Voice to Facial Image Construction using Eigenface

Shreyashri Biswas
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
sbiswas2@andrew.cmu.edu

Adeesh Bhargava
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
adeeshb@andrew.cmu.edu

Joseph Bajor
School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213
jbajor@andrew.cmu.edu

Hanzhi Yin
School of Music
Carnegie Mellon University
Pittsburgh, PA 15213
hanzhiy@andrew.cmu.edu

Abstract

We propose a novel approach to generating discernible faces of individuals from voice alone. Previous literature on the topic has focused on creating high fidelity representations of the speaker, but at the cost of computation speed and training difficulty. Oftentimes, the generator component is the most complicated section of such models. We propose an architecture that uses eigenfaces to provide the generative component of a voice to face model. By applying PCA to a large sample of aligned faces images, we obtain a set of eigenvectors that can be combined to create faces of arbitrary identities. Faces can be represented as a vector of eigenvalues which are multiplied by our eigenface vectors to recreate an image. This enables us to create a model that can create faces from voices with greatly improved efficiency while maintaining the ability to generate identifiable features comparable to larger models.

1 Introduction

The goal of voice profiling is to determine a person's biophysical characteristics, such as gender and age, from their speech. This paper explores the idea of reconstructing a person's face from their voice, taking into account the statistical relationship between a person's voice and their facial features. While certain physical and demographic factors can be inferred from a person's voice, it is currently not possible to accurately recreate their entire face from an audio clip but a good approximation can be made.

The task of reconstructing a face from a voice is challenging because it is a cross-modal problem, meaning it involves information from multiple senses (in this case, hearing and sight). This makes it difficult to disambiguate the various factors that affect a person's voice, such as their face, from other factors. Additionally, the relationship between a person's face and their voice is not well understood, so it is unclear which features of the voice encode information about a person's facial features. Furthermore, the information needed to reconstruct a face may be spread across different spoken sounds, so a long recording is typically needed to capture enough information to accurately reconstruct a face.

Though existing architectures manage to create high-fidelity representations of the target faces, they rely on relatively inefficient and difficult-to-train architectures, such as Generative Adversarial Networks (GANs) (1) and large-scale Convolutional Neural Networks (CNNs) (2). The capability to

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

generate high-fidelity faces will lose if the generated features do not accurately represent the speaker. Knowing this, we are interested in exploring approaches that focus on improving the efficiency of these systems while preserving the ability to generate decipherable, distinct faces matching the major characteristics of the original speaker by exploring various encoding mechanisms that work on voices and faces. In our work, we have used simple methods for dimentionality reduction, namely Principal component analyses, to generate a subset of eigenvectors and eigenvalues that can be used to accurately reconstruct the 15,000 faces used to generate the initial set of face images. These eigenvectors are called eigenfaces, and our theory is that at a large enough scale, they can be used to reconstruct arbitrary faces not seen in the original data. We apply this concept to the field of voice to face reconstruction to compare our hyper-efficient approach to state of the art models when it comes to reconstruction identifiability.

2 Related Works

Eigenface (3) is being considered as one of most prominent early works in the face recognition space. Using principle component analysis (PCA), this research shows that eigenface reduce the dimension of the dataset. Larger number of faces can be represented by reconstructing them from a few eigen faces. This reduces the computations drastically compared to other approaches for face recognition. Most existing approaches can be broken down into two major components, as with most cross-modal conversion tasks. The first is audio encoding into a low-dimensional intermediate representation, from which a decoder is used to create the final face. Additional components, such as a discriminator network, may be necessary to construct the loss depending on the decoder architecture. The second component, the decoding task, is where existing systems spend most of their time computing. Currently, architectures like GANs dominate the space regarding high-fidelity synthesis (1) (4). Lower fidelity systems that emphasize distinctive facial feature generation vs. photo-realistic reconstruction have used the penultimate layer of pretrained face detection CNNs to define a facial feature vector from which to generate the final face representation. Wen et. at (1) proposed a computational frame based GANs where the network learns to generate faces from voices by matching the identities of generated faces to those of the speakers. However, simpler methods of reconstructing faces from a low-dimensional feature vector exist. One common approach is eigenfaces (5), which are facial vectors derived from a lower dimensional set of feature vectors obtained using principal component analysis (PCA) on a high-dimensional face dataset.

While originally used for simplifying facial recognition tasks (5), the resulting eigenface vectors can be used to reconstruct most faces from the original dataset and produce arbitrary faces if the dataset and the resulting eigenface vectors are large enough. Since eigenface representations usually have a lower dimension than common existing systems, generating faces can become a trivial task that does not require tedious computation. For implementing an eigenfaces based solution, the primary and the most trivial requirement is for a huge dataset of faces. Several researches have shown that Eigenfaces is a robust PCA algorithm for face recognition but requires a lot of data. (6) (7) (8) (9)

In order for a robust face recognition algorithm using eigenfaces, it's imperative to have a large volume of face data. The VGG face data (10) presents a publicly available very large scale database of 2.6 million images for over 2600 people. The dataset specifically contains the labeled dataset of 2622 celebrity and at least 1000 images per identity.

The second part of the problem statement describes the creation of faces from voice inputs. Hence it's imperative to make the model correlate the faces with the respective voice input. This requires audio-visual data which can be used for creating the voice embedding. We are using VoxCeleb1 (11) and VoxCeleb2 (12) datasets for this purpose. The VoxCeleb1 has 100,000+ utterances for 1,251 celebrities which were extracted from Youtube Videos and The VoxCeleb2 has 1 million+ utterances of 6,112 celebrities which were extracted from Youtube videos.

3 The Model Structure

Our proposed model is illustrated in Figure 1, which consists three parts: a voice embedder that generates voice embeddings, a voice-to-face converter, which is simply an multi-layer perceptron (MLP) that maps a voice embedding to an eigenface, and an eigenface reconstructor that takes an eigenface to a facial image.

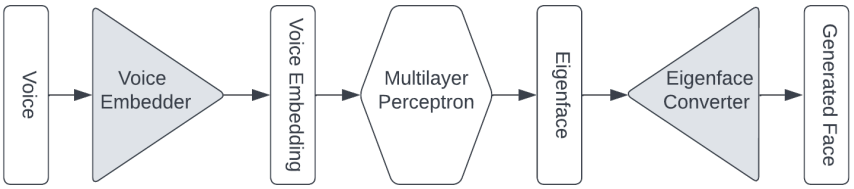

Figure 1: *Proposed Network Structure. The model firstly takes a voice utterance and transform it to a voice embedding vector. Then, the voice-to-face converter converts the voice embedding to an eigenface. Finally, the eigenface reconstructor converts the eigenface back to a face image. Both the voice embedder and the eigenface reconstructor should be trained independently, and the eigenface reconstructor is purely algorithmic. When training the MLP, the voice embedder and the eigenface converter do not participate in backpropagation.*

The voice embedder should only capture salient information of an utterance from a person. It should be trained independently without any prior knowledge of other model components, and it should only be responsible to extract information from voice. The voice embedder model structure and parameters are borrowed from (1), which is a very light convolutional neural network (CNN).

The eigenface reconstructor should only convert an eigenface to an actual facial image and vice versa. Same to the voice embedder, it is trained independently without any prior knowledge of other model components, and it should only be responsible to face conversion. Since eigenface is essentially principle component analysis (PCA), the eigenface converter is merely a linear transformer, not even affine.

The voice-to-face converter should convert a voice embedding to an eigenface. It is also trained independently. However, training the voice-to-face converter requires the existence of the voice embedder and the eigenface reconstructor. We compute the mean-square error (MSE) on the generated eigenface, which can be transformed from actual face images if the eigenface converter is complete. When training the MLP, only the MLP participates in backpropagation. The voice embedder and the eigenface converter do not participate.

4 Experiments

We use the `VoxCeleb` (13) and `VGGFace` (14) dataset pre-processed by Wen, et al. (1). There are 1251 identities in total, with 1225 identities have both voice clips and face images. For each identities, there are multiple voice clips, and there are multiple face images. All images in the pre-processed dataset are with dimension 128 times 128, and almost all faces are front-faced. Throughout the training, all images are converted from RGB to grayscale, meaning that our model can only produce grayscale images. All voice clips have been converted to mel-spectrograms with 64 features.

For the eigenface converter, we randomly selected fifteen thousands images. We choose the first five thousand principle components.

For the voice-to-face converter, we split those identities with both voice data and image data to form train/validation/test split. The identity-ratio is 8:1:1. The MLP has six fully-connected linear layers, with each to have hidden-width 5120, which is close to the number of principle components we selected. In between each linear layers, one-dimensional `Batchnorm` (15) and Gaussian Error Linear Units (16) are interpolated. The loss is computed by mean-square error (MSE) loss on eigenfaces. For each generated eigenface, we compute MSEs between the generated eigenface and all generated identity's eigenface. Then, we take the mean. We used `AdamW` (17) optimizer with betas to be $(0.9, 0.999)$ and the weight decay to be 0.01. The learning rate is 10^{-3}.

The eigenface converter used `scikit-cuda` (18) to compute the principle components. All other components are implemented using `PyTorch` .

5 Results and Analysis

5.1 Subjective Evaluation

In our earlier experiments with eigenface reconstructions, we have already established that it would be possible to create a distinguishable reconstruction of an out of sample face given a large enough set of eigenfaces generated from a diverse set of images.

Our model is able to make use of this concept and can generate distinct faces given voice input from different unseen voices. Though the output will always look like a face, we did notice various modes of failure during our experiments.

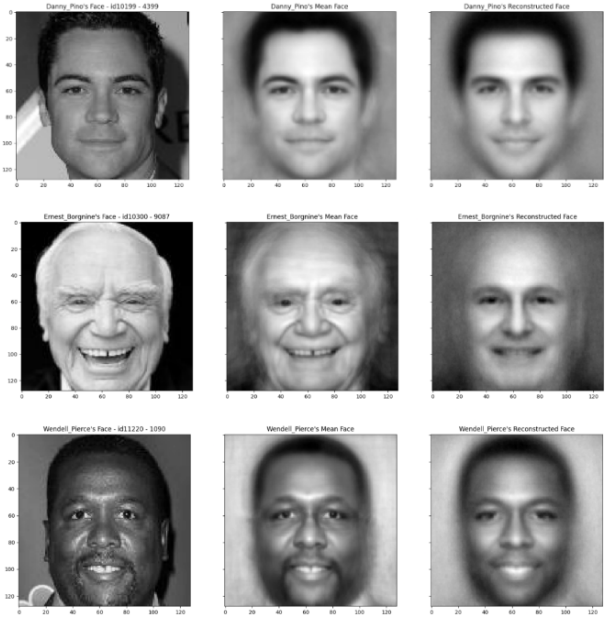

Figure 2: *Successful examples of reconstructed faces. Most distinctive features such as age, gender, and race are properly reconstructed*

Starting with what works, our best-case-scenario results show that the model is sometimes able to create reconstructions that properly convey many of the most distinctive features of each identity, such as gender, race, and to some degree age. Features that have little relevance to voice, such as accessories and facial hair, do not make it across in the model. This is an expected result that falls in line with other models of similar nature.

Looking at some of the cherry-picked best case reconstructions, we can still see some evidence of the model's drawbacks, such as it's inability to create representations that closely match the facial structure of the individual. Though it gets many of the high level features right, things like eye spacing and nose size can still vary (as can be seen in Wendell Pierce's reconstruction).

We also see that sometimes the model can go beyond just basic features and can reconstruct some more distinctive elements of certain people. In 3, we can see that the model has somewhat been able to reconstruct the distinctive ear shape of Aamir Khan, though the face itself has become slightly generic in the reconstruction.

In many outputs, however, the points of failure for our model are on clear display. Easily the most common mode of failure is the tendency for the model to reconstruct highly generic approximations that usually are only distinguishable as male or female. This makes sense, as that is usually the most obvious differentiator when it comes to voice. Many outputs will fall into one of a few generic male or female reconstructions, which can be mistaken for the individual when viewed side-by-side but are difficult to identify when viewed alone.

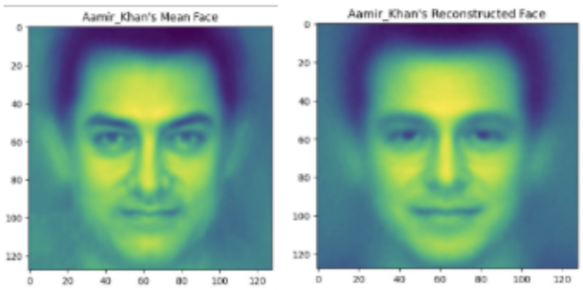

Figure 3: *The model can sometimes come close to visualizing very distinctive features of certain individuals, such as Aamir Khan's ears*

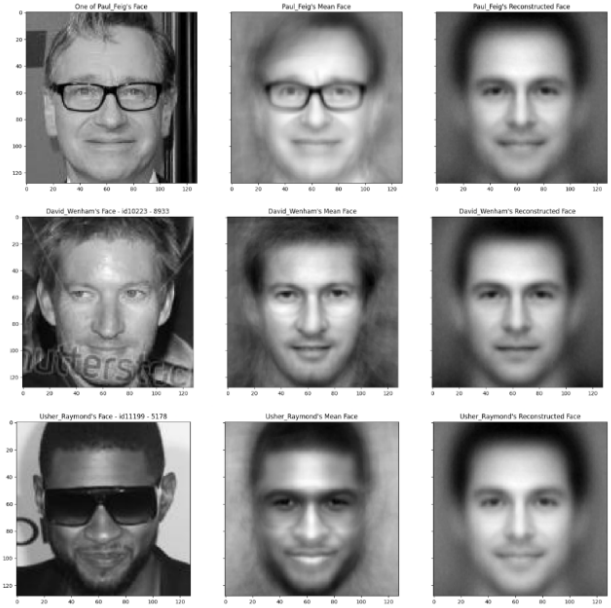

Figure 4: *The model's most common mode of failure is producing an overly generic reconstruction that may miss all of the important features seen in the better reconstructions*

5.2 Objective Evaluation

We pass generated face images (not the eigenface) and face images in the dataset to a trained `VGGFace 2` classifier to get face embeddings. We compare embeddings' cosine similarities to evaluate how closely reconstructed faces and face images in the dataset are related. The trained VGGFace 2 classifier is selected from `https://github.com/cydonia999/VGGFace2-pytorch` with architecture type `resnet50_scratch`, which is a trained-from-scratch model using the ResNet 50 structure (19).

We firstly take a reconstructed face, calculate cosine similarities with each face images of the same identity, and take the mean. Then, for each identities' cosine similarities, we take the mean. Finally, for all identities' cosine similarities, we take the mean.

We compare our reconstructed face with Wen et al.'s GAN model (1). The result is displayed in Table 1.

Ours	Wen, et al.'s
0.5907	0.4148

Table 1: *Mean cosine similarities of the generated face from us and from Yan, et al.*

6 Conclusion

We presented a novel voice-to-face construction model using eigenface.

Though there are some clear ways in which the reconstructions fall short of an ideal outcome, our approach obtains results that are oftentimes comparable or better than existing implementations, even ones of greater complexity. This is a testament to the effectiveness of our approach and shows that using eigenfaces as a sole basis for facial reconstruction can provide some promising results. More research is needed to pin down the reasoning behind the disparity in performance between certain identities, but as a first generation model we consider the results to be promising and worthy of further research. The simplicity of the model and architecture means that there is plenty of flexibility to improve performance through increased model complexity. However, we feel that the approach is benefited by it's current simplicity as it allows for implementation on edge hardware and more flexibility in how it may be deployed, something that cannot be said for more complicated GAN based or VaE based architectures. Potentially, further improvements to performance can be made in the way of a reduced MLP and improved voice encoder to enable the system to perform real time inference.

References

[1] Yandong Wen, Rita Singh, and Bhiksha Raj, "Reconstructing faces from voices," May 2019, arXiv:1905.10604 [cs, eess].

[2] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik, "Speech2Face: Learning the Face Behind a Voice," 2019, pp. 7539–7548.

[3] Matthew A Turk and Alex P Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*. IEEE Computer Society, 1991, pp. 586–587.

[4] Amanda Cardoso Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto, "WAV2PIX: Speech-conditioned Face Generation using Generative Adversarial Networks.," in *ICASSP*, 2019, pp. 8633–8637.

[5] M.-H. Yang, N. Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces," in *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, Sept. 2000, vol. 1, pp. 37–40 vol.1, ISSN: 1522-4880.

[6] Matthew Johnson and Andreas Savakis, "Fast l 1-eigenfaces for robust face recogntion," in *2014 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. IEEE, 2014, pp. 1–5.

[7] Giji George, Rainu Boben, B Radhakrishnan, and L Padma Suresh, "Face recognition on surgically altered faces using principal component analysis," in *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. IEEE, 2017, pp. 1–6.

[8] Mayank Agarwal, Himanshu Agrawal, Nikunj Jain, and Manish Kumar, "Face recognition using principle component analysis, eigenface and neural network," in *2010 International conference on signal acquisition and processing*. IEEE, 2010, pp. 310–314.

[9] Ega Bima Putranto, Poldo Andreas Situmorang, and Abba Suganda Girsang, "Face recognition using eigenface with naive bayes," in *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*. IEEE, 2016, pp. 1–4.

[10] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

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

[12] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[13] Arsha Nagrani, Samuel Albanie, and Andrew 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.

[14] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[16] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," 2016.

[17] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," 2017.

[18] Lev E. Givon, Thomas Unterthiner, N. Benjamin Erichson, David Wei Chiang, Eric Larson, Luke Pfister, Sander Dieleman, Gregory R. Lee, Stefan van der Walt, Bryant Menn, Teodor Mihai Moldovan, Frédéric Bastien, Xing Shi, Jan Schlüter, Brian Thomas, Chris Capdevila, Alex Rubinsteyn, Michael M. Forbes, Jacob Frelinger, Tim Klein, Bruce Merry, Nate Merill, Lars Pastewka, Li Yong Liu, S. Clarkson, Michael Rader, Steve Taylor, Arnaud Bergeron, Nikul H. Ukani, Feng Wang, Wing-Kit Lee, and Yiyin Zhou, "scikit-cuda 0.5.3: a Python interface to GPU-powered libraries," May 2019, `http://dx.doi.org/10.5281/zenodo.3229433`.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.