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

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

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

ABSTRACT

Visual speech is highly related to auditory speech due to the co-occurrence and synchronization of lip motion in speech production. This paper investigates this correlation and proposes a cross-modal speech co-learning paradigm. Specifically, two cross-modal boosters are introduced based on an audio-visual pseudo-siamese structure to learn the modality-transformed correlation implicitly. Inside each booster, a max-feature-map embedded Transformer decoder is proposed for modality alignment and enhanced feature generation. Experimental validation on LRSLip3, GridLip, LomGridLip, and VoxLip datasets demonstrates that our proposed method finally achieves an average of 60% and 20% relative improvement over independently trained audio-only/visual-only and baseline 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 dynamic temporal behavior that could be aligned with auditory speech [4]. When a person speaks, his/her voice, lip motion, and the 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), Gaussian mixture model (GMM), and hidden Markov model (HMM) 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 this work, we study the cross-modal correlation between auditory and visual speech. We refer to this task as the cross-modal co-learning. The key challenge of cross-modal co-learning is modality transfer [14], i.e., modeling one modality aided by exploiting knowledge from another modality using the transfer of knowledge between modalities. Furthermore, we investigate audio-visual lip biometrics from the perspectives of concurrency and transfer, which is lacking in current studies.

The remaining of this paper is organized as follows. Section 2 describes the related work and our motivation. Section 3 demonstrates the proposed MaxFormer, which manages the cross-modal temporal alignment and knowledge transfer. Section 4 shows our audio-visual lip (AVL) database, experimental setup, and results. Code and dataset will be released upon publication[1].

2. RELATED WORK

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 [15] to extract audio speaker embedding. Though, 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.

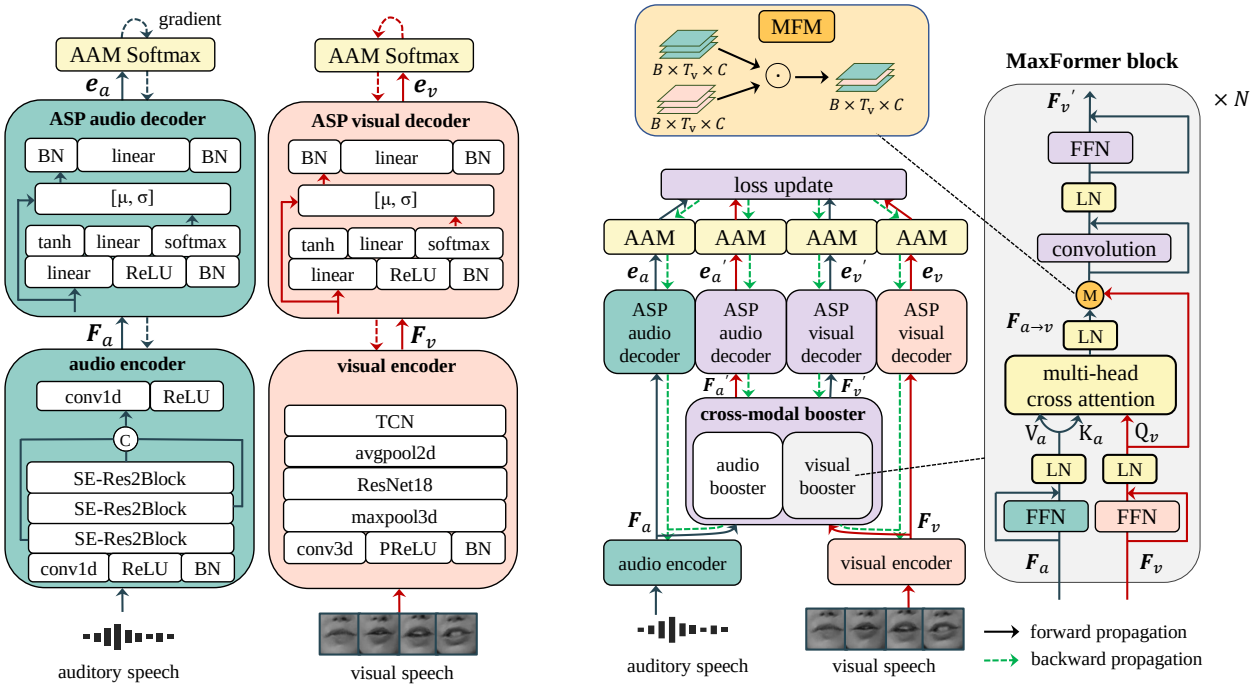
To realize modalities transfer, we need to tackle two problems. First, frame lengths of audio and visual modalities are unaligned due to their different sampling rates. Second, the transferred auditory/visual speech may import new knowledge and feature noise from the other modal feature space at the same time.

Other recent works on multi-modal correlation studies on text, vision, and speech include Linear, bilinear projection [16], gate network [17], and cross-attention [18]. Among these methods, cross-attention is an elegant approach for aligning two temporal sequences of different lengths.

3. CROSS-MODAL SPEECH CO-LEARNING

Figure 1 illustrates our baseline and the proposed cross-modal auditory/visual speech co-learning systems. The baseline systems uses 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 Figure 1(b). The whole network consists of two encoders, two cross-modal boosters, and four decoders.

[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.

3.1. Encoders and decoders

In the audio encoder, the extracted fbank feature ($B \times F_a \times T_a$, corresponding to batch, feature, frame sizes) is first processed by a series of conv1d, ReLU, and batch normalization units. The output frame-level features are concatenated after three SE-Res2Blocks. Finally, we use a conv1d with ReLU activation functions to transform the high-dimensional feature maps to low-dimensional $\boldsymbol{F}_a$ ($B \times C_a \times T_a$, corresponding to batch, channel, frame sizes).

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

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.

3.2. 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. The audio and visual booster are symmetric in our co-learning network and share a similar model structure. Each cross-modal booster contains a stack of MaxFormer blocks, illustrated in Figure 1(b).

Take the visual MaxFormer block as an example. Inside each MaxFormer block, The frame-level audio feature F_a and visual feature 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 pair (query, key, and value). Then, the multi-head cross-modal attention calculates the correlation across modalities and aligns the value to the query. The transferred feature $F_i^{a\to v}$ derived from the i-th single head are detailed as follows:

$$F_i^{a\to v} = \text{softmax}\left(\frac{(Q_v W_i^{Q_v})(K_a W_i^{K_a})^T}{\sqrt{d}}\right) V_a W_i^{V_a} \quad (1)$$

where W_i is the corresponding single head weight. Then the outputs from all attention heads are concatenated and the transferred modality feature $F_i^{a\to v}$ is obtained as fllows:

$$F_{a\to v} = [F_1^{a\to v}, \ldots, F_{\text{m}}^{a\to v}]W^{a\to v} \quad (2)$$

To solve the second issue highlighted in Section 2, we introduce the max feature map (MFM) operation [23]. The MMF 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 $F_{a\to v}$ and the original visual speech compete to compose the enhanced visual speech $F_v^{'}$, shown as follows:

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

where $\mathcal{F}_{\theta_1}(\cdot)$ is the network parameters before MFM and $\mathcal{G}_{\theta_2}(\cdot)$ is the network parameters after MFM, including convolution [24], LN and FFN modules.

Table 1. Dataset statistics of compiled audio-visual lip database.

	Subset	#Speakers	#Utterances	Scenario	Video Resolution	Source Task
LRSLip3-Trainval	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	61%M, 39%F	1,092,009	YouTube	224x224	Speaker Verification
VoxLip1-Test [22]	Test	55%M, 45%F	4,874	YouTube	224x224	Speaker Verification

3.3. Cross-modal Co-learning loss

During the training stage, our co-learning loss $\mathcal{L}_{co}$ involves four additive angular margin (AAM) softmax [25] 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 $\mathcal{L}_v$ cater to the audio and visual speaker embeddings. $\mathcal{L}_a^{'}$ and $\mathcal{L}_v^{'}$ denote the transferred audio and visual embeddings.

3.4. Modality fusion

Fusion of 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 types of strategies. The audio- and visual-driven fusion strategy calculate 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 fixed weights here considering contribution of main modality, complementary auxiliary modality, and transferred auxiliary modality. It could also be tuned according to the specific scene.

4. EXPERIMENTS

4.1. 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]. Verification trials are drawn from the cross-pairs of all test utterances randomly, as shown in Table 2.

Table 2. Details of constructed trial pairs.

Trial Name	#Trials	#Target	#Non-target
LRSLip3-O	13,064	3,064	10,000
LomGridLip-O	31,269	867	30342
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

4.2. 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 and RIR toolkits) and spec augmentation 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.

4.3. LRSLip3 results

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 the audio and visual embeddings are fused. As for our proposed cross-modal co-learning method, we train the model both from scratch and pretrained.

Table 3 compare the performance between the baseline and the proposed co-learning systems. Our proposed system can reach EERs of 1.37%, 10.27%, 0.90%, and 0.45% on LRSLip3-O, VoxLip1-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.

Another interesting finding is that the visual speech performs better than the auditory speech on LomGridLip and GridLip sets, and vice versa. The LRSLip3 and VoxLip were collected from the Internet (poor visual resolution), and the LomGridLip and GridLip were collected using lab cameras (high visual resolution). Experimental results indicate the benefit of choosing visual- or audio-driven fusion according to data conditions. With a good quality camera, we can utilize visual speech (lip biometrics) for near-field audio-visual authentication, e.g., with the mobile phone or at the bank counter.

4.4. VoxLip results

As Table 3 shows, the cross-modal booster seems to have a mismatch on VoxLip1, which both biometrics 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 VoxLip1. Therefore, we train on the VoxLip2 and co-learn from pretrained.

Table 4 shows the results. Without AS-norm, the baseline fusion system achieves an EER of 1.08% and a minDCF of 0.0833 on Vox1-O-29. With the cross-modal knowledge, the system performance could further reach an EER of 0.86%, which relatively reduce 20.3%.

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	5.02%	0.2814	4.95%	0.3291
	visual-only	6.78M	4.22%	0.1756	18.08%	0.6850	1.97%	0.0946	2.13%	0.1241
	avfusion		**1.86%**	**0.0647**	**11.03%**	**0.5423**	**0.92%**	**0.0507**	**0.80%**	**0.0471**
co-learn from scratch	audio	6.19M	3.85%	0.2194	15.65%	0.6270	4.96%	0.2998	5.00%	0.3590
	visual	6.78M	4.08%	0.1526	18.20%	0.6756	1.78%	0.1119	2.71%	0.1596
	audio-transferred	0.89M	1.73%	0.0777	15.61%	0.6350	1.73%	0.1225	1.40%	0.0926
	visual-transferred	0.89M	2.16%	0.1019	15.79%	0.6687	1.31%	0.0627	1.45%	0.1006
	audio-driven fusion		**1.37%**	**0.0460**	**11.00%**	**0.5384**	-			
	visual-driven fusion		-				**0.69%**	**0.0455**	**0.85%**	**0.0538**
co-learn with pretrained	audio	6.19M	4.16%	0.2220	15.12%	0.6215	5.19%	0.2949	5.13%	0.3796
	visual	6.78M	3.92%	0.1607	17.62%	0.6641	1.61%	0.0882	1.80%	0.1184
	audio-transferred	0.89M	2.55%	0.0994	16.32%	0.6225	2.27%	0.0932	0.80%	0.0568
	visual-transferred	0.89M	3.64%	0.1461	17.52%	0.6483	1.73%	0.1009	2.21%	0.1395
	avfusion		1.70%	0.0612	10.77%	0.5246	0.81%	0.0429	0.74%	0.0464
	audio-driven fusion		**1.50%**	**0.0504**	**10.27%**	**0.5159**	-			
	visual-driven fusion		-				**0.69%**	**0.0336**	**0.45%**	**0.0311**

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

System	Vox1Lip-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**	-	

5. CONCLUSION AND FUTURE WORK

This paper investigated the correlation between speech and lip motion, and proposed an audio-visual speech co-learning paradigm. A cross-modal booster structure has been presented to learn the modality-transformed correlation. Our main contribution is the cross-modal co-learning paradigm that realize knowledge transfer between auditory and visual speech. We generated enhanced features via an MFM-embedded MaxFormer. Experiments on multiple test sets reveal 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.

6. REFERENCES

[1] Juergen Luettin, Neil A Thacker, and Steve W Beet, "Speaker identification by lipreading," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 1, pp. 62–65.

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

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[4] Themos Stafylakis and Georgios Tzimiropoulos, "Deep word embeddings for visual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4974–4978.

[5] Xin Liu and Yiu-ming 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] Chen-Zhao Yang, Jun Ma, Shilin Wang, and Alan Wee-Chung 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] Shi-Lin Wang and Alan Wee-Chung Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognition*, vol. 45, no. 9, pp. 3328–3335, 2012.

[8] Maycel Isaac Faraj and Josef Bigun, "Speaker and speech recognition by audio-visual lip biometrics," in *The 2nd International Conference on Biometrics, Seoul Korea*. Citeseer, 2007.

[9] Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, Chang Zeng, and Jianwu Dang, "Deeplip: A benchmark for deep learning-based audio-visual lip biometrics," in *2021 IEEE*

Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021, pp. 122–129.

[10] Bowen Shi, Abdelrahman Mohamed, and Wei-Ning Hsu, "Learning lip-based audio-visual speaker embeddings with av-hubert," *arXiv preprint arXiv:2205.07180*, 2022.

[11] Chao Zhang, Zichao Yang, Xiaodong He, and Li 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] Leda Sarı, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf, "A multi-view approach to audio-visual speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6194–6198.

[13] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, "Audio-visual speaker recognition with a cross-modal discriminative network," *Proc. Interspeech 2020*, pp. 2242–2246, 2020.

[14] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha, "Multimodal co-learning: challenges, applications with datasets, recent advances and future directions," *Information Fusion*, vol. 81, pp. 203–239, 2022.

[15] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[16] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.

[17] Yanmin Qian, Zhengyang Chen, and Shuai Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[18] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3884–3892.

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

[20] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic, "Towards practical lipreading with distilled and efficient models," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7608–7612.

[21] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[22] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

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

[24] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," *arXiv preprint arXiv:2203.15249*, 2022.

[25] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos 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] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[27] Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy 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] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu 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.