speech activity for each frame. We find the best threshold value for each model on MISP development set (DEV) and applied it on MISP evaluation set (EVAL). Given reference VAD that was generated by merging speaker turns in the reference diarization, we adopted the same post-processing strategy in [31]. The accuracy of speaker diarization system in this track is measured by diarization error rate (DER) [33] which is calculated as: the summed time of three different errors of false alarm (FA), missed detection (MISS) and speaker errors (SpkErr) divided by the total duration time.

Table 1: *Diarization results on MISP*

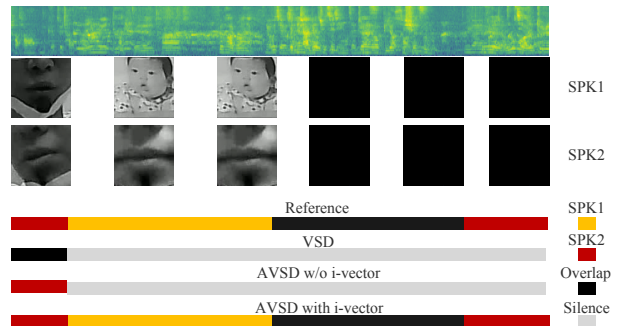
Set		DEV					EVAL				
Reference VAD		with			w/o		with			w/o	
Modality	System	FA	MISS	SpkErr	DER	DER	FA	MISS	SpkErr	DER	DER
Audio	VBx	0.00	25.79	7.56	33.35	40.21	0.00	26.25	7.44	33.69	40.82
	TS-VAD	4.30	11.94	11.67	27.91	-	4.27	12.77	11.92	28.95	-
Visual	VSD	4.91	6.74	2.94	14.59	20.63	4.20	6.60	2.28	13.07	19.64
Audio-Visual	¹ AVSD w/o i-vector	3.60	5.06	2.38	11.04	-	2.35	5.90	1.80	10.05	-
	² AVSD with i-vector	3.41	5.05	2.10	10.57	-	3.07	5.39	1.56	10.01	-
	³ + Joint training	3.32	4.67	2.14	10.12	11.68	2.96	4.97	1.56	9.49	10.99
Fusion	DOVER-Lap of 1, 2, 3	2.98	4.68	2.05	9.71	-	2.38	5.09	1.38	8.85	-

Figure 2: *DER comparison of different lip ROI missing rates without reference VAD on MISP EVAL set.*

4.2. Experimental Results

We presented the effects of the proposed audio-visual speaker diarization model with/without reference VAD in Table 1. For audio-only modality, we adopted the VBx [34] and TS-VAD trained on the middle-field audio of MISP. We reported the visual-only modality result ‘VSD’ with our visual network model of the first stage in Section 3.6. For ‘AVSD w/o i-vector’, we only spliced A-embeddings and V-embeddings as the inputs of embedding combining block. We showed ‘AVSD w/o i-vector’ results with both freezing and unfreezing (joint training) visual network parameters. The audio-only systems performed poorly on MISP mainly because the high overlap ratios in conversations and background TV talkers. The visual-only system still can not handle the lip missing problem in speech regions and lip wiggling problem in silent regions, which lead to higher MISS and FA respectively. With i-vector and joint training, the audio-visual systems can achieve better performance. For brevity, we default ‘AVSD with i-vector’ to be jointly trained hereinafter. Besides, DOVER-Lap [35] of audio-visual systems can bring further improvements. Moreover, benefited from multi-modal information, our audio-visual model is more robust than unimodal systems. Table 1 also shows the diarization results on MISP without reference VAD. We trained a VAD model [8] with MISP data for VBx. For visual-only and audio-visual systems, we only performed thresholding on VSD and AVSD (with i-vector) without any other post-processing strategies. Compared to unimodal results, AVSD is less affected by the absence of reference VAD where DER got worse by 1.5% absolutely while it was 7.13% and 6.57% in audio-only and visual-only systems, respectively.

We also explored the impact of missing lip ROI on different systems by directly removing lip ROI randomly with different degrees. For VSD, the missing lips were assigned silence. For AVSD, the silent lip ROI segments from training set were used to complement the missing parts of lip inputs. We estimate the

Figure 3: *An example including lip wiggling and lip missing problems in audio-visual recording.*

i-vectors with the VSD results corresponding to the lip missing rate. Figure 2 shows the DERs of VSD, AVSD without (w/o) i-vector and AVSD with i-vector on different lip ROI missing rates. The proposed AVSD method is much more robust than VSD method when the visual modal is missing. Meanwhile, the i-vector plays an important role in the AVSD model.

Figure 3 is an example including lip wiggling and lip missing problems. At the beginning, VSD regarded the speech segment as overlapping because SPK1’s lip was wiggling while AVSD (with/without i-vector) detected correctly by audio information. Then SPK1 spoke alone, both VSD and AVSD w/o i-vector failed to detect speech because SPK1’s lip was misdetected as the kid in the wall photo. At the end, AVSD with i-vector correctly detected overlapping speech where both SPK1 and SPK2’s lips were not detected.

5. Conclusions

In this paper, we proposed an end-to-end audio-visual neural-network-based speaker diarization method. Benefited from multi-modal information, our method has the abilities to handle overlap segments and distinguish between speech and non-speech. The i-vectors make audio-visual model much more robust when lip ROIs are missing. In the future, we will explore the multi-channel far-field audio and lower quality of lip ROIs from far-field video in end-to-end AVSD method.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.

7. References

- [1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 356–370, 2012.
- [2] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, vol. 1. IEEE, 2003, pp. I–I.
- [3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.
- [4] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.
- [5] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 413–417.
- [6] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.
- [7] Medennikov and *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," *Interspeech 2020*, Oct 2020. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2020-1602>
- [8] N. Ryant, K. Church, C. Cier, J. Du, S. Ganapathy, and M. Liberman, "Third dihard challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.
- [9] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019. [Online]. Available: <https://doi.org/10.1145/3343031.3350535>
- [10] G. Jocher and *et al.*, "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6222936>
- [11] P. Liu and Z. Wang, "Voice activity detection using visual information," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–609.
- [12] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Visual voice activity detection in the wild," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 967–977, 2016.
- [13] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.
- [14] A. Noulas, G. Englebienne, and B. J. Krose, "Multimodal speaker diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 79–93, 2011.
- [15] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.
- [16] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.
- [17] J. S. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," *arXiv preprint arXiv:1906.10042*, 2019.
- [18] R. Ahmad, S. Zubair, H. Alquhayz, and A. Ditta, "Multimodal speaker diarization using a pre-trained audio-visual synchronization model," *Sensors*, vol. 19, no. 23, p. 5163, 2019.
- [19] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.
- [20] J. Xia, A. Rao, Q. Huang, L. Xu, J. Wen, and D. Lin, "Online multi-modal person search in videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 174–190.
- [21] E. El Khoury, C. S enac, and P. Joly, "Audiovisual diarization of people in video content," *Multimedia tools and applications*, vol. 68, no. 3, pp. 747–775, 2014.
- [22] I. Kapsouras, A. Tefas, N. Nikolaidis, G. Peeters, L. Benaroya, and I. Pitas, "Multimodal speaker clustering in full length movies," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2223–2242, 2017.
- [23] 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, 2017.
- [24] E. Zhongcong Xu, Z. Song, C. Feng, M. Ye, and M. Z. Shou, "Ava-avd: Audio-visual speaker diarization in the wild," *arXiv e-prints*, pp. arXiv–2111, 2021.
- [25] H. Chen, H. Zhou, J. Du, C.-H. Lee, J. Chen, S. Watanabe, S. M. Siniscalchi, O. Scharenborg, D.-Y. Liu, B.-C. Yin, J. Pan, J.-Q. Gao, and C. Liu, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022.
- [26] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [29] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6319–6323.
- [30] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "Nara-wpe: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.
- [31] M. He, X. Lv, W. Zhou, J. Yin, X. Zhang, Y. Wang, S. Niu, Y. Cao, H. Lu, J. Du *et al.*, "The ustc-ximalaya system for the icassp 2022 multi-channel multi-party meeting transcription (m2met) challenge," *arXiv preprint arXiv:2202.04855*, 2022.
- [32] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.
- [33] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.
- [34] 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.
- [35] D. Raj, P. Garcia, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "DOVER-Lap: A method for combining overlap-aware diarization outputs," *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021.