NOISE-TOLERANT AUDIO-VISUAL ONLINE PERSON VERIFICATION USING AN ATTENTION-BASED NEURAL NETWORK FUSION

Suwon Shon, Tae-Hyun Oh, James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

{swshon,glass}@mit.edu, taehyun@csail.mit.edu

ABSTRACT

In this paper, we present a multi-modal online person verification system using both speech and visual signals. Inspired by neuroscientific findings on the association of voice and face, we propose an attention-based end-to-end neural network that learns multi-sensory association for the task of person verification. The attention mechanism in our proposed network learns to conditionally select a salient modality between speech and facial representations that provides a balance between complementary inputs. By virtue of this capability, the network is robust to missing or corrupted data from either modality. In the VoxCeleb2 dataset, we show that our method performs favorably against competing multi-modal methods. Even for extreme cases of large corruption or missing data on either modality, our method demonstrates robustness over other unimodal methods.

Index Terms— person verification, recognition, multi-modal, cross-modal, attention, missing data.

1. INTRODUCTION

From cognitive and neuroscience studies on the integration of face and voice signals in humans, it has been observed that the face-voice association is treated differently in human's brain compared to other paired stimuli [1], and that this perceptual integration plays an important role and is actually leveraged for person recognition processing [2]. Inspired by these findings, computational models have been recently introduced to understand whether, and to what extent, such models can leverage associations between different modalities. To investigate this multi-modal association, Nagrani et al. [3], Horiguchi et al. [4] and Kim et al. [5] presented a face-voice cross-modal matching task by learning a shared representation for both modalities. Neural network based cross-modal learning is explored to distill common or complementary information from large-scale paired data. In particular, Kim et al. showed that their computational model has similar behaviors to humans.

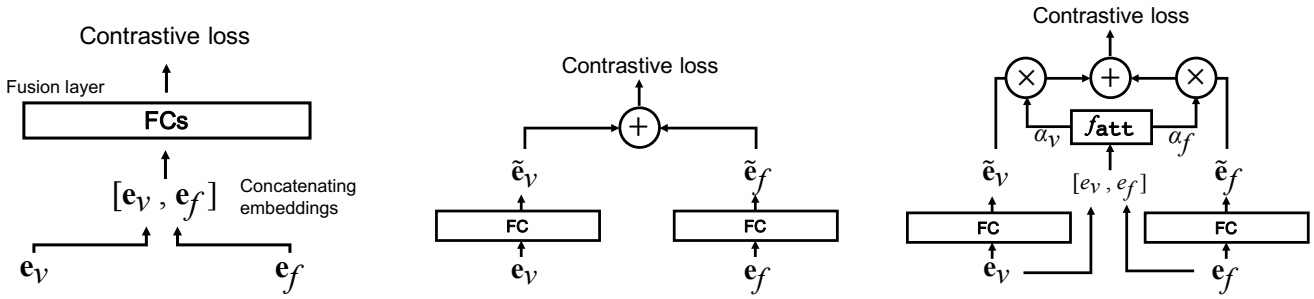
Based on these explorations of multi-modal computational learnability, we investigate the use of multi-modal neural networks for a more specific and challenging task, i.e., person verification. There has been some work that investigates person verification using multi-modal biometric data [6, 7, 8, 9, 10, 11]. These methods typically consist of independent face and voice unimodal recognition modules that are trained separately, followed a score fusion of respective scores from unimodal modules. These methods also typically run in an off-line manner, whereby multiple frames of the face and several seconds of speech are used to maximize recognition performance; thus, there is an inherent latency embeded in the methodologies. On the other hand, feature-level fusion has been uncommon in the person verification. The feature-level fusion has been more commonly adoped in audio-visual speech recognition [12, 13] from a simple concatenation of the feature to end-to-end system [14, 15] with synchronized audio-visual feature. In this work, we shed light on the feature level fusion in the multi-modal person recognition.

In this paper, we explore an online audio-visual fusion system for person verification using face and voice. In contrast to previous work on person verification, our proposed fusion method is conducted at the feature level. In particular, we focus on the fusion of synchronized audio-visual data, under the argument that the system should naturally assess the time-varying contribution of each modality according to its instantaneous quality at any point in time. Our method exploits a single video frame of the face and a short span of speech to facilitate online processing applications, while maintaining high performance against prior state-of-the-art. Motivated by the attention [16] and the multi-sensory association mechanism of human brain [1], our fusion method is implemented by a neural attention mechanism, such that it can learn to evaluate saliency of input modality. Due to the inherent robustness of this architecture, we expect stable performance even when there is corrupted information from either face or voice by noise masking or missing information by pre-processing failures on either modality e.g., face detection, voice activity detection, etc. We experimentally verify that our proposed fusion network is indeed robust to corrupted and missing information from one modality. We also analyze the output of the attention layer to see how it behaves under certain characteristics of input.

2. ONLINE PERSON VERIFICATION FROM VIDEO

The verification of a person's identity (ID) is often achieved by using information from a single modality that contains the biometric signal, such as images for face identification and audio for speaker verification. When multiple modalities are available, such as in video recordings of someone speaking, then opportunities exist to explore the fusion of information from both modalities. Both vision and hearing must address challenges due to variation in a persons appearance or voice, or occlusion caused by environmental conditions. In the case of vision, the image of a person's face will appear differently due to physical changes in a person's appearance, emotional state, occlusion due to other objects, and will depend on position and orientation relative to the camera, etc. Likewise, a person's voice can change due to health, or emotional state, and will be affected by environmental noise, reverberation, and channel conditions.

One interesting difference between face and speaker ID technologies is that high-quality face ID can be obtained from a single image of a person's face. In video data, this corresponds to a single instance in time, and can be sampled many times a second. In contrast, to achieve the same level of performance for speaker verification tasks typically requires a much longer sample of speech from the talker (e.g., 10-30sec of speech are typical conditions,



(a) System $\mathcal{A}$ (b) System $\mathcal{B}$ (c) System $\mathcal{C}$

Fig. 1: Neural network based fusion approaches. $\mathbf{e}_v$: speaker embedding, $\mathbf{e}_f$: face embedding

with a few seconds of speech being a much more challenging task). This discrepancy is because, unlike face images, the speech signal is highly time-varying due to the nature of communication production. A random snippet of speech can be dramatically different from another, even when spoken by the same talker due to differences in the acoustic-phonetic sequences presented in samples. The characteristics of a talker's voice are more reliably extracted when the duration of a speech recording contains more examples of the different sounds produced by the talker. For person verification, there is some truth to the mantra that a picture is worth a thousand words!

When processing video data, there will be situations where one modality or the other may be corrupted or altogether missing. A corrupted modality can be caused by a false alarm of a pre-processing step such as face detection or voice activity detection (VAD). For example, a face detector may incorrectly miss a face, or detect the wrong face or region in the video, or the VAD might be activated by background noise that does not contain a human voice. These corrupted inputs could easily confound a multi-modal network, so that its performance could be worse than fusing separate unimodal systems. When one modality is completely missing, one easiest solution in practice would be to switch to apply an alternative backup unimodal system to the uncontaminated modal data. We demonstrate that our multi-modal system performs favorably against this systematic approach even in the complete missing case.

3. AUDIO-VISUAL MULTI-MODAL FUSION

In this section, we describe the proposed multi-modal fusion approach and its voice and face representation subsystems. Our method is distinguished from previous studies by its use of a feature-level fusion approach based on neural network models. Given discriminative face and speaker representations extracted from each subsystem, our attention layer evaluates each contribution of the respective representations. Then, we combine them according to the estimated contributions, so that joint representation is obtained. We learn this whole fusion network for the person verification task without additional supervision for the attention. In the test phase, we compute the similarity of joint representations between the query (enrollment) and test samples to verify identities.

In the following sections, we elaborate the proposed fusion approach and the speech and face sub-systems used in our experiment.

3.1. Proposed Fusion Approach

We develop a multi-modal attention model that can pay attention to the salient modality of inputs, while producing a powerful fusion representation appropriate for the person verification task. The humans' multi-sensory capability inspires this. Among diverse facets of the human multi-sensory system, the presence of the selective attention [16] allows humans to first pick salient information even from crowded sensory inputs. The human attention mechanism dynamically brings salient features to the forefront as needed without collapsing holistic information into blurry abstraction.

The realization of this attention mechanism in deep neural networks has achieved successes in various machine learning applications. Our attention network is similar to the soft attention [17] which is differentiable. While most previous work applies spatial or temporal attention, e.g., [18], our attention is extended to be attentive across the modality axis. Given face and speaker embeddings, $\mathbf{e}_f$ and $\mathbf{e}_v$, we define the attention score $\hat{a}_{\{f,v\}}$ through the attention layer $f_{\mathtt{att}}(\cdot)$ as

$$\hat{a}_{\{f,v\}} = f_{\mathtt{att}}([\mathbf{e}_f, \mathbf{e}_v]) = \mathbf{W}^\top [\mathbf{e}_f, \mathbf{e}_v] + \mathbf{b}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^m$ are the learnable parameters of the attention layer, m and d denote the number of modality to fuse and the input dimension of the attention layer respectively, and $\mathbf{e}_f$ and $\mathbf{e}_v$ will be discussed in the next subsection. Then, the fused embedding $\mathbf{z}$ is constructed by the weighted sum as

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

where $\tilde{\mathbf{e}}$ denotes the projected embeddings to a co-embedding space compatible with the linear combination. To map $\tilde{\mathbf{e}}_{\{f,v\}}$ from $\mathbf{e}_{\{f,v\}}$, we used a Fully Connected (FC) layer with 600 hidden nodes, i.e. $\tilde{\mathbf{e}} \in \mathbb{R}^{600}$. We do not used non-linearity in the FC layer. We train the attention networks by the contrastive loss on the joint embedding $\mathbf{z} \in \mathbb{R}^{600}$. For each training step, we used 60 positive and negative pairs, a total of 120 pairs for each mini-batch, and all pairs were sampled from the VoxCeleb2 development set.

The proposed attention networks allow us to naturally deal with corruption or missing data from either modality. In our framework, the attention networks spontaneously learn to assess the quality of given multi-modal data implicitly. For example, if the audio signal is largely corrupted by surrounding noise, the attention network would switch off the voice representation path and would only rely on the face representation, and vice versa. In this way, as long as at least one modality provides appropriate information for the task, this model is able to perform the person verification.

Relationship with Other Fusion Methods In the context of the multi-modal person verification, the traditional score-level fusion by

the logistic regression has been investigated up to these days [6, 7, 8, 9, 10, 11]. These score fusion methods do not leverage any large capacity deep neural networks which are capable of dealing with non-trivial fusion strategy. One can come up with a simple extension based on the above approaches, where FC layers are stacked on top of the concatenated speaker and face embeddings, $\mathbf{e}_v$ and $\mathbf{e}_f$, as shown in Figure 1-(a), i.e. System $\mathcal{A}$. We used 2 FC layers with 1,200 and 600 hidden nodes and ReLUs for non-linearities in the first FC layer. This can be regarded as a feature level fusion similar to Nagrani et al. [3]. A downside of this would be the fact that the performance of the system is degraded by corrupted modal data.

Another neural network based fusion can be accomplished as shown in Figure 1-(b). FC layers are stacked on top of respective embeddings, $\mathbf{e}_v$ and $\mathbf{e}_f$, without any nonlinear activation function. This layer simply projects each modality embedding onto a joint audio-visual subspace. Then, the projected embeddings, $\widetilde{\mathbf{e}}_v$ and $\widetilde{\mathbf{e}}_f$, are combined by the summation operation, and used for the contrastive loss function as we did. The summation based ensemble considers both modalities contribute equally, typically yielding a mean representation which can be easily biased with a large contamination [19].

Our method adaptively estimates the weights of each modal embedding to construct a joint representation. Either of weights can be turned off if the embedding would degrade the end performance. This feature is not only robust but also able to deal with missing or a large corruption of the data.

3.2. Voice and Face Representations

To obtain discriminative embeddings for face and voice, $\mathbf{e}_f$ and $\mathbf{e}_v$, we exploit the existing deep neural network based representations.

Voice embedding Voice embeddings generally exploit a large dataset including augmented data with background noise. A voice embedding can be extracted from one of the hidden layers from a neural network trained to classify N speakers in the training dataset. In a previous study, we proposed a frame-level voice embedding to extract robust speaker information by modifying a pretrained DNN structure [20]. For training, the VoxCeleb1 development dataset was used. Details can be found in [20] since we used the same system. Frame-level voice embeddings are extracted every 10ms using a 25ms frame window. Before fusion, a total of 10 and 100 successive voice embeddings are average-pooled across temporal axis to create a voice embedding which spans 115ms and 1015ms, respectively since a single frame-level voice embedding spanning 25ms is too short to extract voice characteristics reliably.

Face embedding Our face embeddings are extracted by using FaceNet [21] pre-trained on CASIA-WebFace.[1] Since the provided face region annotations in the VoxCeleb datasets are coarse, we re-align and crop faces by the face and landmark detectors in Dlib.[2]

4. EXPERIMENTS

In this section, we evaluate the proposed method with various baselines on the evaluation setup described in Sec. 4.1. In Sec. 4.2, we compare our person verification performance with several multimodal fusion approaches as well as unimodal methods in the ordinary scenario that both modal data is given. Then, we demonstrate the robustness of the proposed method against corrupted data in Sec. 4.3. Moreover, we analyze the behavior of the attention layer according to interpretable attributes, including head pose and facial appearance traits, in Sec. 4.4.

	l=0.115 sec		l=1.015 sec	
Systems	EER	mDCF	EER	mDCF
Voice embedding ($\mathbf{e}_v$)	41.27	0.999	14.50	0.863
Face embedding ($\mathbf{e}_f$)	8.03	0.631	8.03	0.631
Score-level fusion	7.83	0.623	5.78	0.491
System $\mathcal{A}$	7.74	0.634	5.52	0.478
System $\mathcal{B}$	7.81	0.625	5.56	0.472
System $\mathcal{C}$ (Proposed)	**7.46**	**0.611**	**5.29**	**0.456**

Table 1: Person verification performance on VoxCeleb2 test set. l is a length of audio segment to extract voice embedding.

4.1. Experimental Environment

In our experiments, we used the VoxCeleb1 & 2 datasets [22, 23], which include multimedia data with a reliable pre-processing step to obtain face regions and voice segments. VoxCeleb1 & 2 have more than 1,281,352 utterances from 7,365 speakers and both datasets have development and test set splits. For verification performance measurement, we made a test trial set using the VoxCeleb2 Test set which contains 36,693 video clips from 120 speakers. We made 300 positive trials (i.e., the same speaker from different clips) and 300 negative trials (i.e., different speaker) trials per speaker, for a total of 71,790 trials.[3]. We used the cosine similarity to measure the similarity of two embeddings.

Voice and face embeddings were extracted in 600 and 512 dimensions respectively. For training the fusion networks ($\mathcal{A}, \mathcal{B}, \mathcal{C}$), we extracted 1 frame per second and its corresponding audio segment with 0.115 and 1.015 secs, respectively. Both embeddings were L2-normalized before feeding into the fusion network. To test, we extract a single frame and its corresponding audio segment pair randomly in each video clip. Thus, a total of 36,693 still images and 0.115 secs (or 1.015 secs) audio segments are used for the test trials. Note that such short speech segments have been barely experimented in speaker verification studies due to its challenging regime. The performance was measured in terms of Equal Error Rate (EER) and minimum Detection Cost Function (mDCF) ($P_{target} = 0.01$) [24].

4.2. Fusion Performance

As shown in Table 1, the voice embedding shows significantly worse performance than the face embedding. This is natural because we only use 0.115 sec, 1.015 sec which is too short segment to extract reliable representations from text-independent speech. The score-level fusion was done by the logistic regression based calibration [25] on the VoxCeleb2 development set. The Systems $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ show neural network-based fusion approaches. While Systems $\mathcal{A}$ and $\mathcal{B}$ show slightly better performance than the score-level fusion on EER, our System $\mathcal{C}$ show a notable gain in both EER and mDCF.

4.3. Effect on Corrupted and Missing Modality

To see the performance under a corrupted or missing modality of either voice and face, we generated random noise drawn from a stan-

[1] https://github.com/davidsandberg/facenet
We used this reproduced open model, which has been improved by the maintainers with several modifications. The modifications include the dimension change of the last layer from 128-D to 512-D. We use the last 512-D FC7 layer activation of this FaceNet version as the face embedding.

[2] http://dlib.net

[3] The number is slightly less than 72,000 because there are a few individuals who have less than five video clips.

	Voice null embeddings				Face null embeddings			
	Random		Zeros		Random		Zeros	
Systems	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF
Score fusion	8.05	0.633	8.03	**0.631**	49.99	0.999	41.27	0.999
System $\mathcal{A}$	8.51	0.712	7.59	0.648	38.81	0.999	35.51	0.999
System $\mathcal{B}$	8.76	0.748	7.51	0.637	37.74	0.999	**34.12**	0.999
System $\mathcal{C}$ (Proposed)	**7.77**	**0.626**	**7.50**	0.633	**37.23**	0.999	34.22	0.999

(a) $l = 0.115$ sec

	Voice null embeddings				Face null embeddings			
	Random		Zeros		Random		Zeros	
Systems	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF
Score fusion	8.19	0.634	8.03	**0.631**	28.18	0.995	14.5	**0.863**
System $\mathcal{A}$	8.64	0.732	7.64	0.649	15.42	0.960	13.27	0.897
System $\mathcal{B}$	8.69	0.724	**7.61**	0.647	16.52	0.970	14.55	0.901
System $\mathcal{C}$ (Proposed)	**7.89**	**0.623**	7.65	0.636	**12.64**	**0.905**	**12.23**	0.871

(b) $l = 1.015$ sec

Table 2: Performance under corrupted and missing modality on either voice and face. l is a length of audio segment to extract voice embedding.

dard normal distribution and a zero vector. Random noise mimics corrupted embeddings obtained from an image without a face or audio without a voice due to the failure of a pre-processing step. The zero vector simulates the missing modality case. This could be handled by switching the multi-modal system to unimodal system. However, we were interested in the scenario where we only used a single universal system and measured the performance when either modality did not exist. In Table 2, the proposed system $\mathcal{C}$ shows better performance in both corrupted and missing modality conditions by assessing the quality of embeddings in the attention layer.

Interestingly, the neural network based fusion systems, particularly the proposed fusion approach, obtain better performance than the one using a unimodal embedding, even for the case that information is only partially available. In the neuroscience study, it has been observed that unimodal perception benefits from the multisensory association of ecologically valid and sensory redundant stimulus pair [26]. As an extension of this observation, we can interpret as the fusion network learns the association of the multisensory data, and it is able to extract robust feature even without multisensory data.

4.4. Analysis of the Attention Layer

We analyze the behavior of the attention layer in our networks. In order to parse what information it has learned and its behavior according to interpretable attributes, we conduct control experiments with facial appearance attributes.

By measuring the probabilities of face/voice attention weights conditioned by an attribute in the test set, we investigate the existence of the statistical correlation between the attribute and the attention, and its tendency. We obtain the attributes of the VoxCeleb2 test set by using the state-of-the-art, Rude et al. [27] and Feng et al. [28] for 40 facial appearance attributes (defined in [29]) and 3D head orientation, respectively. We focus on the relationship between *the behavior of attention weights* and attributes, considering the fact that Kim et al. [5] already showed the connections of face/voice representations with certain demographic attributes.

As a statistical measure, given an attribute A, we measure the expectation of the probability $\mathbb{E}P(\alpha_f > \bar{\alpha}_f | A{=}\text{true})$, where $\bar{\alpha}_f$ denotes the global mean of the face attention over all the test data, and likewise for voice. Since the probability estimate follows the expectation of the Bernoulli trial, we use 95% binomial proportion (Wald) confidence interval. While the attribute estimation methods have low-failure rate profile, due to subtle outlier effects, we conservatively regard 95%-confidence lower bound estimates as significant signals if greater than 60% (greater than random chance).

Head orientation	$\lvert\theta\rvert<30°$		$30°<\lvert\theta\rvert<60°$		$60°<\lvert\theta\rvert$	
	V (%)	F (%)	V (%)	F (%)	V (%)	F (%)
Yaw	43	57	46	54	44	56
Pitch	44	56	41	59	42	58
Roll	44	56	43	57	47	53

(a) Head orientation attributes. (V: voice, F: face)

Facial Attibutes	Voice (%)	Face (%)	95% C.I.
Bald	**74.89**	25.11	$\pm$ 4.02
Blond Hair	32.17	**67.83**	$\pm$ 1.51
Goatee	**70.06**	29.94	$\pm$ 1.38
Mustache	**72.96**	27.04	$\pm$ 1.73
Sideburns	**65.60**	34.40	$\pm$ 1.81
Straight Hair	29.65	**70.35**	$\pm$ 1.09
Wearing Hat	**72.62**	27.38	$\pm$ 2.14

(b) Facial appearance attributes

Table 3: The expectation of $P(\alpha_v > \bar{\alpha}_v | A{=}\text{true})$ and $P(\alpha_f > \bar{\alpha}_f | A{=}\text{true})$, where A denotes attributes. C.I. stands for the (Wald) confidence interval. For head orientation, the front face is represented by all the angle of yaw, pitch and roll equal to 0°.

From Table 3a, we could not find any correlation between head orientation and attention weights. We postulate that the FaceNet embedding is learned to be sufficiently head orientation invariant, so the attention layer turns out to be insensitive to the quality of the embedding according to the orientation. Table 3b shows the 7 attributes of which lower bounds are above 60%. It is interesting that, in the case that a person is with the temporary attributes, such as "Wearing Hat," "Sideburns," "Goatee" and "Mustache," the fusion system is likely to concentrate on the voice with a much higher chance than random. We postulate that temporary factors act as a noise, thus the network relies on the other modality in that case. Also, the strong attributes like "Bald," "Blond hair" and "Straight Hair" show correlation with attention weights.

5. CONCLUSION

Motivated from the recent studies about the multi-modal association, we proposed a feature-level attentive fusion network for audio-visual online person verification task. The temporally synced face image and voice segment assumption encourages the network to learn about the quality of the embedding to verify a person's identity. The learned embeddings of both modalities share a compatible space (co-embedding space) by virtue of the simple linear combination rule to obtain the fused representation. Besides the better performance than the traditional score-level fusion, it has a large advantage to handle severe conditions such as the presence of the corrupted and missing modality. The attention mechanism is also analyzed to understand the correspondence between attention weights and interpretable attributes of visual perception. In addition to visual appearance traits, it would be interesting to further investigate the attention behavior in terms of speech characteristics, such as pitch, language, dialect, as a future direction.

6. REFERENCES

[1] K. Von Kriegstein and A.-L. Giraud, "Implicit multisensory associations influence voice recognition," *PLoS biology*, vol. 4, no. 10, pp. e326, 2006.

[2] B. A. S. Hasan, M. Valdes-Sosa, J. Gross, and P. Belin, "Hearing faces and seeing voices": Amodal coding of person identity in the human brain," *Scientific reports*, vol. 6, 2016.

[3] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] S. Horiguchi, N. Kanda, and K. Nagamatsu, "Face-voice matching using cross-modal embeddings," in *ACM Multimedia Conference*, 2018.

[5] 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 (ACCV)*, 2018.

[6] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal Person Recognition using Unconstrained Audio and Video," in *International Conference on Audio and Video-Based Person Authentication*, 1999, pp. 176–181.

[7] 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*, 2006, pp. 258–269.

[8] W. Thomas and Kie, "Multimodal Person Recognition for Human-vehicle Interaction," *IEEE MultiMedia*, vol. 13, no. 2, pp. 18–31, 2006.

[9] T. Hazen and D. Schultz, "Multi-modal user authentication from video for mobile or variable-environment applications," in *Interspeech*, 2007.

[10] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao, "Audio-visual Celebrity Recognition in Unconstrained Web Videos," in *ICASSP*, 2009, pp. 1977–1980.

[11] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-Visual Person Recognition in Multimedia Data from the IARPA Janus Program," in *ICASSP*, 2018, pp. 3031–3035.

[12] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual Speech Recognition," Tech. Rep., IDIAP, 2000.

[13] J. Kratt, F. Metze, R. Stiefelhagen, and A. Waibel, "Large Vocabulary Audio-Visual Speech Recognition Using the Janus Speech Recognition Toolkit," in *Joint Pattern Recognition Symposium*, 2004, pp. 488–495.

[14] R. Sanabria, F. Metze, and F. De La Torre, "Robust end-to-end deep audiovisual speech recognition," *ArXiv e-prints arXiv:1611.06986*, 2016.

[15] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end Audiovisual Speech Recognition," in *ICASSP*, 2018.

[16] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201, 2002.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[18] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[20] S. Shon, H. Tang, and J. Glass, "Frame-level Speaker Embeddings for Text-independent Speaker Recognition and Analysis of End-to-end Model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[21] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet : A Unified Embedding for Face Recognition and Clustering," in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

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

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[24] "The NIST 2016 Speaker Recognition Evaluation Plan", Available: https://www.nist.gov/document/sre16evalplanv13pdf.

[25] N. Brümmer and D. A. Van Leeuwen, "On calibration of language recognition scores," *IEEE Odyssey 2006: Workshop on Speaker and Language Recognition*, pp. 1–8, 2006.

[26] K. von Kriegstein and A.-L. Giraud, "Implicit multisensory associations influence voice recognition," *PLOS Biology*, vol. 4, no. 10, pp. 1–12, 09 2006.

[27] E. M. Rudd, M. Günther, and T. E. Boult, "MOON: A mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision (ECCV)*. 2016, Springer.

[28] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *European Conference on Computer Vision (ECCV)*, 2018.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision, (ICCV)*, 2015, pp. 3730–3738.