ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Persons facial image synthesis from audio with Generative Adversarial Networks

Huzaifa Maniyar[1], Suneeta V. Budihal[2] and Saroja V. Siddamal[3]

ABSTRACT

This paper proposes to build a framework with Generative Adversarial Network (GANs) to synthesize a person's facial image from audio input. Image and speech are the two main sources of information exchange between two entities. In some data intensive applications, a large amount of audio has to be translated into an understandable image format, with automated system, without human interference. This paper provides an end-to-end model for intelligible image reconstruction from an audio signal. The model uses a GAN architecture, which generates image features using audio waveforms for image synthesis. The model was created to produce facial images from audio of individual identities of a synthesized image of the speakers, based on the training dataset. The images of labelled persons are generated using excitation signals and the method obtained results with an accuracy of 96.88% for ungrouped data and 93.91% for grouped data.

Article information:

Keywords: Generative Adversarial Network(GAN), Image Synthesis, Speech Processing, Deep Learning

Article history:
Received: May 12, 2021
Revised: June 21, 2021
Accepted: December 6, 2021
Published: May 28, 2022 (Online)

DOI: 10.37936/ecti-cit.2022162.246995

1. INTRODUCTION

The issue of audio profiling to deduce a person's bio-physical parameters like age, gender, different health conditions and many more from the audio are considered. The task poses many issues such as a feasibility to learn an individual's somatic parameters, and to actually synthesize the full face from their audio. Constructively, audio profiling with our architecture answers this question. Given an unheard audio wave from an unknown individual, the image of the face must be constructed that has numerous matches with individual speaker in terms of identity is the challenge.

It allows many new uses in the various media space such as improving sound in existing recordings in groups where somebody is talking for instance, blogging recordings or news bondage recordings, empowering video-visiting in quiet territories like libraries or in uproarious situations. The test of discourse age recommended in investigation was that discourse models lost pitch data and therefore have low coherence in the reproduced discourse. In the proposed model, a complete learning based model addresses the high dimensional sound-related spectrogram with its sound age process. Contrasted with existing models LPC, LSP, Customary Spectrogram and sound-related spectrogram jell, both the pitch and reverberation data of sound signed values by a 128 dimensional element at each time step. In the event that the sound is of 8KHz, sound-related spectrogram will respect a great many parameters. Additionally, these sound-related spectrograms [1] are profoundly associated, which makes it very difficult for the system to get familiar with the sound-related spectrogram directly.

2. LITERATURE SURVEY

There are many different approaches of obtaining face images from audio and it is related to many research fields. With an audio profiling architecture, many useful features can be extracted from the audio speech signal of a particular person like age, gender, anthropometric measurements, emotional characteristics, identity, etc. Presently there exists an impor-

[1]The author is with School of Electronic and Communication Engineering, Student of Engineering, KLE Tech., Hubballi, Karnataka, India-580031, Phone(+91)9448821704, E-mail:huzzumaniyar@gmail.com
[2]The author is with School of Electronic and Communication Engineering, Professor of Engineering, KLE Tech., Hubballi, Karnataka, India-580031, Phone(+91)9448821704, E-mail: suneeta_vb@kletech.ac.in
[3]The author is with School of Electronic and Communication Engineering, Professor of Engineering, KLE Tech., Hubballi, Karnataka, India-580031, Phone(+91)9480369502, E-mail:sarojavs@kletech.ac.in

tant field of research for deriving personal profile features from an individual audio. GAN is being used to synthesize the images of the face by noise. It doesn't contain any other constraint knowledge as the input. However, it uses a distinct label as a constraint, but it works with a closed-set framework. Moreover, the difficulty encountered is the trouble within the model generation, where the reference image is to be generated based on target identity. The task is challenging, as the input speech signal is of different modality from the output factor to be generated using GAN.

The authors of paper [2] discussed a deep neural network model which was used to extract a person's facial image. It was demonstrated that generating facial images can provide better comprehensive views of voice and face correlations. This led to many applications for predicting specific features. The authors of paper [3] discussed recent research in the area of profiling humans from voices, which seeks to deduce and describe the speaker's entire personality and surroundings from speech. It describes how the human voice is unique in its ability to both capture and influence human personality. The research in paper [4] demonstrated that given an audio file, it is possible to decide which of two facial images is of the speaker. There are standard datasets available publicly for face recognition from static images (VGGFace) and identification of people from audio (VoxCeleb). A CNN architecture was used for binary and multi-way cross-modal face and audio matching. The work in paper [5] highlights the investigation and identification of sensitive face joint embedding with speech. This embedding enables cross-modal retrieval from voice to face and vice versa. The researchers in paper [6] showed attribute-guided face generation. From a low-resolution facial image, an attribute vector that can be extracted from a high-resolution image. The proposed technique generates a high-resolution image satisfying the given attributes. To address this problem, we apply a new condition for CycleGAN and propose conditional CycleGAN.

In voice to face match up [7], the model used is cross modal matching [8] between audio and the face. The model aims at matching the input modality to a library of multiple inputs, which are used for training and to predict the most suitable and expected results, which may or may not be from other output modality [9]. The existing architectures had become a progressively popular solution in past years. The technique uses an N-way classification problem, which is assigned to grasp the simple embedding technique/method for cross-modal inputs. The talking face model [10] is the most recent work that has been carried out. It is used to generate the continues frames of images which are synchronised with the audio input and the lip movement in those frames. Here the input is a still image of a person wherein the lip movement will be synchronised with the audio irrespective of the identity of the speaker in the given audio clip. The task of talking faces [11] is significantly different as we are concentrating on the speaker identity, but not on the content of speech. The model of talking face concentrates on the speech content of the audio and ignores the individual speaker identity. Many recent surveys have been made by many different authors [12], but the approach claimed is different in paper [13]. The authors of paper [14] used an audio sequence of a person to generate a photo-realistic output video of a target person in synchronism with source audio. An audio-driven facial generation is developed using DNN that employs a latent 3D face model space. The Model inherently learns temporal stability through the 3D representation. Through an iterative design process [15], a design pipeline was developed to generate visualisations of audio using Pix2Pix, a conditional adversarial neural network. Through a process of extracting audio features [16], converting these into simple grid images, and feeding them into a trained machine learning model, a new visual interpretation of the audio can be experienced in the form of images and videos.

We propose a learning based GAN model to take audio as input and predict face highlights, which are changed into understandable discourse. The objective of the model is to build the network using GANs, which will be able to generate the image of the corresponding audio/voice recording input. The generator of GAN processes in terms of speech embedding them in a vector by the voice embedding network. GAN is trained by a discriminator and a classifier. The discriminator detects the image, which is generated by the generator as realistic or fake. The classifier ensures that the individual identity of generated output image matches the true identity of the speaker.

The formulated problem is to design and develop a learning based model using GAN to synthesize facial images for input audio signals.

The following contributions are made towards the stated problem.

- We have introduced a network for generating faces from a voice in an audio using profiling to determine the relationship between audio and face modalities.
- We have proposed and built a simple and effective framework based on GANs for this. The task was accomplished with good accuracy.
- We have also proposed a classifier which is used to validate the synthesized image of a labelled person using a cross-modal matching method.
- The framework is able to synthesize faces with individual speaker identity for input audio.

The remaining part of paper organization is as follows. Section 3 deals with proposed system design, where the design specifications along with the schematics and theory behind each block is explained along with implementation details. Results are discussed in section 4 and conclusion is in section 5.

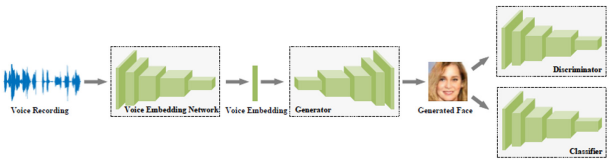

Fig.1: *The GANs-based advanced model used for face generation from audio speech showing the voice embedding network, the generator, the discriminator, and the classifier.*

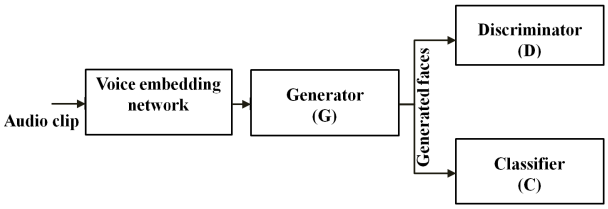

Fig.2: *The functional block diagram: voice embedding network, generator, discriminator, and classifier.*

3. THE PROPOSED FRAMEWORK

At every point, sound can be represented by a single value amplitude (a) as a simple continuous function of time (t), so that a=f(t). For a signal is to be divided into a finite set of numbers, we need to choose a sampling rate. If the sampling rate is 16000 Hz, it means that we will write down the value of the amplitude in intervals of 1/16000 seconds, so that an audio clip of 2 seconds will result in a single vector of 32000 numbers. While images can be viewed as matrices (or tensors), audio signals can be considered as simple vectors. The designed architecture is supposed to synthesize a facial image for a given raw audio clip. We have used some specific notations to describe the different parameters used to build the network. Symbol 'R' represents voice recordings, along with the superscript or subscripts to identify specific individual recordings. Similarly, 'I' is used to represent face images. The identity of a particular individual which provides audio or face data is labelled as y. The true identity of (the subject of) audio recording R is y^R and the image of the person's face I is y^I. The functions that fit an audio recording or face to a particular identity are labelled as y^R = label(R) and y^I = label(I) respectively. Additional symbols will become apparent as we introduce them later.

We used the architecture shown in Figure 1 for the proposed model, which decomposes F(R; Θ) into a series of a pair of components, F_e(R; Θ_e) and F_g(e; Θ_g). F_e(R; Θ_e): $R \rightarrow e$ is a voice/audio embedding function with the variable Θ_e which takes in an input a audio recording R and gives an output of an embedding vector 'e' that holds all the necessary quality information in R. F_g(e; Θ_g) : $e \rightarrow I$ is a function of a generator that takes an embedding vector as input and generates an image of face $\widehat{f}$. The main goal is to train the GAN model F(R;Θ) (with the variable Θ) to take in an input audio recording R and give the image of a face as output $\widehat{f}$ = F(R; Θ) that belongs to the particular individual speaker R, i.e. such that label($\widehat{f}$) = label(R).

3.1 Training the network

Assume the availability of images of faces and voice data from a set of cases Y = {y_1; y_2; . . . ; y_n}. Similarly, we also have a set of audio recordings R = {r_1; r; . . . ; r_n}, with identity labels Y^R = {y_1^R,y_2^R, . . . ,y_p^R} and a set of faces F = {f_1,f_2,. . . ,f_q} with identity labels Y^f = y_n^f such that y^R belongs to Y for all y^R belongs to Y^R and y_f belongs to Y for all y^f belongs to Y^f, where p and q may not be equal. In addition to the above, we define two group of labels L = {l_1; l_2; ·;l_M} and $\widehat{L}$={$\widehat{l}_1$; $\widehat{l}_2$; $\widehat{l}_3$;....$\widehat{l}_n$ for all i values $\widehat{r}_i$=0}, corresponding to Y^f and Y^R respectively. L is a group of labels that specifies that all faces in F are "real". $\widehat{L}$ is a group of labels that specifies that any faces generated from any r belonging to R are synthetic or "fake". The number of training videos are 145,569 and the number for testing was 4,911. Number of audio clips were 1,092,009 for training and 36,237 for testing.

3.2 GAN framework

In the model as shown in Figure 2, for training we consider and impose two requirements. First, for any actual voice input R, the output $\widehat{f}$ of the generator must be a genuine/lifelike image of the face. Second, it must be associated with the same individual as the audio clip, i.e. label($\widehat{f}$) = f^v. We will use a GAN architecture model to train $F_e(:;\theta_e)$ and $F_g(:;\theta_g)$. This will be required to avoid the loss that can be used to study the model feature parameters.

In the model we have designed an adversarial objective. The first is the discriminator F_d, which is used to determine the face that is generated by the generator. Whether the face is real/actual or generated by the generator, is a parameter used to assign the real/fake label. The loss function for the discriminator F_d is given by $L_d(F_d(f)$. The second adversarial objective is the classifier F_c, which is used to assign a label to any real or the generated face. The loss function for the classifier used for the discriminator is given by $L_d(F_c(f))$. The generator is provided with the audio clip input 'R' to generate a face, which is classified into r with an identity label.

In the implementation of model, the voice embedding network is represented by F_e, where R is used

to represent the MFCC Mel-Spectrographic form of the audio signal. An n-dimensional vector e is obtained by the final layer of convolution pooled over time. The Generator function is represented as F_g. f and $\widehat{f}$ are RGB images with $w \times h$, of same resolution. The Discriminator and Classifier are represented by Fd and Fc, respectively. The loss functions L_d and L_c are the two components of the cross-entropy loss.

4. RESULTS AND DISCUSSIONS

The architecture used in this paper is made up of Generative Adversarial Neural networks. A GAN network basically consists of a structure that is more similar to a structure of a VGG. A deep neural network is specially made for image and audio/video processing. The network, once given an input, tries to learn the patterns within it and predict the output on the basis of how it has been trained and what features it was able to extract from the training dataset. During training, care has to be taken so that the network does not learn too much regarding the data given to it, to avoid over fitting. The parameters such as Loss and Val loss are frequently calculated in every epoch to minimize the error and maximize the accuracy of the output face. Error decreases as the loss function decreases, as shown in Figure 3. This is an indication that over fitting could be avoided at that point when the Loss function stops decreasing.

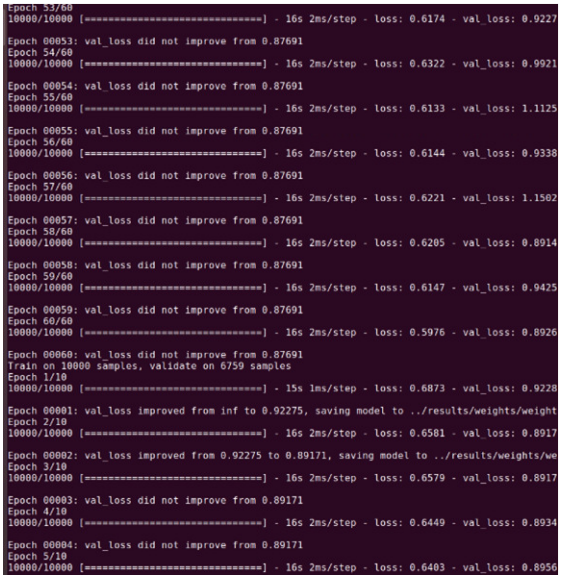

Fig.3: *The implemented result (kernel window) for the loss and v-loss.*

4.1 Pre-processing pipeline for audio and face

For pre-processing of audio segments of the data set, web RTC separates the speech containing part of audio recording. Separated pre-processing pipelines for audio data are implemented for the segments of audio speech and images. A voice activity detector is used. This is an interface from the WebRTC architecture to separate speech containing regions of the recordings for audio segments. Log mel-spectrograms with 25ms analysis window, in which 10ms is hop between the frames, is used. We have randomly cropped an audio clip of around 2 to 8 seconds, which is used for training the GAN, a full recording is considered for testing the network. For the image of face data, [17] facial landmarks in all the images of face are determined. The cropped RGB images of face with the size of $3 \times 64 \times 64$ are acquired through similarity transformation as shown in Table 1.

Table 1: *Dataset details with the total number of speakers with audio and image with identity labels*

	Training	Validation	Testing	Total No.
No. of audio segments	1,13,322	14,182	21,850	149,354
No. of images (face)	1,06,584	12,533	20,455	139572
No. of subject	924	112	189	1,225.

4.2 Layer dimensionality and activation functions

A GAN obtains the previously mentioned faces of size $64 \times 64 \times K$ as information to the system. It utilizes VGG-like piles of little 3×3 responsive fields in its convolutional layers. The design involves five back to back convolutional, softmax pool squares with $32{\times}64{\times}128{\times}256{\times}512$ kernels. These are trailed by two completely associated layers, which contain 512 neurons each. The final layer of CNN has the size 18, which compares to the size of the MFCC vector to get output as shown in Figure 4. Back propagation is done with the help of the mean square error loss function.

Voice Embedding Network			Generator		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$64 \times t_0$	Input	-	$64 \times 1 \times 1$
Conv $3_{/2,1}$	BN + ReLU	$256 \times t_1$	Deconv $4 \times 4_{/1,0}$	ReLU	$1024 \times 4 \times 4$
Conv $3_{/2,1}$	BN + ReLU	$384 \times t_2$	Deconv $3 \times 3_{/2,1}$	ReLU	$512 \times 8 \times 8$
Conv $3_{/2,1}$	BN + ReLU	$576 \times t_3$	Deconv $3 \times 3_{/2,1}$	ReLU	$256 \times 16 \times 16$
Conv $3_{/2,1}$	BN + ReLU	$864 \times t_4$	Deconv $3 \times 3_{/2,1}$	ReLU	$128 \times 32 \times 32$
Conv $3_{/2,1}$	BN + ReLU	$64 \times t_5$	Deconv $3 \times 3_{/2,1}$	ReLU	$64 \times 64 \times 64$
AvePool $1 \times t_5$	-	$64{\times}1$	Deconv $1 \times 1_{/1,0}$	-	$3 \times 64 \times 64$
Discriminator			**Classifier**		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$3 \times 64 \times 64$	Input	-	$3 \times 64 \times 64$
Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$	Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$
Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$	Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$
Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$	Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$
Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$	Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$
Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$	Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$
Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$	Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$
FC 64×1	Sigmoid	1	FC $64 \times k$	Softmax	k

Fig.4: *Layer dimensionality and activation functions used for GAN and classifier architectures, leaky ReLu and ReLu are used for classifier and discriminator.*

We have used a ReLu function as the activation function in all layers except in the last two layers where we have used a sigmoid activation function as output activation function. We are using the Adam Delta Optimizer as the optimizer in the network. The Learning rate Eeta is 0.0002. The learning rate exactly tells us what step or what amount of step has to be taken while training. We have used 0.5 drop out in the convolutional layers and 0.99 dropout in the fully connected layers in the architecture. The network was trained with one speakers data (1000 files) for about 10 hours in the network. We ran the network for 60 epochs and from there on we additionally ran the network for 10 more epochs to increase the accuracy and decrease the error.

4.3 Toolkit used

The previous sections were about the theoretical, architectural and other aspects of the implementation. In this section we will be going through the technical and coding aspects related to the work. The code was developed in python language. We have made use of several libraries.

The GAN based model for face synthesis from raw audio is built with the PyTorch library, which is used for computer vision and natural language processing applications. As discussed earlier, the dataset is used for training in a mini batch size of 128 for about 60 epochs. The training is stopped, once the loss function and Val loss functions stop decreasing. Later it was found that only mouth region recognition lead to less accurate outputs compared to that of face region detection. Usually, a face region is cropped with the help of a mask that has 2 holes within it as shown in Figure 5.

Fig.5: *Synthesizing the face by training the GAN architecture*

Face detection is also an important aspect of the model. Face detection is done with the help of OpenCV that has a cascade based detector for faces. From the face that has been detected using this method further we then try to derive the feature vector needed. Usually people take out only mouth regions and used to extract the features. At that point the convolution slides over to the following pixel and rehashes a similar procedure until all the picture pixels have been processed.

Many trial and error methods of experiments are carried out to get out the best possible results of the model, by trying different parameters and different modalities while training.

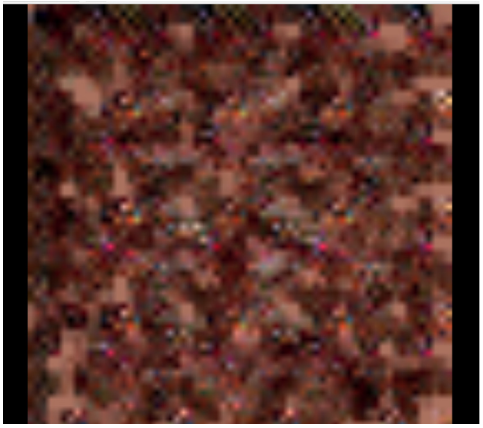

Fig.6: *The result obtained for the audio sample with different noise was blurry and unrecognisable*

Fig.7: *The result obtained for the audio sample where GAN was trained and was able to map the manifold of speech signal to the manifold of the face.*

4.4 Qualitative result of the model

As a method of experiment, the model was trained with different levels and types of noise signals and for the different durations of time(like 5s, 8s,10s,...). The results obtained for the experiment were missing the faces, which were blurred as shown in Figure 6, and were alike for all of the different audio sample inputs. In the next experiment, the model was trained with the audio input with no noise added to the regular speech signal of the audio clip. The result obtained for the training was acceptable, with recognisable face output when GAN was trained and learned to map the manifold of speech signal to the manifold of face as shown in Figure 7.

4.5 Quantitative results of the model

We have constructed a model, which was trained using the sigmoid activation function with all the specifications of training discussed in the previous

sections. It was able to synthesize a face when an audio/ speech was given as an input to the model. The graph in Figure 8 gives the validation loss and loss accuracy of the model trained with the sigmoid activation function. Using transposed convolutions to upscale the audio from noise to 16000 long vectors impacted the synthesized audio. That notoriously produces checker board artefacts in images and the equivalent in audio signals. The discriminator could easily learn to spot fake audio based on this artefact alone, deteriorating the whole training process. This paper proposes a smart solution called Phase Shuffle, but we avoided transposed convolutions and use up sampling layers (Nearest Neighbours) followed by normal convolutions instead. Other parameters like sampling rate, sampling frequency, normalization factor, and the epoch number per cycle, etc. were optimized for performance.

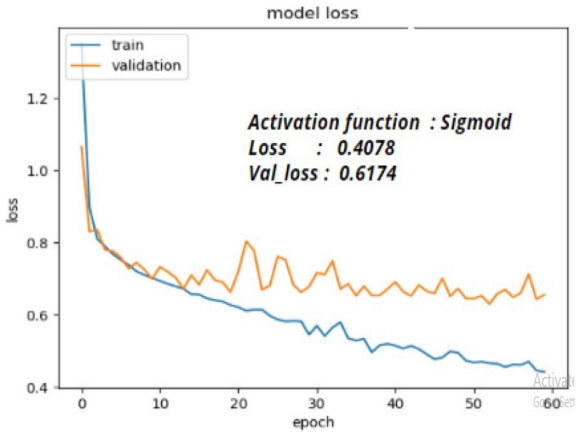

Fig.8: *The training and validation loss of the model.*

A comparison of accuracy of the proposed architecture and other models is given in Table 2. We can differentiate and compare the results obtained for the model where the model is able to reconstruct the faces for input audio. Table 2 gives the accuracies for two different groups of data. One model is trained for all the data contained in the dataset irrespective of any groups like gender, age, etc.(i.e., ungrouped by features). The other is for data that is grouped depending upon the features like gender, age, etc.

Table 2: *Accuracy of the results obtained by the model architecture from the literature survey and of the model we trained.*

Sl.No.	% Accuracy of Ungrouped by any feature (Training data/Testing data)	% Accuracy of Grouped by features like gender, age, etc.(Training data/Testing data)
SVHF[18]	– / 81.01	– / 65.02
DisMNets-I [19]	– / 83.45	– / 70.97
DisMNets-G [20]	– / 72.90	– / 50.35
Proposed model	96.88 / 76.09	93.97 / 59.70

5. CONCLUSION

Our work tested the feasibility of reconstructing a face from audio using a GAN architecture to mechanically learn relevant visible features. We trained the network with 1,06,584 images of faces from 924 different people and 113,322 speech segments from the voxceleb2 and VGGface datasets. We trained and tested the model with other speakers to cross check the performance, and altered the activation function in the last layers to validate the accuracy and performance. The important result of the paper is that the model is able to synthesize the faces for the corresponding audio signals with an accuracy of 96.88%.

References

[1] T.-H. Oh, T. Dekel, and C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein and W. Matusik, , "Speech2Face: Learning the Face Behind a Voice," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7531-7540, June 2019.

[2] T. Dekel, C. Kim, I. Mosseri, W. T. Freeman and M. Rubinstein, "Speech2face: Learning the face behind a voice," *IEEE conference on computer vision and pattern recognition*, 2019.

[3] S. Pavaskar and S. Budihal, "Real-Time Vehicle-Type Categorization and Character Extraction from the License Plates," *Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing*, vol. 768, 2019.

[4] A. Nagrani, S. Albanie and A. Zisserman, "Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8427-8436, 2018.

[5] A. Nagrani, S. Albanie and A. Zisserman,"Learnable pins: Cross-modal embeddings for person identity," *computer science bibliography*, 2018.

[6] Y. Lu, Y.-W. Tai and C.-K. Tang, "Conditional C-GAN for attribute guided face image generation," ,2017.

[7] Mohamad Hasan Bahari, ML McLaren, DA Van Leeuwen, et al. "Age estimation from telephone speech using i-vectors," 2012.

[8] Y. Wen, M. A. Ismail, W. Liu, B. Raj and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *International Conference on Learning Representations ICLR 2019*, pp. 1-17, 2019.

[9] J. Bao, D. Chen, F. Wen, H. Li and G. Hua, "Towards Open-Set Identity Preserving Face Synthesis," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6713-6722, 2018.

[10] V. B. Suneeta, P. Purushottam, K. Prashantkumar, S. Sachin and M. Supreet, "Facial Expression Recognition Using Supervised Learning," *Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing*, vol. 1108, pp. 275-285, 2020.

[11] A. Jamaludin, J. S. Chung and A. Zisserman, "You said that?: Synthesising talking faces from audio:," *International Journal of Computer Vision*, pages 01-13, 2019.

[12] R. Singh, "Reconstruction of the human persona in 3D and its reverse," *In Profiling Humans from their Voice, chapter 10. springer nature Press*, 2020.

[13] R. Huang, S. Zhang, T. Li and R. He, "Beyond face rotation: Global and local perception GAN for photo realistic and identity preserving frontal view synthesis," *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 2439-2448, 2017.

[14] J. Thies, M. Elgharib, A. Tewari, C. Theobalt and M. Nießner, "Neural Voice Puppetry: Audio-Driven Facial Reenactment," *Computer Vision (ECCV), Lecture Notes in Computer Science*, vol. 12361, 2020.

[15] R. Zhang, P. Isola and A. A. Efros, "Colorful image colorization," *In European conference on computer vision*, pp. 649-666, 2016.

[16] S. Willcox, "Artificial Synaesthesia: An exploration of machine learning image synthesis for soundscape audio visualisation," *Victoria University of Wellington*, 2021.

[17] H. Zhou, Y. Liu, Z. Liu, P. Luo and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," *AAAI Conference on Artificial Intelligence (AAAI) 2019*, 2019.

[18] Y. Wen, R. Singh and B. Raj, "Face Reconstruction from Voice using Generative Adversarial Networks," *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[19] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem and a dataset of 230,000 3D facial landmarks," *In International Conference on Computer Vision*, 2017.

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

Huzaifa Maniyar completed her schooling in Hubli. She received B.E degree in Electronics and Communication Engineering, M. Tech in Digital Electronics. Her areas of research include machine learning and communication. She has authored few publications in national and international conferences and journals.

Suneeta V. Budihal completed her schooling in Hubli itself. She received B.E degree in Electronics and Communication Engineering, M. Tech in Digital Electronics and PhD in wireless communication. She is serving as Professor at KLE Technological University, in ECE department, since 22 years. She is a life member of ISTE. She has authored many publications in national and international conferences and journals. Her research areas include wireless communications and Machine learning for wireless communication communications. She has organized STTP and FDPs funded by AICTE and VGST.

Saroja V. Siddamal completed her schooling in Hubli. She received B.E degree in Electronics and Communication Engineering, M. Tech in Digital Electronics and PhD Co-Processor design and its architecture. She is serving as Professor at KLE Technological University, in ECE department, since 22 years. Her areas of research include VLSI Signal Processing, machine learning and coprocessor design. She has authored many publications in national and international conferences and journals. She has completed funded projects and 2 patents are published to her credit.