Interspeech 2022
18-22 September 2022, Incheon, Korea

MSDWILD: MULTI-MODAL SPEAKER DIARIZATION DATASET IN THE WILD

†*Tao Liu*[1], †*Shuai Fan*[2], *Xu Xiang*[2], *Hongbo Song*[2], *Shaoxiong Lin*[1], *Jiaqi Sun*[1], *Tianyuan Han*[1], *Siyuan Chen*[1], *Binwei Yao*[1], *Sen Liu*[1], *Yifei Wu*[1], ‡ *Yanmin Qian*[1] , ‡ *Kai Yu*[1]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, X-LANCE Lab, Shanghai Jiao Tong University
[2]AISpeech Ltd, Suzhou China

{liutaw, Johnson-Lin, jotaro, hty19980927, chensiyuan925, yaobinwei, sen.liu, yifei.wu, yanminqian, kai.yu}@sjtu.edu.cn,
{shuai.fan, xu.xiang, hongbo.song}@aispeech.com

Abstract

Speaker diarization in real-world acoustic environments is a challenging task of increasing interest from both academia and industry. Although it has been widely accepted that incorporating visual information benefits audio processing tasks such as speech recognition, there is currently no fully released dataset that can be used for benchmarking multi-modal speaker diarization performance in real-world environments. In this paper, we release *MSDWild**, a benchmark dataset for multi-modal speaker diarization in the wild. The dataset is collected from public videos, covering rich real-world scenarios and languages. All video clips are naturally shot videos without over-editing such as lens switching. Audio and video are both released. In particular, *MSDWild* has a large portion of the naturally overlapped speech, forming an excellent testbed for cocktail-party problem research. Furthermore, we also conduct baseline experiments on the dataset using audio-only, visual-only, and audio-visual speaker diarization.

Index Terms: speaker diarization, multi-modality, audio-visual

1. Introduction

Speaker diarization [1, 2] identifies the talkers and their talking duration, solving the problem of 'who spoke when.' Inspired by multi-modal complementarity [3] and its success in several audio-visual tasks [4, 5], multi-modal speaker diarization [6, 7, 8] takes advantage of both audio and visual modalities, exploring multi-modal fusion and further improving the performance.

However, existing multi-modal speaker diarization datasets [7, 9, 10] are constrained in meetings, TV programs, or movie sources, which are recorded cooperatively and contain a mismatch to the in-the-wild scenario. In the scenario, people are talking spontaneously with frequent switching and sudden interrupting, leading to a naturally overlapped speech during the talking. Besides, distances from the talking face to the camera vary, resulting in various face resolutions and recording distances. In addition, not all faces are front, and plenty of them are side faces. Lastly, natural interference also exists, including various noises and room reverberations.

Figure 1: *Several examples from AMI [7], AVA-AVD [12], and our MSDWild methods. Compared with existing datasets from constrained sources (e.g., AMI is from meetings, AVA-AVD is from movies), our dataset contains daily casual conversation in the wild, with at least two speakers talking in turn. The bottom of our examples is the speaker diarization timeline; different colors represent different speakers.*

In this paper, our objective is to collect a multi-modal speaker diarization dataset in those in-the-wild settings. To achieve this, we build a collecting pipeline. Compared with collecting pipeline [10], we have two major differences. First, we use the keyword 'VLog' to gather informal talks rather than 'panel debate' or 'discussion,' which are often used in conjunction with formal talks. Second, to filter out videos with over two speakers talking in turn, we do it manually without using any pre-trained audio-visual algorithms (e.g., SyncNet [11]). We think those algorithms may lead to dataset bias. Our dataset examples, compared with AMI [7] and AVA-AVD [12], are shown in Figure 1.

Active speaker detection, which judges whether the audio matches the face, is a significant sub-task for audio-visual speaker diarization methods [13, 14, 10, 8, 12]. Most of those methods train on a separate audio-visual dataset (e.g., VoxCeleb2 [4]), not directly on the audio-visual speaker diarization dataset, which leads to a mismatch between the sub-task and the final inference performance. Our dataset also provides videos with cropped faces, labeled with speaking or not speaking, for training audio-visual sub-task.

† These authors contributed equally to this work.
‡ Yanmin Qian and Kai Yu are the corresponding authors.
* The dataset is available at https://x-lance.github.io/MSDWILD.

10.21437/Interspeech.2022-10466

Table 1: *Comparison with existing audio-visual speaker diarization datasets.* ***overlapped****: The overlapped speech rate per video.* ***#speakers****: The min/average/max speaker number per video.* ***#SC****: The average speaker changes times per minute.* ***#noise****: Whether the videos contain the noise.* ***#continuous****: Whether video scenes are continuous without fast camera switching. Speakers in our dataset frequently talk in turn, and the overlapped ratio is highest compared with other datasets.*

dataset	source	#videos	duration	speech%	overlapped	#speakers	#SC	noise	continuous	language
AMI [10]	Meeting	170	100h	80.91%	13.57%	3 / 4 / 5	7.8	✗	✓	En
VoxConverse [10]	TV show	448	63h50m	90.7%	3.6%	1 / 5.6 / 21	3.28	✗	✓	En
AVA-AVD [12]	Movie	351	29h15m	45.95%	4.4%	2 / 7.7 / 24	9.6	✓	✗	Multi
MSDWild (Ours)	Daily Conversation	3143	80h3m	91.29%	14.01%	2 / 2.7 / 10	11.8	✓	✓	Multi

Our contributions are three-fold. First, we release a multi-modal speaker diarization dataset in the wild, *MSDWild*. This dataset contains over 3000 video clips with 80 hours. Besides, these datasets can be used for training and testing simultaneously. Second, extensive analyses are conducted, including detailed dataset metrics and comparisons with existing datasets. Third, we also conduct several audio-only, visual-only, and audio-visual baseline methods on our dataset. For audio-visual methods, we also investigate fusion strategies.

We hope that *MSDWild* can provide a real in-the-wild testbed for the speaker diarization community. Besides, *MSDWild* is also suitable for exploring the ability of multi-modal audio-visual fusion for speaker diarization, better solving the problem of 'who spoke when.'

2. Related works

Multi-modal speaker diarization dataset. There exist some audio-visual datasets [7, 9, 12, 10] related with our proposed dataset. The AMI meeting corpus [7] is an audio-visual meeting recording dataset. The total length of the data set is 100 hours, and all audios are recorded in English by eight microphone arrays. The AMI meeting corpus focuses on the meeting scenario. VoxConverse [10] proposes a multi-modal speaker diarization dataset. The total length of VoxConverse is around 64 hours. VoxConverse first downloads videos from YouTube by keywords(e.g., panel debate, discussion) and then uses an automatic creation pipeline to filter required videos. We can not analyze it further because VoxConverse does not release its visual part (as of June 2022). AVA movie dataset [9] is an audio-visual dataset, originally intent for active speaker detection. Based on this dataset, AVA-AVD [12] forms a new speaker diarization dataset by re-annotating the speaker diarization labels. Those datasets are collected from constrained sources. Our dataset focus on daily casual conversation, forming a complete in-the-wild dataset. The comparsion is shown in Table 1.

Multi-modal speaker diarization. Multi-modal speaker diarization tries to utilize visual information to improve audio-only performance in the wild. Several methods [13, 14, 10, 8, 12] are investigating how to make fully use of visual modality. Multiple faces may exist simultaneously in real scenarios, leading to a permutation between one audio and multiple faces. Multi-modal speaker diarization requires some pre-processing methods, such as face detection and face tracking, to crop facial motions from videos, solving the problem 'who.' Then audio-visual relation or synchronization task is applied to the videos with cropped faces, solving the problem 'when.' This task tries to find the relation between audio and facial motions, mainly lip motion. The theoretical basis derives from viseme-phoneme mapping: a correlation exists between the minimal pronouncing unit, phoneme, and the corresponding lip motion unit, viseme. Multiple algorithms [11, 8, 15] are proposed to find the relation. There are two typical multi-modal fusion architectures: two-stream and fused. Two-stream methods [11, 8] jointly train audio and visual by metric learning [16] (e.g., contrastive loss and triplet loss [17]). TalkNet [15], a classic fused method, learns the audio-visual relation by the neural network. We will test those two architectures on our dataset.

3. Data collection

Stage 1. Searching and downloading videos. To search videos, we use 'VLog' keyword to search and download movies from YouTube. Before downloading those videos, we further check whether those videos have at least one scene with two speakers talking. Videos with all visible talking faces are preferred. To improve the language diversity, we change the website location or use Google Translate to translate those English keywords into different languages such as Chinese, Thai, Korean, Japanese, German, Portuguese, and Arabic.

Stage 2. Scene detecting. Post-editing, such as multi-camera switching, video clip stitching, and abnormal speed playback, is often combined with downloaded videos, resulting in discontinuous video scenes. So we use PySceneDetect [18] to split the video into separate clips. Each video clip is under the same scene without a quick scene change. This stage aims to avoid face tracking method failure and match the real scenario.

Stage 3. Manual filtering. After scene detecting, not every scene in videos satisfies our requirements. There may still be videos with no talking person, only one talking person, or post-editing. We manually remove those videos and keep the remaining videos satisfying the criteria that at least two speakers talk. It is noted that we do not use any pre-trained methods to assist filtering here.

Stage 4. Manual labeling using VIA Video Annotator. VIA Video Annotator [19] is a manual annotation software for videos, which has a video player and a timeline. We mark different timelines for different speakers and add temporal segments for each speech duration. The opening and closing of the lips mark the beginning and end of a speech segment. Only speech is labeled while other human sounds such as laughing and singing are ignored. In addition, single words (e.g., 'yes' or 'no') and off-screen speeches are also be labeled.

Stage 5. Double checking. To reduce faults and improve the quality of labels, one annotator checks another annotator's diarization. Verification criteria are followed by [10]. The boundary differences between the labeled and the ground-truth segments must be shorter than 0.1 seconds. With a pause time over 0.25 seconds, segments should be split and considered as separated ones.

Stage 6. Talking faces labeling. Apart from labeling speaker diarization, we also crop out talking faces from raw videos. Those videos can be used for audio-visual relation training. We first use a detecting method, S^3FD [20], and IoU

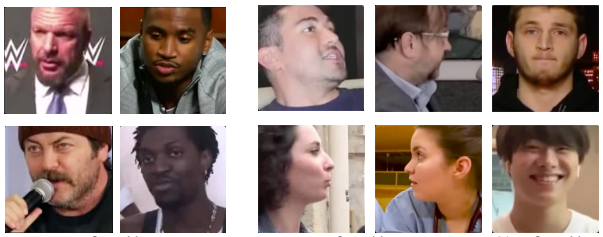

Figure 2: *Comparison with VoxCeleb2 [4], a celebrity interview dataset, talking face in our dataset has more casual head gestures and has an extra not-speaking status. Those talking faces are used to train an active speaker detection task, a sub-task for multi-modal speaker diarization.*

tracking to get all talking face videos. Then the annotators label videos into two groups: speaking and not-speaking. We only filter videos over two seconds for efficiency because the long period can generate the short. Finally, we get around 62 hours of speaking videos and 40 hours of not speaking videos. Those talking face videos are pretty similar to videos in VoxCeleb2 [4] but have two differences. First, those cropped faces are extracted from in-the-wild videos mentioned above. Those videos contain more casual head gestures. Second, we have not-speaking segments: lips are not moving. VoxCeleb2 only has speaking segments. Examples are shown in Figure 2.

4. Dataset description

Our dataset, *MSDWild*, contains 3143 video clips with 84 labeled hours. Daily casual conversation occupies the majority of our dataset. Compared with the formal conversation: news report or debate, the daily casual conversation has three features: frequently talking in turn, various head gestures, and various background noises or room reverberations.

In multi-modal speaker diarization, we find that the algorithm performances are extremely affected by the speaker number. So we separate videos with speaker numbers over than four to form a many-talker set. The rest, with speaker numbers from two to four, forms a few-talker set. Furthermore, we randomly divide the few-talker set into two sets: a training set and a testing set. Due to its limited size, the many-talker set is only used for testing. Therefore, our dataset forms three sets: few-talker training set, few-talker testing set, and many-talker testing set.

The number of video clips in those three sets is 2476, 490, and 177. The corresponding labeled duration is 69.09, 10.58, and 4.51 hours. However, the many-taker set's average utterance number and the overlapped speech ratio are significantly greater than the few-talker set. The many-talker set has much more speech alternations and overlapped speeches. The detailed dataset metrics are listed in Table 2.

Table 2: MSDWild *dataset metrics.* ***#speakers****: The min/average/max speaker number per video.* ***#videos****: Video numbers.* ***labeled (h)****: The total labeled duration in hours.* ***#utters****: The average utterance number per video.* ***overlap (%)****: The average overlapped speech per video.* ***speech (%)****: The average speech occupation. Our dataset is divided into three parts: few-talker train, few-talker val and many-talker val set.*

	Few-talker$_{Train}$	**Few-talker** $_{Val}$	**Many-talker**$_{Val}$
#speakers	2 / 2.54 / 4	2 / 2.61 / 4	5 / 5.86 / 10
#videos	2476	490	177
labeled (h)	69.09	10.58	4.51
#utters	36.1	30.28	43.32
overlap (%)	13.42	15.74	19.52
speech (%)	91.48	91.71	87.27

5. Experiments

5.1. Evaluation metrics

We report two metrics: DER(Diarization Error Rate) and JER(Jaccard Error Rate). DER is the summary of missed speech (MS) time, false alarm (FA) time, and speaker error (SE) time to the reference time. We also calculate errors in overlapped speech part (OVL). JER, initially proposed by DIHARD [21], is the average of each speaker's MS and FA rate. Compared with DER, JER is more strict when some speakers dominate the conversation. A 0.25-second forgiving collar is used for all metrics, and overlapped speeches are taken into account.

5.2. Audio-only method

We run Pyannote [22], a publicly available open-source toolkit[1] for speaker diarization, on our dataset. Pyannote uses a standard audio-only pipeline that includes an SincNet-based [23] model for voice activity detection, ECAPA-TDNN [24] for embedding extraction, and agglomerative clustering for speaker clustering.

5.3. Visual-based pipeline

Visual-only and audio-visual share the same pipeline [13, 14, 8]: face detection, face tracking, and active speaker detection. We adopt S^3FD [20] as the face detection method. In the face tracking stage, we use IoU tracking for faces in adjacent video frames and use Arcface [25] to cluster different tracks when they belong to the same speaker. Because those videos exist audio not belonging to the talking faces, we need active speaker detection methods to capture the correspondence between audio and lip motion. An example is shown in Figure 3. We investigate three methods: visual-only, two-stream audio-visual, and fused audio-visual. Before illustrating those methods in detail, we first describe the experiment setup.

5.3.1. Experiment setup

Visual representation. The inputs of our visual encoder are videos with cropped faces. First, we transform the RGB image to gray-scale in order to save the computation, and the frame rate of the video is converted to 25 Hz. So the input formulation is (T_v, W_v, H_v), representing video frame time, width, and height. Both W_v and H_v use 112 here. Followed by [15], we also use 3D Conv and ResNet-18 to encode each frame into a 512-dimensional embedding. Finally, TCN [26] is followed to capture the inter-frame relation.

Audio representation. The sample rate of our audios is 16k Hz. We first convert audios into MFCC (Mel Frequency Cepstral Coefficient) features, with the window length: 25 ms, the window step: 10 ms, and the number of cepstrum: 13. Then SENet [27] is followed to encode audio features.

Data augmentation. The visual data augmentation uses random horizontal flip, random cropping, random rotation, and random sampling low resolution: 32×32, 64×64, or 96×96.

[1] https://github.com/pyannote/pyannote-audio

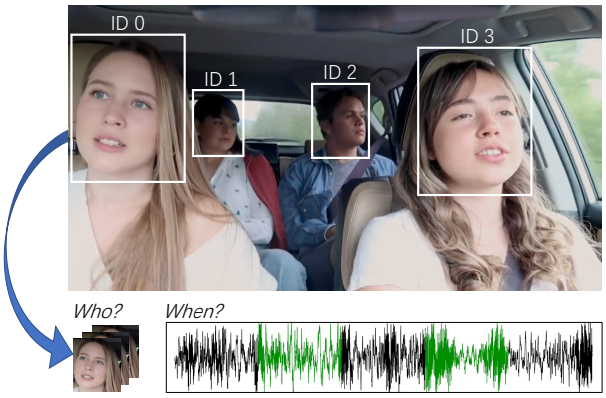

Figure 3: ***Audio-visual Speaker Diarization Pipeline.*** *Best view in color. First, the face detection and face clustering stage are to identify the speaker, solving the problem 'who.' Then, audio-visual active speaker detection is to figure out the speaking period of each talking face, solving the problem 'when.' Waveform in green color represents the speaking period of ID 0 in this example.*

The audio data uses MUSAN [28] and room reverberation [29] for data augmentation.

Training and sampling strategy. We use Adam [30] to optimize our network for training with a learning rate of 0.0001 and weight decay of 0.95. The training epoch is 30. For sampling strategy, same as TalkSet [15] format, we use a similar format to generate sync pairs and asynchronous pairs. However, we have two differences. One is that audio segmentation is from our dataset, not from VoxCeleb2 [4]. The other is adding additional not-speaking videos, lips not moving, from our dataset. All methods are trained on those training samples except for the visual-only method, which lacks audio modality and only uses synced pairs for training.

5.3.2. Implementation of visual-based pipeline

Visual-only method. We solely employ visual representation to train a binary classification network for each video frame in the visual-only method: zero for labeling not-speaking and one for labeling speaking. The binary cross-entropy loss is used to optimize this network.

Two-stream audio-visual method. Two-stream [11, 8] uses a parallel feature extraction network for individual modalities. Two-stream processes segment-level input and outputs an embedding for the whole segment. Longer input duration will lead to a coarse boundary, and shorter input causes lower accuracy. So there exists a trade-off in choosing the segment duration. The segment duration, used in our experiment, is 0.4 and 1 second (for T_a and T_v). Contrastive loss is utilized to optimize this network during training, and L2 distance is employed as the similarity metric.

Fused audio-visual method. The fused architecture [15] implicitly models the relation between multi-modalities by training a weighted network. We use concatenation to fuse multi-modalities here. The input length of the fused model differs from the two-stream model in that it is trained and assessed at the frame level. The fused model classifies the talking state for each video frame, but the segment-level model only produces one outcome. As a result, the fused technique has a better time resolution and can better model inter-frame relations. Same as the visual-only method, the cross-entropy loss is also utilized here to optimize this network. During inference, video frames with a probability over 0.5 are regarded as speaking while others are not speaking.

Table 3: *The baseline results of audio-only, visual-only, and audio-visual methods on the* ***few-talker*** *val set.*

Method	MS	FA	SE	OVL	DER	JER
Audio-only [22]	5.5	3.16	13.3	5.05	21.96	61.0
Visual-only	9.13	12.47	1.71	3.12	23.32	41.22
Audio-visual(Two-stream) [11, 8]	8.49	14.3	1.8	3.66	24.59	44.71
Audio-visual(Fused) [15]	7.27	4.0	0.93	3.26	12.2	35.01

Table 4: *The baseline results of audio-only, visual-only, and audio-visual methods on the* ***many-talker*** *val set.*

Method	MS	FA	SE	OVL	DER	JER
Audio-only [22]	12.64	5.54	24.97	12.34	43.15	84.28
Visual-only	11.72	34.54	7.45	8.09	53.71	62.71
Audio-visual(Two-stream) [11, 8]	14.6	31.08	6.91	7.9	52.6	63.7
Audio-visual(Fused) [15]	14.2	7.45	4.2	6.59	25.86	54.79

5.4. Result and analysis

Table 3 and Table 4 show the final result. First, the overall DER and JER show that our dataset is challenging, especially in the many-talker condition. Second, the two-stream audio-visual methods perform even worse compared with the audio-only method. The fused audio-visual method improves 9.76% and 17.29% absolutely in the few-talker and many-talker val set, respectively. Besides, audio-visual methods can improve speaker diarization performance in speaker error (SE) rate and error rate in overlapped speech part (OVL), especially in many-talker condition. However, they perform poorly in missed speech (MS) and false alarm (FA). Those show potential in audio-visual methods, and more efficient methods remain to be explored. Third, compared with audio-only methods, methods with visual modality, including visual-only and audio-visual ones, improve greatly in JER, which shows that visual modality can better solve the speaker diarization in the wild. This is because there is a high variance in the speaker time in the wild, and the JER metric can capture the variance.

6. Conclusion

In this paper, we propose *MSDWild*: a novel multi-modal speaker diarization dataset in the wild. The dataset contains spontaneously daily conversations on 'unconstrained' conditions. We also test several baseline methods for speaker diarization. However, there exist multi-modal methods to be explored, and experiments reveal that there is still potential for improvement, particularly in the multi-talker situation.

7. ACKNOWLEDGEMENTS

This work was supported by State Key Laboratory of Media Convergence Production Technology and Systems Project (No. SKLMCPTS2020003), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), National Natural Science Foundation of China (Grant No. 92048205), and Suzhou Science and Technology Planning Project (No. ZXT2020003).

8. References

[1] 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," *Computer Speech and Language*, vol. 72, p. 101317, 2022.

[2] H. Aronowitz, W. Zhu, M. Suzuki, G. Kurata, and R. Hoory, "New advances in speaker diarization," in *International Speech Communication Association (Interspeech)*, 2020.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *International Speech Communication Association (Interspeech)*, 2018.

[5] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 15 490–15 500.

[6] W. Kang, B. C. Roy, and W. Chow, "Multimodal speaker diarization of real-world meetings using d-vectors with spatial features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6509–6513.

[7] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th international conference on methods and techniques in behavioral research*, vol. 88. Citeseer, 2005, p. 100.

[8] Y. Ding, Y. Xu, S.-X. Zhang, Y. Cong, and L. Wang, "Self-supervised learning for audio-visual speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4367–4371.

[9] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4492–4496.

[10] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," in *International Speech Communication Association (Interspeech)*, 2020.

[11] J. S. Chung and A. Zisserman, "Lip reading in profile," 2017.

[12] 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.

[13] 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.

[14] R. Ahmad, S. Zubair, and H. Alquhayz, "Speech enhancement for multimodal speaker diarization system," *IEEE Access*, vol. 8, pp. 126 671–126 680, 2020.

[15] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935.

[16] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[17] W. Ge, "Deep metric learning with hierarchical triplet loss," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 269–285.

[18] B. Castellano, "Pyscenedetect: Video scene cut detection and analysis tool." [Online]. Available: https://github.com/Breakthrough/PySceneDetect

[19] 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

[20] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 192–201.

[21] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *International Speech Communication Association (Interspeech)*, 2019, pp. 978–982.

[22] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020.

[23] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[24] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "Ecapa-tdnn embeddings for speaker diarization," 2021, arXiv:2104.01466.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[26] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[29] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent.*, 2014, pp. 1–15.