Interspeech 2024
1-5 September 2024, Kos, Greece

AFL-Net: Integrating Audio, Facial, and Lip Modalities with a Two-step Cross-attention for Robust Speaker Diarization in the Wild

YongKang Yin[1,†], *Xu Li*[2,*], *Ying Shan*[2], *YueXian Zou*[1,*]

[1]ADSPLAB, School of ECE, Peking University, China
[2]ARC Lab, Tencent PCG

yinyongkang@stu.pku.edu.cn, {nelsonxli, yingsshan}@tencent.com, zouyx@pku.edu.cn

Abstract

Speaker diarization in real-world videos presents significant challenges due to varying acoustic conditions, diverse scenes, the presence of off-screen speakers, etc. This paper builds upon a previous study (AVR-Net) and introduces a novel multi-modal speaker diarization system, AFL-Net. The proposed AFL-Net incorporates dynamic lip movement as an additional modality to enhance the identity distinction. Besides, unlike AVR-Net which extracts high-level representations from each modality independently, AFL-Net employs a two-step cross-attention mechanism to sufficiently fuse different modalities, resulting in more comprehensive information to enhance the performance. Moreover, we also incorporated a masking strategy during training, where the face and lip modalities are randomly obscured. This strategy enhances the impact of the audio modality on the system outputs. Experimental results demonstrate that AFL-Net outperforms state-of-the-art baselines, such as the AVR-Net and DyViSE.

Index Terms: multi-modal speaker diarization, cross-attention, lip movement

1. Introduction

Speaker diarization is a cutting-edge technology that identifies "who spoke when" in audio or video recordings [1]. This task involves segmenting the input into homogeneous regions based on speaker identity, playing a crucial role in applications like meeting analysis, broadcast news indexing, transcription services and spoken language understanding [2].

Traditional speaker diarization primarily utilized audio features like spectral characteristics and prosodic patterns for speaker identification and segmentation. Early works [3] followed a multi-stage approach, encompassing speech enhancement, voice activity detection (VAD), speaker feature extraction, similarity scoring, and clustering. Subsequent studies [4,5] employed deep learning networks for end-to-end speaker diarization, with later research addressing associated issues, such as uncertain speaker count [6]. In recent years, multi-modal speaker diarization has integrated audio and visual features to improve performance. These studies employed various methods, such as CNN enhancement [7], self-supervised learning [8], etc. Some works incorporated additional features like speaker i-vectors [9] and localization features [10, 11] to enhance system performance. Other studies focused on specific scenarios, such as real-world conferences and in-the-wild settings [12, 13]. Despite achieving reasonable performance, these methods encountered limitations in complex and noisy environments, resulting in suboptimal outcomes.

In real-world scenarios like movies, challenging scenes with off-screen speakers hinder the direct use of audio-visual speaker diarization (AVSD) systems. The AVR-Net [13] was proposed by using a modality embedding that reflects face visibility, making AVSD adaptable to situations where the face is not visible. While the AVR-Net demonstrated satisfactory performance in real-world scenarios, this study suggests that there is room for further enhancements to improve its overall performance. 1) Besides the face and audio modalities that are considered by the AVR-Net to compute the speaker similarity between segments, lip movement is another key modality but not taken into consideration by the AVR-Net. The dynamic lip movement could reflect a person's pronunciation styles, which are complementary to face and audio modalities and can enhance the discrimination of identities [9]. 2) In the AVR-Net, different modalities may not be sufficiently fused according to the network architecture. The face and audio modalities are not orthogonal and should be conditioned on each other to generate high-level representations, rather than generating the representations independently and simply concatenating these representations of each modality for segment scoring as in [13].

To address the issues above, this work proposes a novel multi-modal speaker diarization network, named AFL-Net, that extends the AVR-Net in the following aspects: 1) Besides the face and audio modalities, we leverage the dynamic lip movement as an additional modality to improve the discrimination of identities; 2) We design a two-step cross-attention fusion module to generate high-level representations of each modality while conditioning on other modalities; 3) Since the visual modalities have the potential to be partially absent (e.g. partially obscured face or incomplete lip movement on side face), we propose a masking strategy to randomly mask out the face and lip movement modalities during training, which increases the impact of the audio modality on system outputs.

A concurrent work, DyViSE [11], bears similarity to ours. However, the proposed AFL-Net distinguishes itself in several ways: 1) The AFL-Net employs a two-step cross-attention method to fuse three modalities, whereas [11] only applies attention between the lip and audio modalities; 2) [11] depends on three pre-trained models for embedding computation and uses cosine similarity for segment scoring. In contrast, our method directly trains a scoring network to output similarity scores between segments; 3) The AFL-Net adopts an additional masking strategy during training. The dedicated architecture of AFL-Net enables it to outperform DyViSE, as will be demonstrated in Section 4.1.

The contributions of this work are as follows: 1) Leveraging the dynamic lip movement as an additional modality to enhance

[†]This work was done when Yongkang Yin was an intern at ARC Lab, Tencent PCG.

* Corresponding authors.

10.21437/Interspeech.2024-764

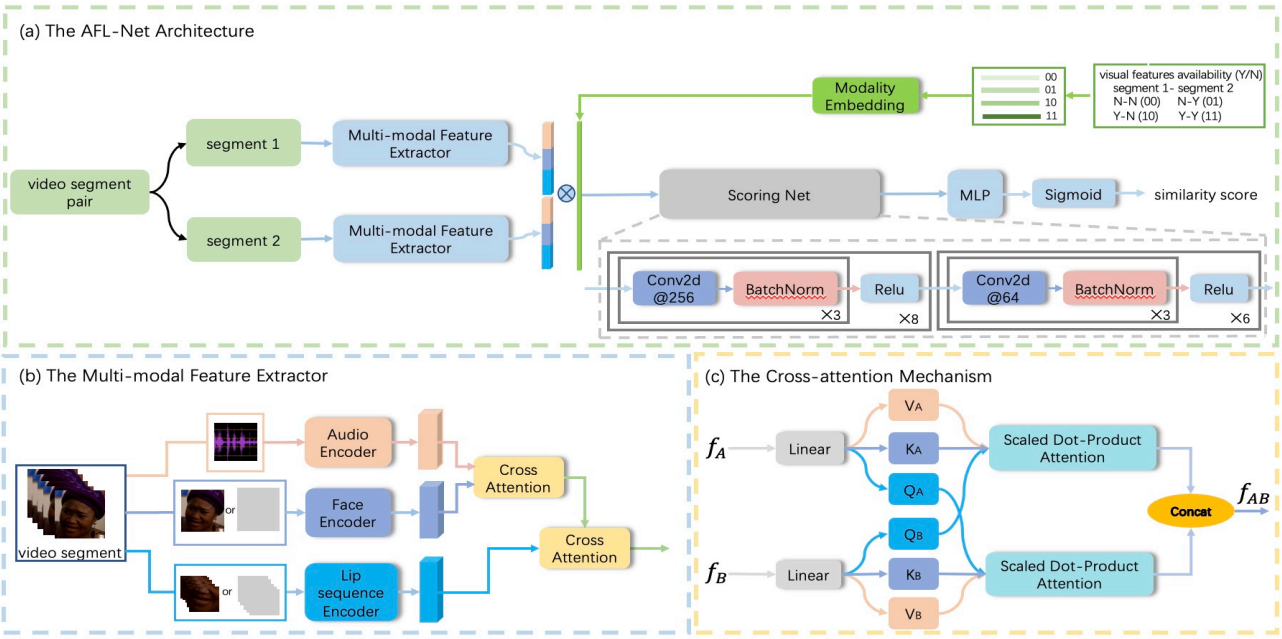

Figure 1: *The proposed system architecture. (a) represents the AFL-Net architecture, where the MLP indicates several linear layers. (b) illustrates the multi-modal feature extractor and (c) demonstrates the cross-attention mechanism.*

the system's identity discrimination; 2) Introducing a two-step cross-attention module to sufficiently fuse different modalities for similarity scoring; 3) Proposing a masking strategy to randomly mask out the visual modalities during training, which leads the system to make decisions more on the audio information and experimentally achieve better speaker diarization performance.

The rest of this paper is organized as follows: Section 2 illustrates the proposed system. Experimental setup and experiment results are demonstrated in Section 3 and 4, respectively. We conclude this work in Section 5.

2. The Proposed Method

The proposed AFL-Net adopts the general architecture of the AVR-Net [13], which trains the network to assign an identity similarity score to each pair of two video segments. During inference, a testing video is initially divided into shorter segments, and the network predicts an identity similarity score between each pair of segments. Subsequently, a clustering algorithm, such as agglomerative hierarchical clustering (AHC), is applied to group segments with high scores, indicating that they likely belong to the same person. This work modifies the scoring network of [13] while adopting the same AHC algorithm for clustering.

The proposed AFL-Net is demonstrated in Fig. 1. In Fig. 1(a), given a pair of two video segments, a multi-modal feature extractor is applied to extract high-level representations for each segment independently, then the two representations are concatenated together for scoring. Since there are instances where visual features are absent, we utilize 4 learnable modality embeddings [13] to indicate whether the visual features are available or not for each segment. These embeddings are randomly initialized and optimized during training. The concatenated representations are multiplied with the selected modality embedding, then fed into the scoring network to generate the similarity score for this pair of segments.

2.1. Multi-modal feature extractor

The multi-modal feature extractor is illustrated in Fig. 1(b). Besides the face and audio modalities adopted in [13], this work additionally leverages the dynamic lip movement as the third modality to be fused in the model because lip movement [14] could reflect a person's pronunciation styles, which are complementary to face and audio modalities and can enhance the discrimination of identities [9]. We utilize three open-sourced pretrained encoders to extract audio [15], face [16], and lip movement [17] features, respectively. All of them have a ResNet-like [18] network architecture and have been trained on large-scale datasets, yielding impressive recognition performance.

Furthermore, to effectively integrate information from different modalities, we employ a two-step cross-attention mechanism. This mechanism allows for the generation of high-level representations for each modality while considering queries from other modalities. The structure of a single cross-attention is shown in Fig. 1(c). Suppose that we have two modality features f_A and f_B, the key, query and value vectors of each modality are first extracted through a linear layer, denoted by K_A, Q_A, V_A, K_B, Q_B and V_B, respectively. The cross-attended fusion of these two modalities is derived by Eq. 1:

$$f_{AB} = \text{softmax}(\frac{Q_B K_A^T}{\sqrt{d_A}})V_A \oplus \text{softmax}(\frac{Q_A K_B^T}{\sqrt{d_B}})V_B \quad (1)$$

where f_{AB} represents the fused feature of the modality A and B, d_A and d_B are the dimensions of the corresponding key vectors, and $\oplus$ denotes the concatenation operation.

Please note that there are multiple fusion implementations for three modalities, including a one-step three-modality fusion or a two-step two-modality fusion. In this study, four fusion implementations were explored, and experimental results

indicate that the optimal implementation involves initially fusing the audio and face modalities, followed by the inclusion of the lip movement modality in the second step, as illustrated in Fig. 1(b). Further details will be discussed in Section 4.3. The implementation is illustrated by Eq. 2 and 3:

$$f_{AF} = \text{softmax}(\frac{Q_F K_A^T}{\sqrt{d_A}})V_A \oplus \text{softmax}(\frac{Q_A K_F^T}{\sqrt{d_F}})V_F \quad (2)$$

$$f_{AFL} = \text{softmax}(\frac{Q_L K_{AF}^T}{\sqrt{d_{AF}}})V_{AF} \oplus \text{softmax}(\frac{Q_{AF} K_L^T}{\sqrt{d_L}})V_L \quad (3)$$

Finally, f_{AFL} is the fused feature of the three modalities.

2.2. Masking strategy

Given that visual information in videos, such as face and lip movement modalities, can be partially absent (e.g., due to obscured faces or incomplete lip movements on the side face), systems that heavily rely on visual inputs may encounter a significant performance decline in such challenging scenarios. To mitigate this issue, we propose a masking strategy, where the face and lip movement modalities are randomly masked out during training. This strategy aims to amplify the impact of the audio modality on the system outputs, compensating for the potential absence of visual cues. In our approach, we randomly mask out the visible face and lip movement modalities at a rate of 0.3 during training. This process also includes updating the corresponding modality embedding. It is important to note that this masking strategy is only applied during training, and the testing data remains unaltered for inference.

2.3. Agglomerative hierarchical clustering

Following [13], we use agglomerative hierarchical clustering (AHC) [19] for speaker clustering. AHC is a bottom-up method that starts with each data point as an individual cluster and iteratively merges the closest clusters until a predefined similarity threshold is reached. In our approach, AFL-Net predicts identity similarity scores between video segments. AHC then groups these segments into clusters, each representing a specific person, without needing to pre-specify the number of speakers.

3. Experimental Setup

3.1. Datasets

The experiments involve three datasets: AVA-AVD [13], VoxCeleb1 [20], and VoxCeleb2 [21]. The AVA-AVD dataset contains 351 video clips (243 for training, 54 for validation, and 54 for testing) from 117 movies, each 5 minutes long with up to 24 speakers. VoxCeleb1 and VoxCeleb2 include over 1 million video segments from more than 7,000 celebrities. Due to its diverse scenes and complex acoustic conditions, AVA-AVD is challenging but relatively small, so it is used for both training and evaluation. To validate the proposed model on a larger scale, VoxCeleb1 and VoxCeleb2 are used as additional training data.

3.2. Model and training configurations

Following [13], we partition the 16-bit 16kHz mono-channel active speaker audio segments, obtained after a speech enhancement process and voice activity detection (VAD) [22, 23], into segments of 0.5 seconds each. Concurrently, we extract facial images and lip sequence images within the same audio time frame, leveraging existing open-source tools [24–26]. From each time slot, we randomly select one face image and ten lip images. In instances where visual features are absent, a zeroing operation is performed as a placeholder, as shown in Fig. 1(b).

The loss function for training the AFL-Net is defined as:

$$\mathcal{L}_{\text{MSE}} = \sum_{i=1}^{N}\sum_{j=1}^{N}(s_{i,j} - 1(y_i == y_j))^2 \quad (4)$$

where N represents the batch size, y_i and y_j represent the identity labels, and $s_{i,j}$ is the predicted score by the AFL-Net. The AFL-Net is trained to predict 1 if two segments are from the same person and 0 vice versa. The Adam optimizer is utilized to update model parameters, with an initial learning rate of 5×10^{-4}. After 6k iterations of training, the model is evaluated on the validation set of the AVA-AVD dataset every 500 iterations. The model parameters that achieve the best performance on the validation set are selected for inference. For the clustering backend, we search the threshold within the range from 0.1 to 0.3 with a step size of 0.01. The threshold that yields the minimum DER in validation is recorded for the testing period.

3.3. Evaluation metric

There are four commonly used evaluation metrics in speaker diarization [27]: missing rate (MR), false alarm rate (FAR), speaker error rate (SpkErr), and diarization error rate (DER). DER is a comprehensive measure that combines MR, FAR, and SpkErr, calculated using Eq. 5.

$$DER = \frac{T_{MS} + T_{FA} + T_{SPKE}}{T_{total}} \quad (5)$$

where T_{total} represents all the testing frames, T_{MS}, T_{FA} and T_{SPKE} represent the misclassified frames that cause the miss, the false alarm and the speaker error, respectively.

The AVR-Net [13] and the DyViSE [11] are selected as baseline models for performance comparison. Both the AVR-Net and the AFL-Net utilize the same VAD tool, resulting in identical performance in terms of MR (2.55%) and FAR (0). Hence, we only demonstrate the SpkErr and the DER results for comparing the AFL-Net and the AVR-Net in Table 1. The DyViSE [11] utilizes a stronger VAD tool [28], resulting in a slightly lower MR of 1.98%. Furthermore, the audio encoder in DyViSE utilizes a substantial 310M-parameter WavLM model, whereas our approach only employs a modest 9M parameters. To ensure a fair comparison, we also integrate the WavLM model into our audio encoder, which leads to improved performance. The comparison of SpkErr and DER results between AFL-Net and DyViSE can be found in Table 2.

4. Experiment Results

4.1. Performance comparison with baselines

The comparison between the AFL-Net and the AVR-Net is depicted in Table 1. It shows that the proposed AFL-Net consistently surpasses the AVR-Net on both the AVA-AVD and combined datasets. Notably, the AFL-Net trained on the AVA-AVD dataset achieves a DER of 23.65%, significantly exceeding the baseline system with a relative DER reduction of 13.8%, thus setting a new state-of-the-art (SOTA) performance. Given that the MR and FAR performance remains identical between the proposed and baseline systems, the DER reduction can be attributed to a decrease in the speaker error rate. This reduction

Table 1: *Performance comparison between AVR-Net and AFL-Net. AVA-AVD refers to models trained solely on the AVA-AVD dataset, while w/ Extra Data indicates models trained on a combination of Voxceleb1, Voxceleb2, and AVA-AVD.*

Models	AVA-AVD		w/ Extra Data	
	SpkErr ↓	DER ↓	SpkErr ↓	DER ↓
AVR-Net	24.88%	27.43%	18.58%	21.13%
AFL-Net	**21.10%**	**23.65%**	**17.10%**	**19.65%**

Table 2: *Performance comparison between DyViSE and AFL-Net (with WavLM) on AVA-AVD dataset.*

Models	FAR ↓	MR ↓	SpkErr ↓	DER ↓
DyViSE [11]	0.0	**1.98%**	20.86%	23.46%
AFL-Net + WavLM	**0.0**	2.55%	**19.57%**	**22.12%**

confirms that the proposed AFL-Net exhibits superior identity discrimination compared to the AVR-Net. Moreover, by integrating large-scale datasets like VoxCeleb1 and VoxCeleb2 into the training process, both systems achieve lower SpkErr and DER performance. Consistently, the AFL-Net achieves a DER of 19.65%, surpassing the AVR-Net with a relative DER reduction of 7.0%. It is also worth noting that AFL-Net experiences a modest increase in size (58.0M total, 11.6M trainable parameters) compared to AVR-Net (47.5M total, 11.4M trainable parameters), but achieves a remarkable DER reduction. We have provided some demo audios to showcase how AFL-Net outperforms AVR-Net in challenging scenarios[1].

The comparison with DyViSE is presented in Table 2. As no open-source codes were made available, we cite the performance reported in [11] directly for comparison purposes. It's noteworthy that even without incorporating the WavLM into our audio encoder, our model still achieves comparable results (DER: 23.65% v.s. 23.46% in [11]), despite our audio encoder being more than 30 times smaller. After integrating the WavLM into the audio encoder, AFL-Net achieves a lower DER of 22.12%, thereby outperforming DyViSE.

4.2. Performance comparison under varying missing rates of the visual features

We conducted experiments to validate the efficacy of AFL-Net under varying rates of visual feature absence, as depicted in Fig. 2. With increasing missing rates, both models show higher DER, highlighting visual modality's impact. AFL-Net's faster decline is due to its use of two visual feature types. Because of the Mask strategy, AFL-Net still outperforms the baseline in high missing rate. The proposed system consistently outperforms the baseline across different missing visual feature rates, with the performance gap narrowing as the missing rate increases. This indicates that AFL-Net surpasses AVR-Net in effectively utilizing visual information to enhance diarization performance. This improvement can be attributed to the integration of the lip movement modality and the two-step cross-attention fusion mechanism.

4.3. Ablation study

Table 3 compares four fusion implementations for three modalities. The results indicate that all three two-step two-modality

[1] https://afl-net.github.io/afl-net/

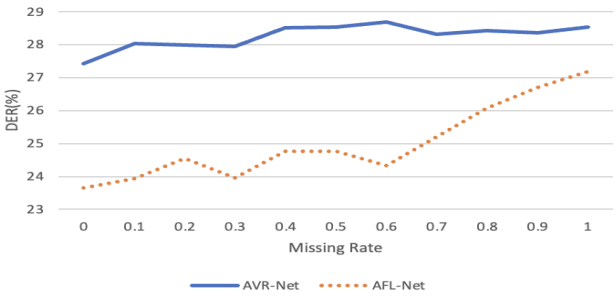

Figure 2: *The DER performance comparison on AVA-AVD dataset under varying missing rates of the visual features.*

Table 3: *Performance comparison among 4 fusion implementations on AVA-AVD dataset. In two-step two-modality fusion strategies, the modalities fused in the first step are enclosed in brackets. A+F+L denotes the one-step three-modality fusion.*

	(A+F)+L	(A+L)+F	(F+L)+A	A+F+L
SpkErr	**21.10%**	22.11%	22.00%	22.59%
DER	**23.65%**	24.66%	24.55%	25.14%

fusion strategies outperform the one-step three-modality fusion strategy. This is likely because the two-step two-modality strategies are easier for the model to learn and yield better results. Furthermore, the best performance is achieved when initially fusing the audio and face modalities, followed by the inclusion of the lip movement modality in the second step.

An ablation study was carried out, as shown in Table 4. Beginning with the AFL-Net, we sequentially remove each modification in the following order: the integration of the lip movement modality, the masking strategy, and the cross-attention mechanism. It's important to note that after the removal of these three modifications, the AFL-Net reverts to the AVR-Net. From the table, we observe a consistent decline in performance following the removal of each modification, thereby validating the effectiveness and necessity of each modification.

Table 4: *Ablation study on AVA-AVD and w/ Extra Data. Results are expressed in percentages (%).*

Models	AVA-AVD		w/ Extra Data	
	SpkErr ↓	DER ↓	SpkErr ↓	DER ↓
AFL-Net	**21.10**	**23.65**	**17.10**	**19.65**
– lip movement	21.79	24.34	17.58	20.13
– masking	22.80	25.35	18.22	20.77
– cross attention	24.88	27.43	18.58	21.13

5. Conclusion

This work introduces the AFL-Net, enhancing AVR-Net in three key ways. Firstly, it introduces a dynamic lip movement modality to improve identity discrimination. Secondly, it incorporates a two-step cross-attention method for effective modality fusion. Lastly, it proposes a masking strategy during training to encourage the system to rely more on the audio modality, which directly relates to the speaker's identity. Experimental results consistently demonstrate that AFL-Net outperforms state-of-the-art baselines, including AVR-Net and DyViSE.

6. 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 & Language*, vol. 72, p. 101317, 2022.

[2] X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou, "Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding," *arXiv preprint arXiv:2311.11375*, 2023.

[3] 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.

[4] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[5] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

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

[7] K. Fanaras, A. Tragoudaras, C. Antoniadis, and Y. Massoud, "Audio-visual speaker diarization: Improved voice activity detection with cnn based feature extraction," in *2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2022, pp. 1–4.

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

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

[10] 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.

[11] A. Wuerkaixi, K. Yan, Y. Zhang, Z. Duan, and C. Zhang, "Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2022, pp. 1–6.

[12] W. Wang, X. Qin, and M. Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9171–9175.

[13] E. Zhongcong Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "Ava-avd: Audio-visual speaker diarization in the wild," *arXiv e-prints*, pp. arXiv–2111, 2021.

[14] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[15] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[16] 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.

[17] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of classification*, vol. 1, pp. 7–24, 1984.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[22] L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin, and C.-H. Lee, "Speaker diarization with enhancing speech for the first dihard challenge." in *Interspeech*, 2018, pp. 2793–2797.

[23] B. Sredojev, D. Samardzija, and D. Posarac, "Webrtc technology overview and signaling solution design and implementation," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2015, pp. 1006–1009.

[24] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[25] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

[26] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[27] 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.

[28] J. Tian, X. Hu, and X. Xu, "Royalflush speaker diarization system for icassp 2022 multi-channel multi-party meeting transcription challenge," *arXiv preprint arXiv:2202.04814*, 2022.