ORIGAMI – Oration to Physiognomy

Mansi Pandya[a], Palak Chavan[b], Richa Sheth[c], Pratik Kanani[d]

[a] Dwarkadas J Sanghvi College of Engineering, Mumbai, India
[b] Dwarkadas J Sanghvi College of Engineering, Mumbai, India
[c] Dwarkadas J Sanghvi College of Engineering, Mumbai, India
[d] Dwarkadas J Sanghvi College of Engineering, Mumbai, India
[a] mansipandyaa29@gmail.com, [b]palakchauhan381@gmail.com, [c]richasheth46@gmail.com, [d]pratikkanani123@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: More than 400,000 people die from homicide and other crimes each year. Despite the law workers doing everything they can these numbers keep rising. In order to provide help to deal with this situation, this paper explains how we can use deep learning to make correlations between faces and sound they produce which can, in turn, be used to track down criminals. Here, we analyze and comprehend how exactly a person's face is created from a short audio clip of his/her voice. A deep neural network is trained and devised using a dataset of people speaking. Various physical features of the speaker like age, gender and race are captured during training by learning voice-face correlation in a self-supervised manner without explicit modelling of these features. Variations in numeric values between reconstructions from audio and original images are compared and evaluated to train the model to determine how the model works. The canonical face of a person is created from the audio clip. The reconstructed image will not be like the true image but will have the most prominent features of the true image.

Index Terms: Artificial Intelligence, Deep Learning, Convolution Neural Network, Speech-to-Face, Spectrogram, 4096-D face feature

1. Introduction

The strong bond between the speech and appearance of a person can be proved by the fact that humans create a mental image of a people they haven't met but only heard, like over a phone call. This is because physical features play a very major role in how a person speaks and sounds. Other factors like - accent, speed, pronunciations, language, which are shared among various nationalities and culture - also help in determining who the speaker is. The reconstructed images from the audio are front-facing, neutral or in a canonical form. The model proposed tries to capture the most dominant features of the speaker and not every single feature. A neural network model is designed to perform the task of converting a complex spectrogram of audio into a feature vector representing a face. This represented face is actually a 4096-D face feature taken from the second last classification layer of a pre-trained face recognition network. This predicted face is then converted to a canonical face by with the help of a separately trained reconstruction mode. The method used does not require additional information and uses natural simultaneous occurrence of speech and faces to train itself in a self-supervised manner. This field of research about making predictions isn't new, there have been numerous papers which explore predicting age and gender from an audio clip of a person. There have also been multiple attempts to predict faces from audio, but these aren't as robust because of the way the model is designed, leading to a lot of limitations. The methods suggested in these models were that attributes would be predicted based on an audio clip. Based on these predicted attributes either an image would be created or an already existing image with all the attributes would be returned. This approach was the limitation because to perform the task mentioned above a robust and accurate classifier would be required to capture each specific attribute and also curbed the predicted face to mirror a prior definitive set of attributes. On the contrary, this paper focuses on

generating a full face from an audio speech rather than predicting attributes and then combining them and is actually the first paper to do so. The model is tested on multiple speakers and numerical evaluation is done based on how well it can capture a true face solely on audio, to what extent it agrees to the reconstructed face in terms of age, sex, race etc.

2. Literature Survey

A. Disjoint Mapping Network (DIM Net)

The system proposed by Yandong Wen et al. [1] demonstrates the use of a Disjoint Mapping Network (DIM Net) which learned common embedding between voices and faces which were separately applied with covariates like gender, race, ID etc. to obtain supervision, instead of using supervision provided through correspondence between speech and face data. Identically dimensioned features for data from each modality were learned from the learning modules and classifier for predicting covariates from the learned feature were a part of DIMNet. During learning, separate data from each modality was presented but the feature representations were forced to be

comparable in unified classifier forces. Once trained, the classifier could be detached, and the learned feature 6 representations would be used to compare data across modalities. After testing the model, they found out that ID is the strongest supervision factor. In short, they showed how DIMNets frameworks were used to perform cross modelling where individual supervisions were used to learn common embedding for voice and face from multiple covariates rather than the current approach of mapping voices to faces directly. Multiple kinds of label information were also made use of with the help of the DIMNets framework.

B. Conditional GAN and cross-modal

The system proposed by Amanda Duarte et al. [2] implements cross-modal visual generation for construction of the speakers faces from their audio signal. To perform this task, they proposed the use of a conditional generative adversarial model. This model required audio clips with no

background noise or other disturbances so instead of using common datasets like Lip Reading in the wild or VoxCeleb, they made their own dataset. This dataset was made from videos of youtubers because youtubers generally use high quality mics for their videos which helps the model capture the expressiveness in both face and sound. Their main contributions to this paper were presenting a conditional GAN which made use of the concept of Generators and Discriminators. The Discriminators were first trained with the dataset to map the voices to their respective faces. Once trained the discriminator was disabled. An image would be passed through the generator and the discriminator would identify it as real or fake and in this way the Generator would be trained to produce very closely related images of the speaker.

C. Generative Antagonistic Systems

The system proposed by Van Leeuwen et al. [3] worked on finding a solution to the task of remaking somebody's face from their model as it presented an issue. The undertaking is structured to respond to the inquiry: given a sound clasp spoken by an inconspicuous individual, can we picture a face that has the same number of basic components, or relationship as conceivable with the speaker, as far as personality?

As a solution to this problem, they suggest a basic yet successful Mathematical casing work dependent on generative antagonistic systems i.e., GANs. Their system figures out how to produce faces from voices by coordinating the personalities of appearances created to those of the speakers, on a preparation set. They assess the presentation of the system by utilizing a firmly related assignment - cross-modular coordinating. Their outcomes show that their model can produce faces that coordinate a few biometric attributes of the speaker, and results in coordinating correctness that are obviously superior to risk.

So generally, commitment made was: They present another undertaking of creating face images from voice audio in voice profiling which can be utilized in investigating the connection among face and voice modalities.

Additionally, a straightforward yet viable structure dependent on generative ill-disposed systems for this assignment. Every part in the structure is all around inspired. They proposed to collectively assess the faces created by utilizing the cross-modular coordinating assignment. The subjective as well as the quantitative outcomes show that their structure is capable of creating faces that have character relationships with the information voice.

D. Conditional Generative Adversarial Networks utilizing auxiliary classifier

The system proposed by Chua-Hung et al. [4] employed methods like phantom standard, projection discriminator and assistant classifier, contrasted and guileless restrictive GAN using which the model creates pictures with better quality as far as both emotional and goal evaluations. The innovation they use to gain proficiency with a sound to-picture generator depends on GAN.

a. Generator

The highlights can be spectrograms, fbanks, and

mel-recurrence cepstral coefficients, and the secret layer yields of the pre-prepared SoundNet model. The yield of the

generator is a picture produced dependent on the info condition.

b. Discriminator

The convolution layers accept a picture as input and yield a scalar addressing the nature of the picture. The last yield of the discriminator is the expansion of the comparability score and the scalar that exclusively comes from convolution layers. The auxiliary classifier imparts loads to the convolution layers in the discriminator, and they are mutually educated.

The generator and the discriminator are prepared iteratively. That is, the generator is fixed, and the discriminator is refreshed a few times to limit LD (loss function of the discriminator). At that point, we fix the discriminator and update the boundaries of the generator likewise a few times to limit LG (loss function of the generator).

c. Cross modal modelling

Arsha Nagrani et al. [5] proposed that the constrained coordinating with assignment can be characterized as follows; let x = {v, f1, f2} signify a set comprising of an anchor voice portion v and two face pictures f1and f2. Each input set x contains one certain and one negative face, where face fi is characterized as sure on the off chance that it has a similar way of life as the anchor voice, and negative in any case. Given pictures and voices of known personality, we can develop a dataset of training model D = {xn, yn}n=1N by basically randomizing the situation of the positive face in each face pair. The learning issues related to maximising likelihood: θ = argmaxθ L (gθ; D), where gθis the parameterised model to be learned. The loss to be limited would then be able to be outlined as a cross-entropy loss on track name positions. We instantiate gθ as a three-stream convolutional neural organization, taking motivation from the oddball network design. The model plan comprises three methodology explicit sub-organizations; two parameter sharing face sub-networks that ingest picture information and a voice sub-network that ingests spectrograms. The three streams are then joined through a combination layer (by means of highlight link) and taken care of into methodology shared completely associated layers on top. The combination layer is needed to empower the organization to build up a correspondence among countenances and voices. The model henceforth has two sorts of layers, methodology explicit (face and voice) layers and more elevated level layers which are divided among the two modalities. The reasoning behind this design is to compel early layers to work in methodology explicit highlights, (for example, edges in face pictures and ghostly examples in sound fragments) while permitting later layers to catch more significant level idle cross-modular factors (like sex, age, nationality and character). The three principal assignments tackled: 1) Static coordinating, which utilizes just actually face pictures, 2) Dynamic coordinating, which includes recordings of countenances during the discourse, and 3) N-way arrangement, which is an augmentation of the coordinating with errand to quite a few appearances (more prominent than two).

Matching of forward, backward, and sine wave synthesis The system proposed by Miyuki Kamachi et al. [6] demonstrates that identical data in respect to character is accessible hybrid-modularly coming out of the frontage and modulation. Utilizing a deferred coordinating to test project,

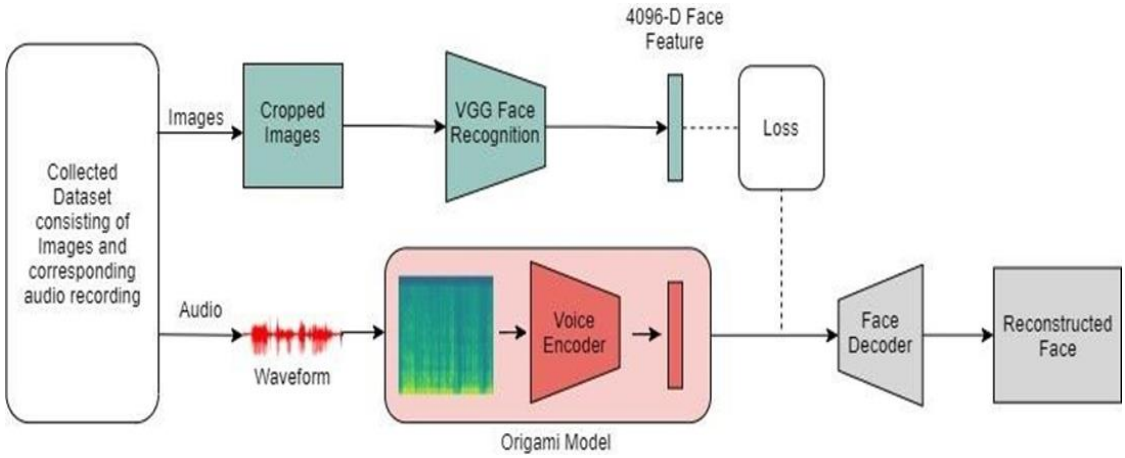

Fig 1: Architecture

AXY, it is displayed that individuals can coordinate with the program of a new frontal A, to a new voice, X or Y, and the other way around, however just when boosts are advancing and are encouraged. The basic job of pace changing data is featured through the capacity to coordinate appearances to modulate consisting of just the granular structural and transient data given via sine convolution discourse. The auditory and visual data obtained from normal discourse are firmly connected, and this connection bears the cost of recuperation of speaker-explicit data across an adjustment in methodology. The over-simplification of the impact across various sentences shows that character coordinating doesn't need redundancy of one or the other construction or substance, however expanded cover may give extra signs. The data utilized is time-fluctuating and bearing ward, proposing that it is intently attached to the elements of characteristic discourse creation. Hence, this data crosses modality as well as overcomes any issues among insight and creation. Forward, backward, and sine wave synthesis - the three exploratory conditions announced feature the significance of generally speaking, bearing ward spatiotemporal

design for the undertaking. On the whole, the utilization of deferred coordinating to an example task, XAB, with a change in tangible methodology from hear-able to visual or the other way around was made between the first and second stages. A face (or voice), X, was introduced in the main stage, and afterward the onlooker was given two voices (or faces) in the subsequent stage. The assignment was to pick which of the improvements in the subsequent stage related to that introduced in the main stage. Hence, this data crosses modalities as well as overcomes any issues among insight and creation.

3. Methodology

The model requires an individual's audio sound with his/her respective faces mapped. We trained our model on an Indian dataset. Instead of using a common dataset like YouTube videos, we've curated our dataset through a google form, in which approximately 100 samples have been collected. We captured the naturalness of every individual's sound and face. Indian samples collated spoke different languages and communicated every dialect of English spoken in the non-native English language which would be accented. Thus, for daily communication, every non-native English individual would speak English with a shadow of their native language over it. The purpose of collecting such diverse information was to capture the particular way in which an individual engages him/herself in the language spoken, expression communicated and the speaking rate. The system

architecture consists of 4 main modules, these modules are executed in a sequential order and are dependent on each other. Fig 1 represents the proposed architecture

a. Pre-processing

The image and audio files collected were pre-processed in different ways. The collected images were run through a code and cropped to capture the most important parts of the face to be fed into a pretrained VGG Face model network. The collected audio files consists of audio clips with different formats like aac, mp3, mp4 etc. which are converted to a wav format to generate corresponding spectrograms.

b. Face Recognition

The Face Recognition module deals with generating 4096-D Face Features from the cropped images. 4096-D Face Features are basically feature vectors that are extracted from the fc7 or the penultimate layer of a pretrained VGG Face Model and contain activations of the hidden layer immediately before the classifier.

c. Training the Voice Encoder

The Voice encoder is a Convolutional Neural Network Model that is trained on the generated spectrograms to produce the 4096-D vector similar to those in the Face Recognition model. The architecture for this CNN Model can be seen in the Table I. The Voice Encoder is trained by minimizing the loss function between the two 4096-D Vectors in such a way that the vectors generated based on the spectrograms of the audio recording are as close as possible to vectors produced by the VGG Face models of the corresponding face image.

d. Autoencoder Model

This project assists with reproducing facial pictures from a short audio clip by utilizing an unsupervised deep learning method or rather a self-supervised deep learning procedure of auto encoders. The auto encoder takes data sources and makes them go through hidden layers that should give a yield like the information. Subsequently, the entire point is to get indistinguishable outcomes fed into input from the yield. There are two segments of an auto encoder for example encoder and decoder. The encoder has the task to compress

Layer	Input	CONV RELU BN	CONV RELU BN	CONV RELU BN	MAX POOL	CONV RELU BN	MAX POOL	CONV RELU BN	MAX POOL	CONV RELU BN	MAX POOL	CONV RELU BN	CONV RELU BN	CONV	AVGPOOL RELU BN	FC RELU	FC
Channels	2	64	64	128	-	128	-	128	-	256	-	512	512	512	-	4096	4096
Stride	-	1	1	1	2x1	1	2x1	1	2x1	1	2x1	1	2	2	1	1	1
Kernel Size	-	4x4	4x4	4x4	2x1	4x4	2x1	4x4	2x1	4x4	2x1	4x4	4x4	4x4	∞ x 1	1x1	1x1

Table I: Voice encoder architecture

the information into more modest encodings. While decoders figure out how to fabricate yield, equivalent to the input utilizing the encodings.

We first load the pictures to prepare the autoencoder. At that point, we encode a picture as a little feature vector with fixed weights. Here, we show the interaction of information compression and the recreation of the encoded information by first structuring an auto-encoder utilizing Keras and afterwards reproducing the encoded information and imagining the remaking. First, every one of the utility capacities is characterized which are required at various steps of the structure. The auto-encoder is characterized and afterwards, each function is called likewise.

For our encoder, we have utilised a famous dimensionality decrease strategy of the Principal Component Analysis. In the engine, PCA endeavours to decay object-feature matrix X into two more modest networks: W and Wminimizing mean squared error: ||(XW) W - X ||22 → w . W

Encoder: X -> Dense(d units) -> code Decoder: code -> Dense(m units) -> X

Where Dense is a fully-connected layer with linear activation: f(X) = W . X + b

Further, we also flatten and unflatten data to be compatible with image shapes. Lastly, we mend everything together in one model.

1. Face encoder

Our encoder takes an input picture and returns a dimensional element vector. We need to pick the encoder cautiously so that is robust to shifts in the areas of pictures. Accordingly, we utilize a pre-prepared model which doesn't refresh its boundaries and standardizes away variety in face pictures that isn't characteristic of the personality of the subject. In this way, the embeddings of the controlled preparing pictures get planned to a similar space. This permits us to just train on the controlled pictures. When utilizing VGG-Face highlights, we utilize the 4096-D "fc7" layer.

2. Face decoder

We use the face decoder model to recreate a certified face image. We tend to train this model using comparative face features isolated from the VGG-Face model as a contribution to the face decoder. A primarily hindered face and its fractional structure are what's created by our face decoder.

We tend to show the impact of face translating on the presentation of a deep face recognition pipeline with hindrances. Our assessment intends to quantitatively assess

the impact of various checks on recognition and (b) what quantity face information is rehabilitated through face unravelling. The principle objective of the face decoder is to repeat the image of a face from a low-dimensional face. we decide to issue any spare assortments (present, lighting, etc), whereas securing the facial credits. This model is organized autonomously and unbroken mounted throughout the voice encoder designing.

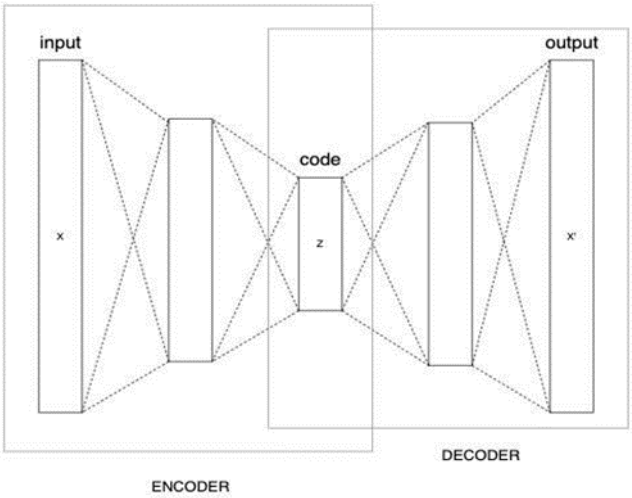

Fig 2: The representation of the autoencoder is as two consecutive Keras models: the encoder and decoder separately.

4. Result Analysis

We assess and mathematically value how—and in what way—our Speech2Face recreations, noninheritable squarely from sound, appear as if the real face photos of the speakers. The trained info contains ninety pictures on that the model was ready, and also the testing info contains around ten entirely new pictures. Our face recovery

execution shows the chance of the real image of the speaker and consequently, we are able to accomplish a decent exactness. The Fig 3 shows implementation of the face decoder that retrieves the reconstructed faces. increasing the computation power and utilizing a complete dataset will assist U.S. with accomplishing a lot of distinguished exactness. We tend to inquiry our info base of face photos by different our Speech2Face forecast of information sound to all or any VGG-Face highlights within the information set. For every

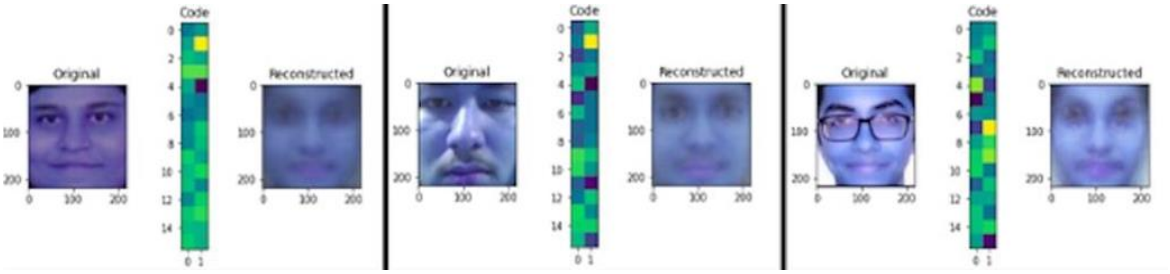

Fig 3: Implementation of Face decoder which retrieves the reconstruction of faces.

inquiry, we tend to get the closest recovered example. Assumptive the speech recommends that the individual is Indian, the associate face is nearest to an Indian. The larger a part of the anticipated folks matches in status and sex. If the real image is not close to the recovered example, this may well be ascribed to associate degree immoderate bound face embrace as an example facial hair or abundance beard growth that the model did not adapt of course attributable to less info. If the standard of the cut photos is poor, the facial highlights will not be legitimate.

Analysis of the initial and reconstructed Faces:

Qualitative results on the dataset are shown in figure higher than. For every example, actuality image of the speaker is shown as a relation to the model, the face reconstructed from the face feature (evaluated from actuality image) by the face decoder, and also the face reconstructed from a 6-seconds audio section of the person's speech, that is our result. Whereas wanting somewhat like average faces, the reconstructed pictures capture made physical info regarding the speaker, like their age, gender, and race. The expected pictures conjointly capture extra properties just like the form of the face or head (e.g., elongated vs. round), that we regularly realize per actuality look of the speaker.

5. Conclusion

Our paper has popularized a scholarly investigation of face generation straightforwardly from the sound account of an individual talking. We delineated that our procedure probably predicts possible appearances with the facing qualities consistent with those of authentic pictures. We have presented the novel assignment of coordinating among appearances and voices. The aftereffects of the trials firmly propose the presence of cross-modular biometric data, prompting the end that maybe our appearances are increasingly like our voices than we might suspect.

Likewise, we can discover the outward appearance delineating the feeling of the individual Accents shape our impression of an individual. The characterization into social classifications, as for example race assumes a significant job is imagining an individual..

References

1. Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, Rita Singh. Disjoint Mapping Network for Cross-modal Matching of Voices and Faces. ICLR 2019 Conference Blind Submission.
2. Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, Xavier Giro-i-Nieto. Wav2pix: Speech-Conditioned Face Generation using Generative Adversarial Networks. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
3. Van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995).
4. Chua-Hung Wan1, Shun-Po Chuang2, Hung-Yi Lee2. Towards audio to scene image synthesis using generative adversarial networks. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
5. Arsha Nagrani, Samuel Albanie, Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

6. Miyuki Kamachi, Harold Hill, Karen Lander, Eric Vatikiotis-Bateson. Putting the Face to the Voice: Matching Identity across Modality. Current biology (2003).
7. Michele Merler, Nalini Ratha, Rogerio Feris, John R. Smith. Diversity in Faces.
8. Tae-Hyun Oh, Tali Dekel Changil Kim, Inbar Mosseri, William T.Freeman, Michael Rubistein, Wojciech Matusik. Speech2Face: Learning the Face behind a Voice. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
9. Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, William T. Freeman. Synthesizing Normalized Faces from Facial Identity Features. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).