INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-modality Matters: A Performance Leap on VoxCeleb

Zhengyang Chen, Shuai Wang, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai
{zhengyang.chen, feixiang121976, yanminqian}@sjtu.edu.cn

Abstract

The information from different modalities usually compensates each other. In this paper, we use the audio and visual data in VoxCeleb dataset to do person verification. We explored different information fusion strategies and loss functions for the audio-visual person verification system at the embedding level. System performance is evaluated using the public trail lists on VoxCeleb1 dataset. Our best system using audio-visual knowledge at the embedding level achieves **0.585%, 0.427% and 0.735% EER** on the three official trial lists of VoxCeleb1, which are the best reported results on this dataset. Moreover, to imitate more complex test environment with one modality corrupted or missing, we construct a noisy evaluation set based on VoxCeleb1 dataset. We use a data augmentation strategy at the embedding level to help our audio-visual system to distinguish the noisy and the clean embedding. With such data augmented strategy, the proposed audio-visual person verification system is more robust on the noisy evaluation set.

Index Terms: person verification, multi-modal information fusion, embedding, data augmentation

1. Introduction

Multiple biometric characteristics could be used to verify a person's identity, where speech and face are two typical ones. Accordingly, face verification and speaker verification are hot research topics in the biometric field. The recent thriving deep learning technologies greatly boost the performance of both tasks. Different architectures [1, 2, 3, 4] and different loss functions [5, 6, 7, 8] have been investigated by the researchers in the past few years, leading to well-performing systems which can even be commercialized for real-world applications.

Despite the success in single modality applications, multi-modal learning has attracted more and more attention from academia and industry. The motivation comes in two folds.

1. The complementary information from different modalities could improve system performance.
2. Models built from multiple modalities tend to be more robust and fault-tolerant, and the failures in the single modality could be fixed or suppressed.

Audio and vision are two most commonly used information sources, and a lot of related multi-modal learning work has been carried out [9, 10, 11]. Researchers investigated to fuse the lip information from video data with the audio features to help speech recognition [12, 13] or speech separation tasks [14, 15, 16, 17]. In the biometric recognition field, many researchers found that simply fusing the scores from the face recognition and speaker recognition systems could obtain impressive results [18, 19, 20, 21, 22]. Authors in [23] tried to fuse the audio-visual information at the embedding level to improve the online person verification system.

Similar to [23], in this paper, cross-modality integration is carried out at the embedding level, where the more powerful segment-level trained speaker embeddings are used. Different fusion strategies and loss functions are investigated and compared in the multi-modal learning framework.

Moreover, to imitate the real-world scenes, we constructed a noisy evaluation set with one modality corrupted or missing. To compensate the performance degradation, a novel embedding-level noise distribution matching (NDM) data augmentation method [24] is proposed, which greatly improved the performance under noisy condition.

All the systems are evaluated on the standard VoxCeleb1 dataset, and our best multi-modal system achieves $0.585\%, 0.427\%$ and 0.735% EER on the three trial lists (VoxCeleb1_O, VoxCeleb1_E and VoxCeleb1_H) respectively, which are the best reported results on this dataset to our knowledge. Furthermore, the NDM based multi-modal system shows the ability to select more salient modality information when evaluated on the noisy evaluation set.

2. Methodology

2.1. Embedding Level Multi-Modality Fusion

In this section, we will introduce three approaches to fuse the face embedding $\mathbf{e}_f$ and voice embedding $\mathbf{e}_v$ to one person identity embedding $\mathbf{e}_p$. As shown in Fig.1, $\mathbf{e}_f$ and $\mathbf{e}_v$ are first transformed to $\tilde{\mathbf{e}}_f \in \mathbb{R}^D$ and $\tilde{\mathbf{e}}_v \in \mathbb{R}^D$ through transform layers $f_{\text{trans\_f}}$ and $f_{\text{trans\_v}}$, respectively:

$$\begin{aligned} \tilde{\mathbf{e}}_f &= f_{\text{trans\_f}}(\mathbf{e}_f) \\ \tilde{\mathbf{e}}_v &= f_{\text{trans\_v}}(\mathbf{e}_v) \end{aligned} \tag{1}$$

The transformed $\tilde{\mathbf{e}}_f$ and $\tilde{\mathbf{e}}_v$ lie in a co-embedding space which are more suitable for the later fusion.

2.1.1. Simple Soft Attention Fusion

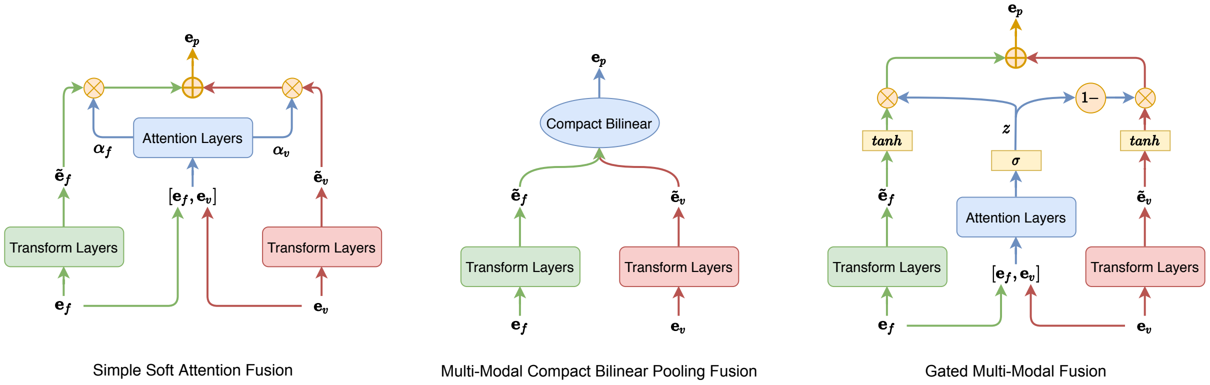
In this section, we first introduce a simple soft attention (SSA) across the modality axis used in [23]. As shown in Fig. 1 (left), given the face and voice embedding $\mathbf{e}_f$ and $\mathbf{e}_v$, the attention score $\hat{a}_{\{f,v\}} \in \mathbb{R}^2$ through attention layers $f_{\text{att}}(\cdot)$ is defined as:

$$\hat{a}_{\{f,v\}} = f_{\text{att}}\left([\mathbf{e}_f, \mathbf{e}_v]\right)$$

[†] corresponding author

This work was supported by the China NSFC project No. U1736202. The author Zhengyang Chen is supported by Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

http://dx.doi.org/10.21437/Interspeech.2020-2229



Simple Soft Attention Fusion

Multi-Modal Compact Bilinear Pooling Fusion

Gated Multi-Modal Fusion

Figure 1: *Three multi-modal fusion strategies at the embedding level*

Then the fusion embedding is calculated by the weighted sum as:

$$\mathbf{e}_p = \sum_{i \in \{f,v\}} \alpha_i \tilde{\mathbf{e}}_i, \text{ where } \alpha_i = \frac{\exp\left(\hat{a}_i\right)}{\sum_{k \in \{f,v\}} \exp\left(\hat{a}_k\right)}, i \in \{f, v\} \tag{2}$$

2.1.2. *Compact Bilinear Pooling Fusion*

Bilinear pooling fully explores the relationship between two vectors using outer product operation and has no training parameters involved. However, the outer product is usually infeasible in practice due to its high dimensionality. The work in [25] introduced a method, called multi-modal compact bilinear pooling (MCB), to approximate the outer product result and reduce the result's dimension at the same time. It's worth noting that there are no training parameters in MCB either. As shown in Fig. 1 (middle), we directly use the compact bilinear pooling to fuse the $\tilde{\mathbf{e}}_f$ and $\tilde{\mathbf{e}}_v$ to $\mathbf{e}_p$. The implementation details about compact bilinear pooling can be found in [25], which is originally used for visual question-answering system.

2.1.3. *Gated Multi-Modal Fusion*

In this section, we use a gate to control the information flow from face and voice modality, which is inspired by the flow control in recurrent architectures like GRU or LSTM, and we call it gated multi-modal fusion (GATE). The work in [26] presents the similar idea to fusion the information from image and text modality. As shown in Fig. 1 (right), given the face and voice embedding $\mathbf{e}_f$ and $\mathbf{e}_v$, a gate vector $\mathbf{z} \in \mathbb{R}^D$ can be calculated:

$$\mathbf{z} = \sigma(f_{\text{att}}([\mathbf{e}_f, \mathbf{e}_v]))$$

And then, we use the gate vector $\mathbf{z}$ to fuse $\tilde{\mathbf{e}}_f$ and $\tilde{\mathbf{e}}_v$ to $\mathbf{e}_p$, and $\odot$ denotes the element-wise product:

$$\mathbf{e}_p = \mathbf{z} \odot \tanh(\tilde{\mathbf{e}}_f) + (1 - \mathbf{z}) \odot tanh(\tilde{\mathbf{e}}_v) \tag{3}$$

2.2. Loss Function

In this section, we would introduce the loss functions we used to optimize proposed multi-modality fusion systems.

2.2.1. *Contrastive Loss With Aggressive Sampling Strategy*

The original contrastive loss is defined as:

$$\mathcal{L}_{con} = \frac{1}{N} \sum_{i, y_i = 1} D_i + \frac{1}{M} \sum_{k, y_k = 0} \max(0, m - D_k) \tag{4}$$

where D is the distance between a pair, N and M are the numbers of positive and negative pairs in a batch. $y = 1$ and $y = 0$ denote the positive and negative pair respectively, and m is the margin. In our experiment, we use cosine similarity to measure the distance of embedding pairs.

The tuned margin m in original contrastive loss makes the loss focus more on "hard" negatives. However, the "hard" positives are not considered. Here, we introduce a more aggressive sampling strategy, and similar idea is also used in [27]. During training, after the forward-propagation of the neural network, we only use a subset of γM "hardest" negatives and γN "hardest" positives ($\gamma \in (0, 1]$) to calculate the loss. Contrastive loss with new sampling strategy can be defined as:

$$\begin{aligned} \mathcal{L}_{con} = & \frac{1}{\gamma N} \sum_{i, y_i = 1} \max(0, D_i - D_{\text{p\_low}}) \\ & + \frac{1}{\gamma M} \sum_{k, y_k = 0} \max(0, D_{\text{n\_high}} - D_k) \end{aligned} \tag{5}$$

where $D_{\text{p\_low}}$ denotes the smallest distance in all "hardest" positives and $D_{\text{n\_high}}$ denotes the largest distance in all "hardest" negatives.

2.2.2. *Additive Angular Margin Loss*

In addition, we also tried the popular angular margin loss [6] in our experiment. For an input with person identity label y_s, the loss is defined as:

$$L_s = -\log \frac{e^{s \cdot \cos(\theta_{y_s} + m)}}{e^{s \cdot \cos(\theta_{y_s} + m)} + \sum_{j=1, j \neq y_s}^{n} e^{s \cdot \left(\cos \theta_j\right)}} \tag{6}$$

where m is the additive margin and s is scale parameter which can help the model converge faster. In our experiment, s is set to 32 and m is set to 0.6 in the fusion system.

2.3. Embedding Level Augmentation for Noisy Evaluation

2.3.1. *Noisy Evaluation Set Construct*

Information from different modalities is not always available or salient enough to do the verification task. In real applications, one modality is often corrupted or missing because of some inevitable external factors, such as the ambient light, the motion of people or the background noise. To address such conditions, we construct a noisy evaluation set based on VoxCeleb1 evaluation set.

For the image data, we use vertical and horizontal motion blur to imitate the motion of person and use the Gaussian blur to imitate other noises. For the audio data, three kinds of noises in Musan [28] are combined with the original data to generate the corrupted audio samples. We also consider the completely missing case of one modality by directly setting the corresponding extracted embedding to zero values. The detailed pipeline to construct this dataset is shown in Algorithm 1.

Algorithm 1: Noisy Evaluation Set Construct

```
Initialize noisy probability $p_{noise} = 0.3$. Here, we use
  $\{1, 2, 3\}$ to denotes 3 different noises which can be added to
  both modalities and use 4 to denote missing modality.
for $recording_i \in$ VoxCeleb1 do
    Randomly sample a value $\eta \in (0, 1)$
    if $\eta < p_{noise}$ then
        Randomly select a noisy type value $k \in \{1, 2, 3, 4\}$.
        Randomly select a modality type in {face, voice};
        if $k == 4$ then
            Set the seleted modality's embedding to zero
              vector;
            Extract another modality's embedding using
              recording$_i$;
        else
            Add noise with noise type $k$ to the seleted
              modality in recording$_i$ to get noisy_recording$_i$;
            $\mathbf{e}_f$ = FaceSystem(noise_recording$_i$);
            $\mathbf{e}_v$ = VoiceSystem(noise_recording$_i$);
    else
        $\mathbf{e}_f$ = FaceSystem(recording$_i$);
        $\mathbf{e}_v$ = VoiceSystem(recording$_i$);
```

2.3.2. Embedding Level Augmentation

To build a system which is more robust to the corrupted audio-visual data, an additional embedding-level augmentation strategy is proposed in this work. In our previous work, we use deep generative models such as generative adversarial network (GAN) [29] or variational autoencoder (VAE) [30] to mimic the distribution of noisy speaker embeddings. Here, instead, a simple statistics based distribution matching algorithm is used.

We randomly selected 100,000 recordings from the training set (1,092,009 recordings) and generated different types of corrupted data. Then, for each noise type, we assume the difference between the noisy embeddings and original embeddings could be described by a Gaussian distribution. After estimating the parameters of the noise distribution, we sample noise from the distribution and directly add it to the original embedding to generate a noisy embedding. We term this embedding-level augmentation method as noise distribution matching (NDM). Compared directly adding noises to the whole training set and extract augmented embeddings, NDM only uses a small portion of the training data and directly augments the embeddings, which saves both time and disk. Besides, we still use the zero vector to imitate the case of modality missing.

3. Experimental Setups

3.1. Dataset

In our experiments, we use visual and audio data from VoxCeleb1 & 2 datasets [31, 32]. For training, we use the DEV part of VoxCeleb2 dataset, which includes 5,994 speakers and 1,092,009 utterances. VoxCeleb1 is used as the evaluation set. Three official trial lists[1] Vox1-O, Vox1-E and Vox1-H are used for evaluation. It is noted that the visual data from official VoxCeleb1 dataset is incomplete, and we downloaded the missing visual data from youtube and make it public[2].

3.2. Experimental Setups

3.2.1. Single-Modality Systems

For audio data, 40-dimensional Fbank features are extracted using Kaldi toolkit [33], with silent frames removed using an energy-based voice activity detector. Then we do the CMN on the Fbank features with sliding-window size 300. For video data, we extract 1 frame per second. Then, we use MTCNN [34] to detect the face landmarks and use a similarity transformation to map the face region to the same shape (3x112x96). Finally, we normalize pixel value of each image to $[0, 1]$ and subtract 0.5 to map the value range to $[-0.5, 0.5]$.

During training, the Fbank features from one utterance is split to chunks with chunk-size from 200 to 400. During testing, we extract one voice embedding for each recording, and multiple face embeddings from one recording are averaged to obtain one single face representation.

In our experiments, the 50-layer SE-ResNet described in [35] is used for the face system and the 34-layer ResNet described in [36] is used for the voice system. Embeddings of both systems are set to dimension 512. AAM loss with a margin $m = 0.2$ are used to optimize both systems.

3.2.2. Multi-Modality System

Face and voice embeddings are extracted from the single-modality systems for all the recordings in the training set. Then, all the embeddings are L2-normalized to construct the new training set for the audio-visual multi-modality system.

For the SSA fusion systems, the transform layers are two fully connected layers both with 512 units, and the attention layer is a fully connected layer with 2 units. For compact bilinear fusion and gated multi-model fusion, the transform layers are both a fully connect layer with 512 units. The attention layers in gated multi-model fusion system are two fully connected with 32 and 512 units respectively. For all the adjacent fully connected layers above, we insert another batchnorm and relu layer in the middle.

4. Results And Analysis

4.1. Evaluation on Embedding Level Multi-Modal Fusion

To fuse the information from the face and voice modalities, different fusion strategies, different loss functions are explored and compared in our embedding-level fusion systems. The results and analysis will be presented in this section.

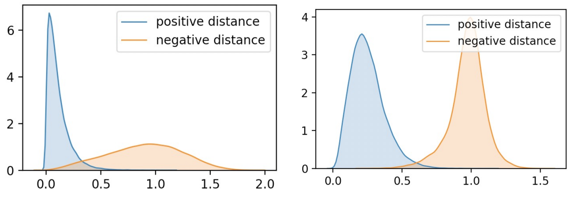
The results of single-modality system are shown at the top in Table. 1. We find that the face and the voice single-modality systems are basically comparable. As shown in the third line of the table, the result of simple score average between these two single-modality systems largely exceeds both single-modality systems, which shows the strong complementary power between audio and visual modalities.

[1] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html

[2] https://github.com/czy97/VoxCeleb1-missing-cropped-face-images

4.1.1. *Loss Functions Comparison*

Firstly, the SSA fusion strategy with the supervision of constrastive loss is investigated, which is also the best system in [23]. However, as shown in middle part of Table 1, in our experiments, the original contrastive loss based system doesn't converge to a good optimal, and the fusion system even obtained a much worse performance than the single-modality systems. To enhance the constrastive loss, the revised version with more aggressive sampling strategy introduced in Section 2.2.1 is adopted, exhibiting a much better result (SSA+Con-new). To give a more intuitive exhibition of the new strategy's effectiveness, the distribution of the distance between the positive and negative pairs is shown in Figure. 2. It shows that the new contrastive loss can enlarge the difference between the positive and negative distance. Furthermore, instead of the constrastive loss, we also used the classification based AAM-Softmax loss for the multi-modal system optimization, which substantially outperforms the constrastive loss. AAM-softmax and new contrastive loss would be mainly used for the following experiments.



(a) *Original Contrastive Loss*

(b) *New Contrastive Loss*

Figure 2: *Distance distribution of the positive / negative pairs*

4.1.2. *Fusion Strategies Comparison*

Different fusion strategies introduced in Section 2.1 are compared in this section, while the AAM-softmax loss or new contrastive loss provides the supervision signal. Results are shown in the middle part of Table. 1. From the results, all three fusion strategies achieve remarkable improvement compared with the single-modality systems, and the gated multi-modal fusion architecture performs the best. However, the simple score averaging still performs the best, which is not consistent with the findings in [23]. The possible reason is that we have much stronger single-modality systems in this work: using the same trial list of VoxCeleb2 test, we achieve 4.08% and 3.43% EER for face and voice, respectively, while the corresponding number in [23] is 14.5% and 8.03% [3]. This big difference can also attribute to the different experimental setups, and we adopted segment-level optimization in our systems, while the authors in [23] used frame-level embedding extractors to enable the online verification.

In addition, when we jointly use the AAM loss and the new contrastive loss, a further improvement is obtained and the performance on Vox1-E and Vox1-H trails exceed the score average result. The results are shown in the penultimate line of Table. 1. Surprisingly, we find the fusion system using proposed models complements the simple score average system. When we further average the score of the GATE+AAM+Con-new fusion system with the averaged score from single-modality systems, the best system performance is obtained. To the best of our knowledge, this is also the best published result for person verification on the VoxCeleb1 evaluation dataset.

Table 1: *Results comparison using different fusion strategies and losses.* ***Con-orig****: original contrastive loss.* ***Con-new****: proposed contrastive loss using more aggressive sampling strategy. The m in* ***Con-orig*** *is set to 0.5 and γ in* ***Con-new*** *is set to 0.05*

Modal	Fusion	Loss	Test Trial (EER %)		
			Vox1-O	Vox1-E	Vox1-H
Face	—	AAM	2.260	1.542	2.374
Voice	—	AAM	2.308	2.234	3.782
Voice + Face	① ScoreAvg	-	0.505	0.432	0.782
	SSA	Con-orig	5.303	4.880	10.30
		Con-new	1.766	1.192	2.452
		AAM	0.670	0.584	1.009
	MCB	Con-new	0.925	0.869	1.661
		AAM	0.803	0.604	0.997
	GATE	Con-new	1.026	1.031	2.199
		AAM	0.670	0.469	0.801
	② GATE	AAM+Con-new	0.585	0.427	0.735
	① + ② ScoreAvg	—	**0.499**	**0.379**	**0.683**

4.2. Evaluation on Corrupted and Missing Modality

To test the fusion system on the more complex real condition with one modality corrupted or missing, results are evaluated using the noisy evaluation set illustrated in section 2.3.1, and the results are shown in Table. 2. From the results, we find that simple score average operation can still significantly improve the performance, and the proposed multi-modality fusion system trained with augmented embedding data achieves the best result for this condition. Besides, the audio-visual fusion system trained only on clean embeddings does not have the ability to distinguish noisy embedding from clean embedding well and achieves slightly poor results. Noted that the results in brackets show that the proposed fusion system trained with augmented embeddings can still perform well on clean evaluation set.

Table 2: *Results (EER %) comparison on noisy evaluation set. We use the* ***GATE+AMM+Con-new*** *fusion system here.* ***Train_Clean****: Fusion system trained on clean embedding.* ***Train_Noise****: Fusion system trained with augmented noisy embedding. The result in brackets is tested on clean evaluation set.*

Trial	Voice	Face	ScoreAvg	GATE+AMM+Con-new	
				Train_Clean	Train_Noise
Vox1-O	11.58	9.855	3.962	6.446 (0.585)	**2.500** (0.659)
Vox1-E	10.77	9.995	3.146	5.928 (0.427)	**2.128** (0.513)
Vox1-H	12.68	11.48	4.777	7.355 (0.735)	**3.251** (0.929)

5. Conclusions

In this paper, we explored different multi-modality fusion strategies and loss functions for person verification system, and it can effectively combine the audio and visual information at the embedding level. Based on the strong single-modal system, our best system achieves **0.585%, 0.427% and 0.735% EER** on the three official trial lists of VoxCeleb1, which is, to our knowledge, the best published results on this dataset. Besides, we also introduce an embedding level data augmentation method, which helps the audio-visual multi-modal person verification system perform well when some modality is corrupted or missing.

[3] We would like to thank Suwon Shon for providing the customized trial list of VoxCeleb2 test

6. References

[1] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. 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.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[5] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," *arXiv preprint arXiv:1906.07317*, 2019.

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

[7] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[8] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[9] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[10] S. Horiguchi, N. Kanda, and K. Nagamatsu, "Face-voice matching using cross-modal embeddings," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1011–1019.

[11] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 276–292.

[12] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

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

[14] C. Li and Y. Qian, "Deep audio-visual speech separation with attention mechanism," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7314–7318.

[15] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[16] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[17] C. Li and Y. Qian, "Listen, watch and understand at the cocktail party: Audio-visual-contextual speech separation," *submitted to InterSpeech 2020*.

[18] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proceedings, International Conference on Audio-and Video-Based Person Authentication*. Citeseer, 1999, pp. 176–181.

[19] J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrus, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando, "Audio, video and multimodal person identification in a smart room," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 258–269.

[20] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, and H. Abut, "Multimodal person recognition for human-vehicle interaction," *IEEE MultiMedia*, vol. 13, no. 2, pp. 18–31, 2006.

[21] T. J. Hazen and D. Schultz, "Multi-modal user authentication from video for mobile or variable-environment applications," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[22] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao, "Audiovisual celebrity recognition in unconstrained web videos," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 1977–1980.

[23] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3995–3999.

[24] X. Gong, Z. Chen, Y. Yang, S. Wang, and Y. Qian, "Speaker embedding augmentation with noise distribution matching," *submitted to InterSpeech 2020*.

[25] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[26] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[27] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 118–126.

[28] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] Y. Yang, S. Wang, M. Sun, Y. Qian, and K. Yu, "Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 205–209.

[30] Z. Wu, S. Wang, Y. Qian, and K. Yu, "Data augmentation using variational autoencoder for embedding based speaker verification," *Proc. Interspeech 2019*, pp. 1163–1167, 2019.

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

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

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[35] 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.

[36] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.