ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095883

CROSS-MODAL AUDIO-VISUAL CO-LEARNING FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Meng Liu[1,2], *Kong Aik Lee*[2,3,*] , *Longbiao Wang*[1,*], *Hanyi Zhang*[1], *Chang Zeng*[4], *Jianwu Dang*[1]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Singapore Institute of Technology, Singapore
[4]National Institute of Informatics, Tokyo, Japan

ABSTRACT

Visual speech (i.e., lip motion) is highly related to auditory speech due to the co-occurrence and synchronization in speech production. This paper investigates this correlation and proposes a cross-modal speech co-learning paradigm. The primary motivation of our cross-modal co-learning method is modeling one modality aided by exploiting knowledge from another modality. Specifically, two cross-modal boosters are introduced based on an audio-visual pseudo-siamese structure to learn the modality-transformed correlation. Inside each booster, a max-feature-map embedded Transformer variant is proposed for modality alignment and enhanced feature generation. The network is co-learned both from scratch and with pretrained models. Experimental results on the test scenarios demonstrate that our proposed method achieves around 60% and 20% average relative performance improvement over baseline unimodal and fusion systems, respectively.

Index Terms— visual speech, co-learning, cross-modal, lip biometrics, speaker verification

1. INTRODUCTION

Audio-visual lip biometrics [1] utilizing auditory speech (i.e. spoken utterances) and visual speech (i.e., lip motion) has raised increasing attention recently [2]. Unlike visual biometrics using static face or iris images [3], lip motion has a dynamic temporal behavior that could be aligned with auditory speech [4]. When a person speaks, his/her voice, lip motion, and spoken words are three closely bound modalities of audio, visual, and text [5] and vary from utterance to utterance. Therefore, it is difficult to forge these identity characteristics at the same time [6]. Due to these advantages, audio-visual lip biometrics could be applied to various mobile applications and highly secured financial identification systems.

Over the past decades, the development of lip biometrics has undergone a significant change from the classic machine-learning approaches to data-centric deep-learning models. In the former, support vector machine (SVM), hidden Markov model (HMM) [1], and Gaussian mixture model (GMM) [2] were used in conjunction with appearance-based [7] and shape-based [8] features derived from lip geometry. In the latter, lip image sequences were usually directly fed into deep neural networks [9]. In [10], audio-visual speaker embedding was extracted from lip sequences with a pretrained AV-HuBERT model. However, these methods focused more on modality fusion but overlooked the correlation between auditory and visual speech [11, 12, 13].

In [9], we presented the DeepLip system that fuses complementary information derived from auditory and visual speech. The visual stream uses a multi-stage convolutional neural network (MCNN) to extract visual speaker embedding. The audio stream employs an x-vector system [14] to extract audio speaker embedding. Although it achieves a satisfactory performance, the concurrency of auditory and visual speech is largely ignored in the preliminary work: no modality-transferred knowledge was learned during training.

To realize modalities transfer, we need to tackle two problems: i) frame lengths of audio and visual modalities are unaligned due to their different sampling rates; ii) the transferred auditory/visual speech may import new knowledge and feature noise from the other modal feature space at the same time. Other recent works on multi-modal correlation studies on text, vision, and speech include linear, bilinear projection [15], gate network [16], and cross-attention [17]. Among these methods, cross-attention is an elegant approach for aligning two temporal sequences of different lengths.

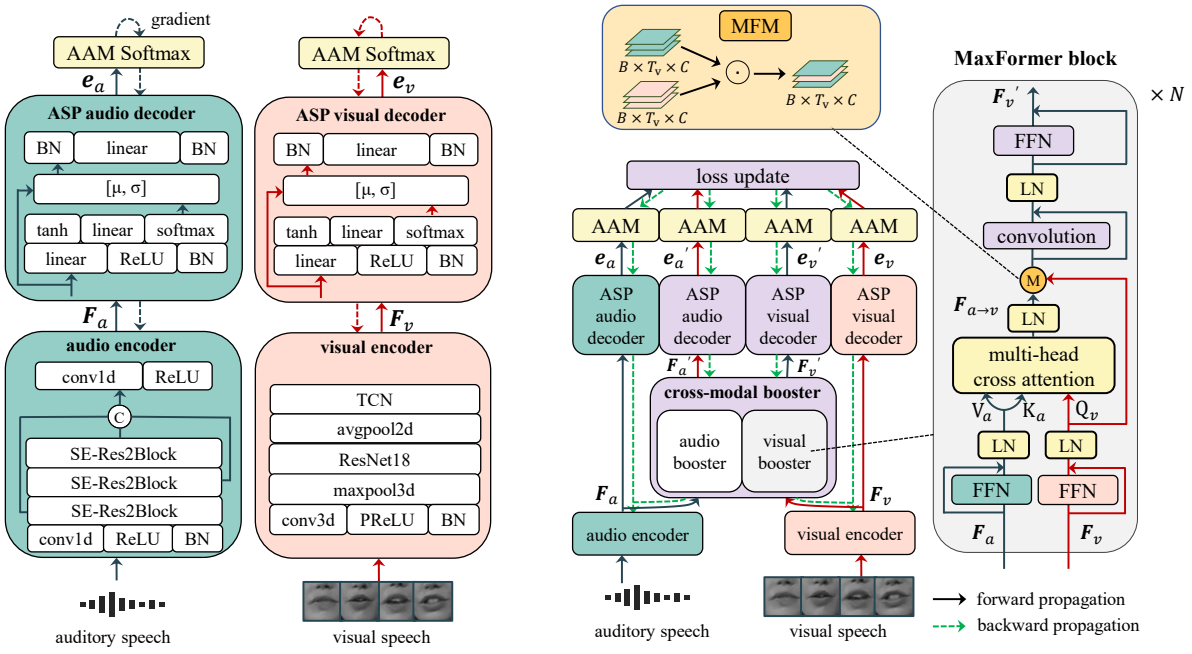
In this work, we study the cross-modal correlation between auditory and visual speech. We refer to this task as cross-modal co-learning. The key challenge of cross-modal co-learning is modality transfer [18], i.e., modeling one modality aided by exploiting knowledge from another modality using the transfer of knowledge between modalities. We propose a MaxFormer, which manages the cross-modal temporal alignment and knowledge transfer. Furthermore, we validate our method on the compiled audio-visual lip (AVL) database. Code and dataset could be found here[1].

2. CROSS-MODAL SPEECH CO-LEARNING

Figure 1 illustrates our baseline and the proposed cross-modal auditory/visual speech co-learning systems. The baseline systems use two independent branches to learn the audio and visual representation without inter-modality interaction. Specifically, the audio-only encoder involves an ECAPA-TDNN [19] structure, while the visual-only encoder uses an MCNN [9]. Based on an audio-visual pseudo-siamese structure, we construct the cross-modal co-learning network, illustrated in Fig. 1(b). The whole network consists of two encoders, two cross-modal boosters, and four decoders.

* Corresponding Authors

[1]https://github.com/DanielMengLiu/AudioVisualLip



(a) Independent training (baseline) **(b) Cross-modal co-learning (proposed)**

Fig. 1. Audio-visual lip biometrics baseline and the proposed cross-modal co-learning framework.

2.1. Encoders and decoders

In the audio encoder, the extracted fbank feature is processed by a series of conv1d, ReLU, and batch normalization units. The output frame-level features are concatenated after three SE-Res2Blocks. We use a conv1d with ReLU activation function to obtain a low-dimensional feature map $\boldsymbol{F}_a \in \mathbb{R}^{C_a \times T_a}$, where C_a and T_a corresponds to channel and frame size, respectively.

In the visual encoder, the gray-scale lip sequence $\boldsymbol{L} \in \mathbb{R}^{H \times W \times T_v}$ is processed by a front-end 3D convolution to extract short-term temporal-spatial visual features, where H, W and T_v denote height, width and frame size, respectively. Then a vanilla 18-layer residual network is used to extract spatial features. After that, we apply average pooling to obtain the visual embeddings on each channel. Finally, a temporal convolutional network (TCN) [20] captures long-term temporal lip movements $\boldsymbol{F}_v \in \mathbb{R}^{C_v \times T_v}$, where C_v and T_v denote the channel and frame size, respectively.

All the decoders in this paper apply the attentive statistics pooling (ASP) [21] decoder, which aggregates the attentive statistics pooling and an embedding affine layer. We use attention without global context [19] for the calculation of ASP. After each decoder, we obtain the output embedding $\boldsymbol{e}$.

2.2. Cross-modal booster

To leverage the information from the other modality and fully utilize the concurrency of the audio and visual modalities, we propose the cross-modal boosters. Since the audio and visual booster are symmetric in our co-learning network and share a similar model structure, we take the visual MaxFormer block as an example. Each cross-modal booster contains a stack of MaxFormer blocks, illustrated in Fig. 1(b).

Inside each MaxFormer block, the frame-level audio feature $\boldsymbol{F}_a$ and visual feature $\boldsymbol{F}_v$ are first affine transformed to d-dimensional features. After a feed-forward network (FFN) and a layer normalization (LN) block, we construct the input features to a cross-modal triplet (query $\boldsymbol{Q}_v$, key $\boldsymbol{K}_a$, and value $\boldsymbol{V}_a$). Then, the multi-head cross-modal attention calculates the correlation across modalities and aligns the key to the query. The transferred feature $\boldsymbol{F}^i_{a\to v}$ derived from the i-th single head ($i \in \{1, \cdots, m\}$, m demotes the number of heads) is detailed as follows:

$$\boldsymbol{F}^i_{a\to v} = \text{softmax}\left(\frac{(\boldsymbol{Q}_v \boldsymbol{W}^i_{Q_v})(\boldsymbol{K}_a \boldsymbol{W}^i_{K_a})^T}{\sqrt{d}}\right)\boldsymbol{V}_a \boldsymbol{W}^i_{V_a}, \quad (1)$$

where $\boldsymbol{W}^i$ is the corresponding single head weight. Then the outputs from all attention heads are concatenated and the transferred modality feature $\boldsymbol{F}_{a\to v}$ is obtained as follows:

$$\boldsymbol{F}_{a\to v} = [\boldsymbol{F}^1_{a\to v}, \ldots, \boldsymbol{F}^i_{a\to v} \ldots, \boldsymbol{F}^m_{a\to v}]\boldsymbol{W}_{a\to v} \quad (2)$$

To solve the second issue highlighted in Section 1, we introduce the max feature map (MFM) operation [23]. The MFM plays the role of local feature selection in Maxformer. It selects the optimal features learned from different modalities at different locations. In the case of backpropagation, it induces a gradient of 0 and 1 to inhibit or activate neurons. The MFM operation can obtain more competitive nodes by activating the maximum value of the feature map, thus reducing feature noise and distortion. The transformed visual speech $\boldsymbol{F}_{a\to v}$ and the original visual speech compete to compose the enhanced visual speech $\boldsymbol{F}'_v$, shown as follows:

$$\boldsymbol{F}'_v = \mathcal{G}_{\theta_2}(max(\mathcal{F}_{\theta_1}(\boldsymbol{F}_v), \boldsymbol{F}_{a\to v})), \quad (3)$$

where $\mathcal{F}_{\theta_1}(\cdot)$ is parameter function of layers before the MFM module and $\mathcal{G}_{\theta_2}(\cdot)$ corresponds to layers after the MFM module, including convolution [24], LN and FFN modules.

Table 1. Dataset statistics of compiled audio-visual lip database (M: male, F: female).

	Subset	#Speakers	#Utterances	Scenario	Video Resolution	Source Task
LRSLip3-Train	Train	4,004	31,982	TED	224x224	Lipreading
LRSLip3-Test	Test	412	1,321	TED	224x224	Lipreading
LomGridLip	Test	24M, 30F	5,400	Lab	720x480	Lipreading
GridLip	Test	18M, 16F	34,000	Lab	360x288	Lipreading
VoxLip2-Dev	Train	3,656M, 2,338F	1,092,007	YouTube	224x224	Speaker Verification
VoxLip1-Test [22]	Test	16M, 13F	4,162	YouTube	224x224	Speaker Verification

2.3. Cross-modal co-learning loss

During the training stage, our co-learning loss $\mathcal{L}_{co}$ involves four additive angular margin softmax (AAMSoftmax) [25] loss functions with equal weights, as follows:

$$\mathcal{L}_{co} = \mathcal{L}_a + \mathcal{L}_v + \mathcal{L}_a^{'} + \mathcal{L}_v^{'}, \tag{4}$$

where $\mathcal{L}_a$ and $\mathcal{L}_v$ represent the loss of the audio and visual speaker embeddings, respectively. $\mathcal{L}_a^{'}$ and $\mathcal{L}_v^{'}$ denote the loss of the transferred audio and visual embeddings, respectively.

2.4. Modality fusion

The auditory speech score s_a, visual speech score s_v, transferred auditory speech score $s_a^{'}$ and transferred visual speech score $s_v^{'}$ follow two symmetric fusion strategies. The audio- and visual-driven fusion are calculated as follows:

$$s_{a-driven} = 0.5 \cdot s_a + 0.25 \cdot s_v + 0.25 \cdot s_v^{'} \tag{5}$$

$$s_{v-driven} = 0.5 \cdot s_v + 0.25 \cdot s_a + 0.25 \cdot s_a^{'} \tag{6}$$

We set the weights considering the contribution of primary, auxiliary, and transferred modalities. The weights could be determined according to the specific scenario.

3. EXPERIMENTS

3.1. Data description

As shown in Table 1, we compile an audio-visual lip biometrics dataset from LRS3 [26], LombardGrid [27], Grid [28], VoxCeleb1 [22], VoxCeleb2 [3]. Almost all the subsets have a similar gender distribution. The LRS3Lip3 and VoxLip subsets were collected from the Internet, covering a wide range of accents, ages, ethnicities, and languages, while LomGridLip and GridLip were collected with lab cameras that the volunteers were stuff and students aging from 18 to 30 years old. The LRS3Lip3 and VoxLip subsets have poor lip resolution due to a transmission loss of the Internet. Verification trials are drawn from the cross-pairs of all test utterances randomly, as shown in Table 2. Since some videos in VoxCeleb1 test list were not available from YouTube, we could only download 29/40 test speakers, with 29,690/37,611 selected pairs. We refer to the new trial as Vox1-O-29.

Table 2. Statistics of constructed trial pairs.

Trial Name	#Pairs	#Positive	#Negative
LRSLip3-O	13,064	3,064	10,000
LomGridLip-O	20,000	4,000	16,000
GridLip-O	20,000	4,000	16,000
Vox1-O-29	29,690	15,057	14,633
Vox1-O	37,611	18,802	18,809

3.2. Experimental setup

We train the network for 40 epochs with a batch size equal to 128, a multi-step learning scheduler (milestones are 10 and 15, gamma is 0.1), an initial learning rate of 0.001, and a weight decay of 1e-7, with the Adam optimizer. The scale and margin of AAMSoftmax are set to 30 and 0.2, respectively. During the training stage, 2 seconds of audio-visual clips are used. For visual clips, we use 50×96×96 pixels and a random horizontal flip for augmentation. For audio clips, we extract 80-dimensional fbanks. Within an epoch, we apply the equal possibility of choosing clean, noisy (generated using MUSAN [29] and RIRs [30]) and spec augmentation [31] samples.

The audio encoder uses a 512-dimensional channel. The visual encoder has two layers of TCN with a kernel size of 5. Each cross-modal booster uses three blocks of MaxFormer, and d is set to 128. All decoders use a 192-dimensional output embedding. Performance will be measured by providing the Equal Error Rate (EER) and the minimum normalized detection cost function (MinDCF) [19] with P_{target}=10^{-2} and C_{FA}=C_{Miss}=1.

3.3. Training on LRSLip3

The models are all trained on the LRSLip3 training set. For the baseline, the audio-only and visual-only systems are trained independently, and then their embeddings are fused. As for our proposed cross-modal co-learning method, we train the network both from scratch and with pretrained audio- and visual-only models.

Table 3 compares the performance between the baseline and the proposed co-learning systems. Our proposed system can reach EERs of 1.37%, 10.27%, 0.83%, and 0.45% on LRSLip3-O, Vox1-O-29, LomGridLip-O, and GridLip-O test sets, and the minDCFs reduce by 28.9%, 4.9%, 20.8%, and 34.0%, respectively. Generally speaking, the transferred audio and visual speech achieved improvement compared with the corresponding unimodal systems. The fusion system further improves the performance, which indicates that our proposed cross-modal boosters have learned new knowledge. However, we notice that the performance of our co-learning model (from scratch) declines a bit on GridLip-O, which may be due to overfitting to the training set.

Another interesting finding is that the visual speech performs better than the auditory speech on the LomGridLip and GridLip sets, and vice versa. The LRSLip3 and VoxLip sets were collected from the Internet with poor visual resolution, and the LomGridLip and GridLip sets were collected using lab cameras with high visual resolution. Experimental results indicate the benefit of choosing visual-

Table 3. Performance comparison between the baseline and cross-modal co-learning systems on various audio-visual lip test sets.

	System	param	LRSLip3-O		Vox1Lip-O-29		LomGridLip-O		GridLip-O	
			EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF
baseline	audio-only	6.19M	3.92%	0.2134	15.18%	0.6269	5.28%	0.2989	4.95%	0.3291
	visual-only	6.78M	4.22%	0.1756	18.08%	0.6850	2.01%	0.0967	2.13%	0.1241
	avfusion		**1.86%**	**0.0647**	**11.03%**	**0.5423**	**0.70%**	**0.0514**	**0.80%**	**0.0471**
co-learn from scratch	audio	6.19M	3.85%	0.2194	15.65%	0.6270	5.38%	0.3086	5.00%	0.3590
	visual	6.78M	4.08%	0.1526	18.20%	0.6756	1.80%	0.1148	2.71%	0.1596
	audio-transferred	0.89M	1.73%	0.0777	15.61%	0.6350	1.68%	0.1135	1.40%	0.0926
	visual-transferred	0.89M	2.16%	0.1019	15.79%	0.6687	1.20%	0.0839	1.45%	0.1006
	audio-driven fusion		**1.37%**	**0.0460**	**11.00%**	**0.5384**	-			
	visual-driven fusion		-				**0.83%**	**0.0414**	**0.85%**	**0.0538**
co-learn with pretrained	audio	6.19M	4.16%	0.2220	15.12%	0.6215	5.54%	0.3242	5.13%	0.3796
	visual	6.78M	3.92%	0.1607	17.62%	0.6641	1.82%	0.0794	1.80%	0.1184
	audio-transferred	0.89M	2.55%	0.0994	16.32%	0.6225	2.09%	0.0994	0.80%	0.0568
	visual-transferred	0.89M	3.64%	0.1461	17.52%	0.6483	1.86%	0.1011	2.21%	0.1395
	avfusion		1.70%	0.0612	10.77%	0.5246	0.82%	0.0519	0.74%	0.0464
	audio-driven fusion		**1.50%**	**0.0504**	**10.27%**	**0.5159**	-			
	visual-driven fusion		-				**0.90%**	**0.0407**	**0.45%**	**0.0311**

or audio-driven fusion according to data conditions. With a high-resolution camera, we can utilize visual speech for near-field audio-visual authentication, e.g., with the mobile phone or at the bank counter.

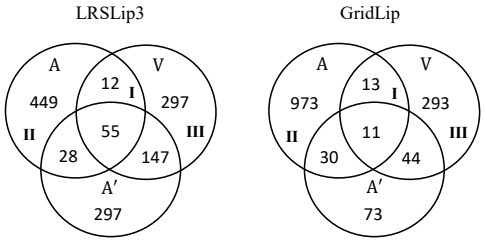


Fig. 2. Analysis on wrong verification predictions using the co-learning model (A: auditory speech, V: visual speech, A': transferred auditory speech).

Furthermore, we analyze the verification predictions on the LRSLip3 and GridLip sets using the co-learning method. As shown in Fig. 2 (left), area I represents twelve verification pairs that are wrongly predicted by both audio and visual speeches but correctly predicted by transferred auditory speech. It is the new knowledge learned via modality transfer and does not exist in the original audio or visual feature space. Area II and III contain new knowledge that benefits for fusion systems.

3.4. Training on VoxLip

As shown in Table 3, the cross-modal booster seems to have a mismatch on the VoxLip1 set, in which both biometric modalities have poor performance; thus, correlation is hard to learn. Furthermore, the train set of LRSLip3 has limited utterances which are challenging to train a robust model for the VoxLip1 set. Therefore, we train on the VoxLip2 set and co-learn with the pretrained model.

As shown in Table 4, without AS-norm, the baseline fusion system achieves an EER of 1.14% and a minDCF of 0.0759 on Vox1-O-29. With the cross-modal knowledge, the system performance could further reach an EER of 0.76% with a relative reduction of 33.3%.

Table 4. Performance comparison between the baseline and co-learning systems on the VoxCeleb1 test sets (w/o AS-norm).

System	Vox1-O-29		Vox1-O	
	EER	minDCF	EER	minDCF
audio-only	1.21%	0.0971	1.16%	0.0877
visual-only	6.75%	0.2189		-
fusion	1.14%	0.0759		-
AV-Hubert-L(AV) [10]	-	-	0.84%	-
audio-transferred	2.89%	0.1500		-
visual-transferred	5.33%	0.1919		-
audio-driven fusion	**0.76%**	**0.0559**		-

4. CONCLUSION AND FUTURE WORK

This paper investigated the correlation between speech and lip motion and proposed a cross-modal speech co-learning paradigm. A cross-modal booster structure has been presented to learn the modality-transformed correlation. Our primary contribution is the cross-modal co-learning paradigm that realizes knowledge transfer between auditory and visual speech. We generated enhanced features via an MFM-embedded MaxFormer. Experiments on multiple test sets revealed the potential application scenario of the proposed audio-visual speaker recognition using lips. In the future, we will continue working on text-dependent audio-visual speaker verification using lips.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62176182. The work of Kong Aik Lee is supported by the Agency for Science, Technology and Research (A*STAR), Singapore, through its Council Research Fund (Project No. CR-2021-005).

6. REFERENCES

[1] J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker identification by lipreading," in *Proc. ICSLP 1996*, vol. 1. IEEE, 1996, pp. 62–65.

[2] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.

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

[4] T. Stafylakis and G. Tzimiropoulos, "Deep word embeddings for visual speech recognition," in *Proc. ICASSP 2018*, pp. 4974–4978.

[5] X. Liu and Y.-m. Cheung, "Learning multi-boosted hmms for lip-password based speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 2, pp. 233–246, 2013.

[6] C.-Z. Yang, J. Ma, S. Wang, and A. W.-C. Liew, "Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1841–1854, 2020.

[7] S.-L. Wang and A. W.-C. Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognition*, vol. 45, no. 9, pp. 3328–3335, 2012.

[8] M. I. Faraj and J. Bigun, "Speaker and speech recognition by audio-visual lip biometrics," in *The 2nd International Conference on Biometrics, Seoul Korea*. Citeseer, 2007, pp. 1–9.

[9] M. Liu, L. Wang, K. A. Lee, H. Zhang, C. Zeng, and J. Dang, "Deeplip: A benchmark for deep learning-based audio-visual lip biometrics," in *ASRU Workshop 2021*, pp. 122–129.

[10] B. Shi, A. Mohamed, and W.-N. Hsu, "Learning lip-based audio-visual speaker embeddings with av-hubert," in *Proc. Interspeech 2022*, pp. 4785–4789.

[11] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[12] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *Proc. ICASSP 2021*, pp. 6194–6198.

[13] R. Tao, R. K. Das, and H. Li, "Audio-visual speaker recognition with a cross-modal discriminative network," in *Proc. Interspeech 2020*, pp. 2242–2246.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP 2018*, pp. 5329–5333.

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.

[16] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079–1092, 2021.

[17] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proc. ACM MM 2020*, pp. 3884–3892.

[18] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: challenges, applications with datasets, recent advances and future directions," *Information Fusion*, vol. 81, pp. 203–239, 2022.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *in Proc. Interspeech 2020*, pp. 3830–3834.

[20] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *Proc. ICASSP 2021*, pp. 7608–7612.

[21] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, pp. 2252–2256.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, pp. 2616–2620.

[23] Z. Yang, M. Jian, B. Bao, and L. Wu, "Max-feature-map based light convolutional embedding networks for face verification," in *Chinese Conference on Biometric Recognition*. Springer, 2017, pp. 58–65.

[24] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. M. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. Interspeech 2022*, pp. 306–310.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR 2019*, pp. 4690–4699.

[26] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[27] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[29] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.

[30] 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 *Proc. ICASSP 2017*, pp. 5220–5224.

[31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, pp. 2613–2617.