Learning to synthesize faces using voice clips for Cross-Modal biometric matching

Pranav Agarwal*, Soumyajit Poddar†, Anakhi Hazarika‡
Electronics and Communication
IIIT Guwahati
Guwahati, India
*pranav2109@hotmail.com, †poddar18@gmail.com, ‡anakhi22@gmail.com

Hafizur Rahaman
Information Technology
IIEST Shibpur
Howrah, India
rahaman_h@yahoo.co.in

***Abstract*—Cross-Modal biometric matching has been a scarcely explored field but carries several important applications and aims to further secure the currently existing security systems. In this paper, a framework for cross-modal biometric matching is presented, where faces of an individual are generated using his/her voice clips and further the synthesized faces are tested using a face classification network. Generative Adversarial Network (GAN) has become a recent trend in deep learning and has been widely used for image synthesis. We explore the advancements of Convolutional Neural Network (CNN) for feature extraction and generative networks for image synthesis. In the experiment, we compare the performance of Variational Autoencoders(VAE), Conditional Generative Adversarial Networks(C-GAN) and Regularized Conditional Generative Adversarial Networks(RC-GAN) and show that RC-GAN that is C-GAN with a regularization factor added to its loss is able to generate faces corresponding to the true identity of the voice clips with the best accuracy of 84.52% while VAE generates a less noise prone image with the highest PSNR of 28.276 decibels but with an accuracy of 72.61%.**

***Index Terms*—Biometric matching, Deep Learning, Convolutional Neural Network, Generative Adversarial Network**

I. Introduction

Biometric Matching has been a long-standing problem and has been widely explored [1]–[3]. Most of the current existing works are limited to a single domain that is face to face or voice to voice matching and multi-domain [4] that is fusing features from different domains to identify the individual but there has been very few works which have explored the cross-domain approach that is voice to face or vice-versa that is mainly due to the lack of availability of datasets corresponding to the face and voice of the same individual in spite of its huge advantage of further securing the current existing security systems as well as several other applications. The recent development of VoxCeleb [5] and VGGFace [6] datasets have to some extent helped in exploring the field.

The main idea of Cross-Modal biometric matching is to use the voice samples and identify the face of the individual to which that voice clip belongs to or vice-versa. This mapping of voice to face or face to voice is indeed possible and has been proven both biologically [7] as well as neuro-psychologically [8]. Biologically it is well known that hormonal changes effect facial attributes as well as voice pitch during puberty in both male and female. Also the shape of the vocal tract effects voice and movement of facial muscles at the same time. Neuro-psychological experiments have also shown that fusiform face area in the occipital lobe which is responsible for face recognition is directly connected to regions in the temporal lobe which is responsible for voice recognition [9]. This shows a common link between voice and face recognition in the brain. Reference [10] show that because of this connection people tend to remember faces corresponding to a given voice clip and an increase in brain activity and accuracy when matching faces to a given voice clip in comparison to matching voices and names.

The practical applications of this technology are not only limited to security systems but have got a wide range of scope. Phonagnosia [9] is a rare disease where the people are impaired with the ability to recognize familiar voices. The people with this disease lose their ability to recognize or imagine the face of even their family members or famous celebrities when heard over phone or radio. Our application can help the sufferer to counter the disease to some extent. Another application can include further securing the telephonic interviews. In telephonic interviews, the interviewer can easily be fooled if a different person gives the interview in place of the original which can be prevented by this system simply by collecting the voice as well as face samples before the interview.

Deep learning in recent times have achieved state of the art result for many applications and have even surpassed human performance in some tasks. But there have been very few works which have explored the field of cross-modal biometric matching using deep learning methods. Reference [11] was the first to use a deep learning approach to solve the task. They trained a separate convolutional neural network for the spectrograms of the voice clips and the faces and fused the features in the later layers which are then used for classification. Reference [12] learned a shared representation for different modalities which is used to find the correspondence between the modalities by mapping them individually to their common covariates.

In contrast, we propose a novel approach for solving the cross-domain biometric matching using generative networks. To the best of our knowledge, this is the first work to synthesize faces using voice clip embeddings as input to the

generative networks. These embeddings are 128-dimensional and are extracted from a sound classification network trained using spectrograms of each voice clip.

The contribution of our work is summarized as follows:

1) We address the cross-modal biometric matching problem from a view of generative models and develop a novel framework for simultaneously synthesizing faces and checking the accuracy of synthesized faces.

2) We compare the performance of different generative models and show that adding pixel loss as a regularization factor to the original adversarial loss of conditional GANs helps in improving the accuracy of the synthesized images corresponding to the identity of the individual as specified by the voice embeddings.

The rest of the paper is organized as follows. In Sec. 2 we discuss related work and background. In Sec. 3 we present our proposed method. We discuss the dataset, experiments and results in Sec. 4 and the conclusion and future work is drawn in Sec. 5.

II. Related Work

Biometric matching: Computationally knowing the identity of an individual is one of the major concern for security reasons and other aplications and previous works [1]–[3] have successfully been able to achieve this. Face recognition by [13], voice recognition by [14] and fingerprint recognition by [15] are some of the widely popular and successfull approaches for predicting the identity of an individual in a single domain.Reference [16] used a multi-modal approach by fusing facial and fingerprint features for person identification and show significant improvement in accuracy. But there have been very few works [11], [12] which have used a cross domain approach that is given a voice recognising the face to which that voice belongs or vice versa. Our works differs from the previous approaches for cross-modal biometric matching in that we use generative networks for face synthesis using voice embeddings and simultaneously checking the accuracy of the synthesized faces. In comparison to the previous two approach we restrict the test of our trained model to test samples from the same class on which training was done. Our work takes motivation from [17] where lip movement is being synthesized of a target subject using voice clips from a different subject.

A. Background

T-Distributed Stochastic Neighbour Embedding (t-SNE) t-SNE [18] helps in visualizing the higher dimensional data. Analysing the higher dimensional features is useful in checking the quality of the trained network. Most of the features have a complex polynomial relationship which is very difficult to analyze with the help of other dimensionality reduction method such as Principal Component Analysis (PCA) since it is linear in nature on the other hand t-SNE makes use of probability distributions which helps in analyzing the structure of the data much better. In more general sense t-SNE operates by minimizing the Kullback-Leibler divergence of the distribution of similarity of two input features in the higher dimensional space to the corresponding features in the lower dimensional space as shown in (1). We used t-SNE for visualizing the 128 dimensional voice and face embeddings for each voice spectrograms and face samples.

$$KL(P\|Q) = \sum_{i,j} p_{i,j} log \frac{p_{i,j}}{q_{i,j}} \tag{1}$$

Generative Networks Unsupervised generative models like Variational Autoencoder and Generative Adversarial Networks [19] have achieved mind-boggling results in the recent times. The main idea behind their working is to learn the distribution of a given training data and then generate new data corresponding to a given input. This has seen various applications like Generating Images from a given textual description, Image Denoising to name a few.

Variational Autoencoder Variational Autoencoder [20] follows a compression-decompression like an algorithm using an encoder-decoder network. The encoder maps the input data to a latent representation which is made to follow a Gaussian distribution. The decoder then samples data from this distribution to reconstruct the input data. Since the decoder maps the lower dimensional data to higher dimensions there is always a loss associated with this approach. The network is trained to optimize the loss represented as,

$$L(\theta, \phi) = \sum_{i=1}^{\infty} (-E_{z \sim q_\theta(z|x_i)}[log p_\phi(x_i|z)] + KL(q_\theta(z|x_i)\|p(z))) \tag{2}$$

where the encoder is represented as $q_\theta(z|x)$ and decoder as $p_\phi(x|z)$ where θ and ϕ are the parameters of the encoder and decoder network respectively and x is the input data while z is the encoded latent vector.The reconstruction loss is represented by the first term in (2) where the expectation of the datapoint i is taken with respect to the distribution of the encoder which helps the decoder in efficient reconstruction. The KL divergence term in the loss function act as a regularizer and measures the amount of similarity between the distribution of q and p. This regularizer leads to a penalty for the encoder whenever the encoded distribution generated does not follow a standard normal distribution. Thus data from the same class have very less difference in their latent vector representation.

Generative Adversarial Network Generative Adversarial Networks or GANs are another kinds of unsupervised generative networks consisting of two networks competing against each other in order to obtain Nash equilibrium. The most famous example to explain the working is that of counterfeiting money. The generator learns to develop fake currency in order to fool the discriminator while the discriminator trains itself to differentiate between the real and fake ones. Hence both of them are involved in a competition to become more efficient than its counterpart. If we consider probabilistically the generator learns the probability distribution of the real input data in order to fool the discriminator that is $p_{Gen}(x) = p_{real}(x)$. So

the main aim of the GANs is to optimize the function given as

$$L(G,D) = E_{x\sim p_{data}(x)} log(D(x)) + E_{z\sim p_z(z)} log(1 - D(G(z))) \quad (3)$$

where x is the real input data while z is some random noise. The discriminator tries to maximize $L(G,D)$ while the generator tries to minimize it. The ideal point is reached when the generator mimics the exact distribution of the real input data while the generator outputs probability 0.5 for any given input whether it is real or fake.

Conditional Generative Adversarial Networks Conditional GAN [21] is one of the variants of GAN where both the networks are modeled based on a given conditional vector y that is the generator tries to generate data given both noise z and y as a condition while the discriminator differentiates real and fake given the input data and the condition y. The loss function for conditional GANs are given as

$$L(G,D) = E_{x\sim p_{data}(x)} log(D(x,y)) + E_{z\sim p_z(z)} log(1 - D(G(z,y),y)) \quad (4)$$

where the discriminator tries to maximize $L(G,D)$ while the generator tries to minimize it.

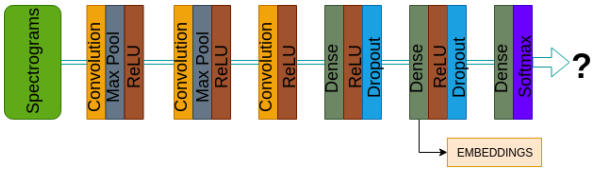

Fig. 1. Sound Classification Network

III. The Proposed Method

In this section we detail our proposed method. In particular we describe each of the subnetworks used which combine together to form the entire framework.

A. Voice Embeddings Generator

We train a voice classification network which takes as input spectrograms of dimension 60*41*2 as described in Section 4-A. This network is 6 layered deep as shown in the Fig. 1 and is trained using Adam Optimizer with a learning rate of 0.001 to optimize the categorical cross entropy loss. Each convolutional layer has 128 kernels with kernel size being restricted to 3*3. The convolutional layers are followed by a 2d Max Pooling of pool size 4*2. The output of the pooling layer is passed through a ReLU activation function.The three fully connected layers after convolutional layers have 512,128 and 9 hidden units respectively. To prevent overfitting 50% dropout and L2 penalty of 0.001 on weights are used in the fully connected layers. Training is done using early stopping. The trained model is saved by removing the last fully connected layer and each of the spectrograms are again passed through the model which outputs a unique 128 dimensional embeddings. These embeddings are visualized using t-SNE as shown in Figure 2 a) which proves the efficiency of these embedings in capturing the identity of each class.

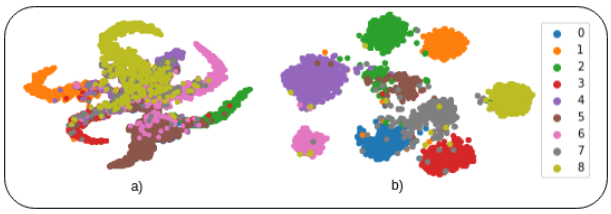

Fig. 2. t-SNE visualization.

B. Face Classification using Embeddings

We use a pre-trained Inception [22] network trained on FaceScrub [23] and CASIA-WebFace [24] to get unique 128-dimensional embeddings for each of the 3,808 faces for the 9 classes, we use for training generative networks from the VGGFace dataset. The input to the network is a min-max normalized 96*96 RGB image and the output is 128-dimensional embeddings for each image. The main aim for extracting these embeddings are that these capture the unique identity information for each image as shown in t-SNE visualization in Figure 2 b) which is useful in calculating the optimal threshold τ for differentiating one class from the another by checking the L2 distance between these embeddings. In generative networks most of the output is a slightly blurred version of the expected output image. So image to image L2 distance between the original image and the synthesized image is avoided as image belonging to the same identity may have large distance while the embeddings focus on facial features even in the presence of noise, hence effective for classification. Further L2 distance of embeddings between faces of same and different classes are computed and a grid-search algorithm is used to find the optimal τ. We obtain the best accuracy of 93.9% for a threshold of 0.77. This same τ is used to classify the synthesized image from the generative networks.

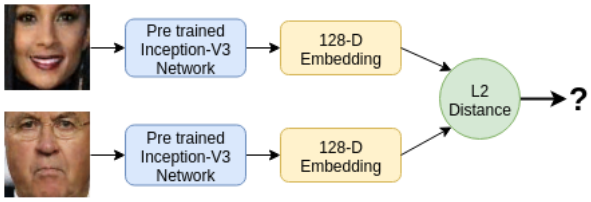

Fig. 3. Face Classification Using Embeddings

C. Voice to Face Decoder

Sound embeddings for each voice clip as described in Section 3- A. is used as a condition for each of the generative network. The training set consists of pairs of sound embeddings and their corresponding faces. Since we are only interested in preserving the identity of the generated face from the voice embeddings we choose a single random image from

each class and pair it with the voice embeddings (F_i, V_i^j), where F_i represents an image of class i and V_i^j is a j^{th} voice embedding of class i , $i \in 0, 1, 2, 3, ..., 8$ is the index for each of the individual in our dataset. The training and network architecture of generative networks are described in detail in Section 4.

Generated faces need to be evaluated both qualitatively and quantitatively. Quantitative estimation is required for checking the accuracy of the trained model while qualitative results finds the best model which creates less noise prone image so that it is interpretable both by humans as well as computational models. 128-dimensional embeddings for the output image of each generative network are obtained using the same approach as described in Section 3-B. Threshold tau of 0.77 which gave the best accuracy is used here as well.

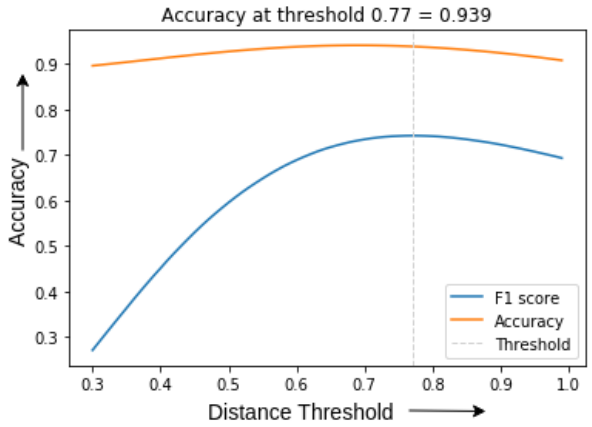

Fig. 4. Accuracy-F1 Score of Face Embeddings for different thresholds

IV. Experiments

In this section, we evaluate the proposed framework for each of the three different generative networks. Firstly, we introduce the VoxCeleb and VGGFace datasets and the pre-processing performed on these datasets. Then we present the implementation and training details and conclude the section with a comprehensive analysis of classification accuracy as well as the quality of the synthesized faces.

A. Dataset and Preprocessing

VoxCeleb [5]. This is an audio-visual dataset and consists of 100,000 utterances for 1,251 individuals. All the audio clips are collected from Youtube videos in a challenging environment conditions. These audio clips are mp3 compressed and are converted to spectrograms as shown in Figures 5 using the same approach as in [25]. Spectrograms have been successfully used for sound classification [26], [27] and is widely used in deep learning for audio classification because of its image like representation which prevents loss of frequency information which is otherwise not possible in the original 1-D time domain representation. The input to our CNNs for voice classification consists of spectrograms of 60 rows(bands), 41 columns(frames) and 2 channels where the 1st channel is the mel-spectrogram and the 2nd channel corresponding deltas. No further normalization is done.

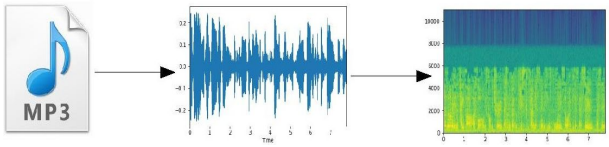

Fig. 5. mp3-wav-spectrogram visualisation

VGGFace [6]. This is a large scale image dataset consisting of around 2.6 million images for 2600 identities. All the images are of random dimension with redundant information other than faces which affect the performance of both the face classification network as well as generative networks. This redundant information is removed as shown in Figures 6 and faces are aligned using [28]. For the face classification network, the aligned images are resized to 96*96 and min-max normalized. The generative network uses a min-max normalized aligned image of size 64*64. Generative networks

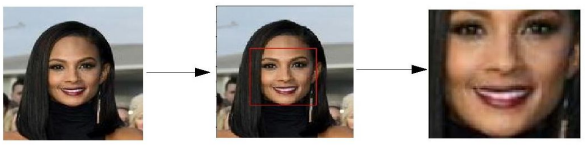

Fig. 6. Cropping Image to remove Redundant information

particularly GANs require huge amount of data for their efficient training. So we restricted the training and testing of our framework to 9 classes which had a minimum of 200 voice samples for each class.

B. Implementation details

We experimented with three variants of the generative network on the above datasets. A variational autoencoder like network where sound embeddings are encoded based on a prior distribution and face are decoded from the sampled data. This approach is different from the original VAE where input and output data belong to the same domain. The advantage of using this approach is that the sound embeddings as extracted in Section 4A are random in nature, hence simply upsampling the embeddings using transposed convolution will not give good results, to overcome this so that the embeddings are close to a distribution they are passed through an encoder which is two layered fully connected layer with 128 hidden neurons followed by a relu activation layer. Thus, given an embedding, the encoder maps it to a latent representation which follows a particular distribution. Further, the decoder learns to generate faces from this latent representation. The decoder is a three-layered network starting with a fully connected layer followed by a two-layer transposed convolution layer as shown in Figure 7. Thus, the latent representation tries to learn the important features from the sound embeddings which is useful in constructing the face of the corresponding individual. The

loss incurred in latent representation and face generation is optimized using Adam Optimizer with a learning rate of 0. 001. The model is trained for 5000 epochs with a batch size of 32.

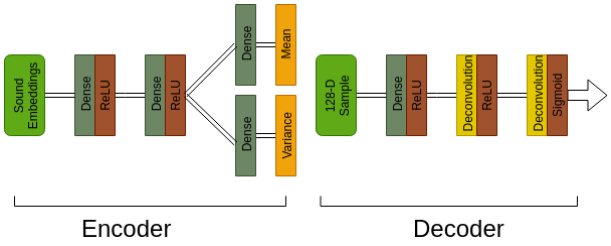

Fig. 7. Encoder and Decoder

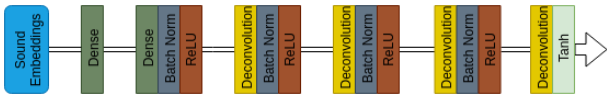

Fig. 8. Generator

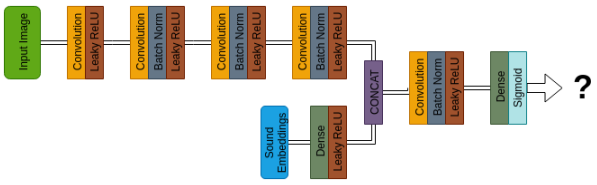

Fig. 9. Discriminator

Previous work have shown the ability of GANs to generate realistic images in comparison to VAEs which suffer from averaging effect. GANs as described above consists of two subnetworks the discriminator and the generator. The input to our generator network for C-GAN is a 128-dimensional sound embedding as a condition. Previous applications have concatenated the conditions with noisy inputs which resulted in images with variations depending upon variation in noise but our aim is to produce faces corresponding to the identity of the voice clip so 128-dimensional sound embeddings are directly inputted to the generator. Further, the embeddings are passed through a two-layered fully connected network followed by a four-layered transpose convolutional layer Figure 8. The discriminator network takes in input the images as well as the sound embeddings. The images are first downsampled using 4 convolutional layers and then concatenated with sound embeddings which are further passed through a fully connected layer and the output is a score suggesting whether the image is a real belonging to that class or not. This way of concatenating the image features with sound embeddings in the later layers is appropriate since the initial 4 convolutional layers learn the high-level features which can then be combined with the condition that is sound embeddings which can be used to make the final decision. Both the Discriminator and Generator networks loss are optimized using Adam Optimizer with a learning rate of 0.0002 and 0.0001 respectively.

The third experiment is done on RC-GAN by varying the loss function of C-GAN. L1 loss between the actual image and the generated image is added to the adversarial loss of the C-GAN. The architecture of the generator and discriminator for RC-GAN remains the same as shown in Figure 8 and Figure 9 respectively.

C. Evaluation Metrics

Being the first work of this kind and due to lack of good datasets we limited our work to VGGFace and VoxCeleb datasets and extensively evaluated the performance of our approach for three different generative networks. We evaluate the working of our trained model by using the test voice embeddings as input to each of the three networks and verifying if the generated face corresponds to the identity as represented by the sound embeddings. For this verification, the generated faces, as well as the actual faces corresponding to each of the sound embeddings is first aligned using facial landmarks resized to 96*96 and min-max normalized and then passed through a pre-trained Inception like network as discussed in the Face Classification section. The output is a 128-dimensional unique embedding. Since the last layer of a neural network generally gives a high-level representation of the input data so our approach of getting these embeddings acts as a unique I.D for each of the individual. L2 distance between the embeddings of the generated face and the actual face corresponding to the identity of the sound embeddings is computed. Threshold value τ of 0.77 is used which gave the best possible accuracy in differentiating one class of face from another as calculated in the Face Classification section. RC-GAN that is a GAN with a regularizer gave the best possible accuracy of 84.52% as shown in the table.

Further, we compared the performance of each of the generative networks based on the quality of the images generated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Signal to Noise Ratio (SNR), Peak Signal to Noise Ratio (PSNR), Perceptual-fidelity Aware Mean Squared Error (PAMSE), Structural Similarity Index Measure (SSIM) as shown in table 2. Comparing the qualitative results Variational Autoencoder like network produce the best quality of image in comparison to the adversarial networks but at the cost of the classification accuracy.

1) RC-GAN vs C-GAN: We compare the performance of RC-GAN with C-GAN based on the same network architecture as described in Section 4.2. RC-GAN outperforms C-GAN both qualitatively as well as quantitatively as shown in Table 1 and 2. Face classification accuracy shows a significant improvement of around 13% for RC-GAN in comparison to C-GAN. Figure 10 shows that RC-GAN generates less noise prone image and much better results which is also proved by the PSNR and MAE value of both the networks. This is because C-GAN even after tuning every hyperparameter outputs a blurred version of the original image. For some instances C-GAN even outputs an average or merged faces of classes having the same colour or gender which is not the case for RC-GAN. This is mainly because RC-GAN penalizes

	VAE	C-GAN	RC-GAN
Accuracy	72.61%	68.09%	84.52%

TABLE I
FACE CLASSIFICATION ACCURACY

	MAE	MSE	SNR	PSNR	PAMSE	SSIM
VAE	0.049	0.011	17.857	28.276	575.408	0.866
C-GAN	0.215	0.075	1.125	11.543	3964.294	0.717
RC-GAN	0.112	0.025	6.221	16.639	1055.896	0.703

TABLE II
IMAGE QUALITY METRICS

the network for both the adversarial loss as well as the pixel loss in comparison to only adversarial loss in case of C-GAN.

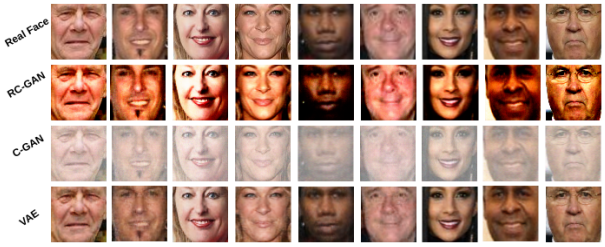

Fig. 10. Qualitative comparison

V. CONCLUSION AND FUTURE WORKS

In this paper, we propose a novel framework for cross-modal biometric matching. We show that sound embeddings effectively captures the fine details of an individual's voice clip and can be used as a condition to synthesize faces for that individual. We compare the performance of our approach on three different generative models and show that adding pixel loss to the existing adversarial loss of GANs preserves the identity of the generated faces hence improving the accuracy of the trained model.

Our framework for cross-modal biometric matching can be used for many different applications, though lack of voice-face pair dataset acts as a hindrance.. We believe if attributes like age, ethnicity is taken into consideration while training the voice to face model will further improve its performance since both these factors have a drastic effect on both voice and facial features. In future our aim is to take into consideration these factors to further develop the current work with advancement in the field and development of efficient datasets.

REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, Jan 2004.

[2] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627–639, April 2007.

[3] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 2002, pp. IV–4072–IV–4075.

[4] A. Ross and A. Jain, "Information fusion in biometrics," vol. 24, 02 2003, pp. 2115–2125.

[5] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[6] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[7] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'putting the face to the voice' matching identity across modality," *Current Biology*, vol. 13, pp. 1709–1714, 2003.

[8] J. J. S. Barton and S. L. Corrow, "Recognizing and identifying people: A neuropsychological review," *Cortex*, vol. 75, pp. 132–150, 2016.

[9] J. C. Hailstone, S. J. Crutch, M. D. Vestergaard, R. D. Patterson, and J. D. Warren, "Progressive associative phonagnosia: A neuropsychological analysis," in *Neuropsychologia*, 2010.

[10] K. von Kriegstein and A.-L. Giraud, "Implicit multisensory associations influence voice recognition," *PLoS Biology*, vol. 4, pp. 1709 – 1714, 2006.

[11] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8427–8436.

[12] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *CoRR*, vol. abs/1807.04836, 2018.

[13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[14] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," vol. abs/1003.4083, 2010.

[15] R. Cappelli, M. Ferrara, and D. Maltoni, "Minutia cylinder-code: A new representation and matching technique for fingerprint recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2128–2141, Dec 2010.

[16] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 450–455, March 2005.

[17] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *ECCV*, 2018.

[18] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[20] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *NIPS*, 2016.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[23] H. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," pp. 343–347, Oct 2014.

[24] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," vol. abs/1411.7923, 2014.

[25] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2015, pp. 1–6.

[26] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," vol. 18, no. 2, Feb 2011, pp. 130–133.

[27] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *NIPS*, 2009.

[28] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.