International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Analysis of Voice to Predict the Physical Attributes of an Individual by Using the Face Reconstruction Approach

[1]Prashant S. Kolhe, [2]Ranjeet Bidwe, [3]Deepak Mane, [4]Bapurao Bandgar, [5]Sunil Shinde, [6]Sunil M. Sangve

Submitted: 25/11/2023 **Revised**: 05/01/2024 **Accepted**: 15/01/2024

Abstract. We hear people's voices on the radio, on telephones, etc. Many features can be extracted from a person's Voice. Factors like age, gender, and ethnicity can characterize a person's Voice. How many can we infer about someone from the way they communicate? In this work, we have tried to recommend a neural community structure that enables us to extract features like age, gender, and ethnicity from a person's speech. Different features of the Voice helped us to extract different features, like the pitch of the Voice giving information about a person's gender. Accent, speed, and pronunciation gave us information about a person's ethnicity. The main goal of this work is to determine the extent of information that can be extracted from a person's speech for the Indian accent datasets. Here, we have used a Voice Encoder and Face decoder model. Voice Encoder keeps track of the vocal features, and the face decoder uses these vocal features to generate the face. The whole neural architecture is inspired by Generative Adversarial Networks (GANs).

Keywords: *Face Recognition, Face Reconstruction, Voice analysis, Feature Extraction, Voice Encoder*

1. Introduction

The physical attributes extracted from a person's Voice can help us reconstruct facial vectors, which can lead to the generation of the face[1].One of the most critical programs is policing. Phone abuse and harassing cellphone calls are old trouble, but it is growing in incidence. And with burner phones and Voice over net Protocol (VoIP) offerings, it is more and more difficult to trace the identification of a caller and seize people making threatening or harassing cellphone calls[2]. Police departments urge recipients of threatening calls to document details about the caller, their gender, approximate age, etc. Our product can do even higher. From a voicemail or snippet of a harassing name, it can generate a composite cartoon of the caller. This can be used to locate or verify their identification, especially if this becomes part of a bigger sample of harassment, like stalking, as many harassing calls are.

It has been observed that major criminals hide their facial identities not to get caught by the police. The topic of face retrieval from voice features has been very novel as of the present time. Additionally, the existing works show promising results on face regeneration from Voice. But they fail to capture the features of the Indian Accent in their works.

Our predominant purpose of this work is to look at what volume we will infer from how a person looks from how they talk. Specifically, from a brief input audio section of a person talking, our approach, without delay, reconstructs an image of the man or woman's face in a canonical form (i.e., frontal-facing, impartial expression). Obviously, there's no person-to-one matching between faces and voices. Hence our aim is only sometimes to predict a recognizable picture of the exact face; instead, to seize dominant facial developments of the person correlated with the input speech. We aim at reading an extra standard, open question: what type of facial statistics can be extracted from speech? We display that our reconstructed face photographs may be used as a proxy to deliver the visible homes of the man or woman, including age, gender, and ethnicity. That is achieved without earlier data or the life of correct classifiers for those forms of first-class geometric features. Further, predicting face photos directly from Voice may support useful applications, including attaching a consultant face to smartphone/video calls based totally on the speaker's Voice. This approach can be incorporated with and can be an extended application of a question-answering system [3].

Therefore, we have proposed a solution to generate physical attributes by constructing the face from the vocal features of an individual. Hence, we propose a

[1]*Department of Electronics and Telecommunication, College of Engineering, Dharashiv-413501, Maharashtra, India*

[2]*Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University) (SIU), Lavale, Pune 412115*

[3,5,6]*Vishwakarma Institute of Technology, Bibwewadi, Pune-411037, Maharashtra, India*

[4]*School of Computer Studies, Sri Balaji University, Tathawade, Pune-411033, Maharashtra, India*

prashantsk68@gmail.com
Correspondence Author: ranjeetbidwe@hotmail.com
dtmane@gmail.com,
bapuraob@yahoo.com
sunil.shinde@vit.edu
sunil.sangve@vit.edu

framework inspired by the GAN architecture[4]to generate facial features.

2.Related Work

Artificial intelligence (AI) is intelligence added to computers by teaching them a collection of scenarios and helping them understand the associated rules or knowledge. It enables machines to solve issues and make decisions for themselves. A wide range of AI algorithms is currently employed in interdisciplinary sectors[5]. We can observe a considerable impact on the outcome due to breakthroughs, especially in machine learning (ML) and deep learning (DL) domains. Analyzing data quickly and producing predictions in real-time [6], [7], it exhibits amazing outcomes in numerous applications[8].AI has also proven its importance in the domain of speech to visual analysis. Several survey papers are published on Visual Speech Recognition (VSR)[9]–[11], and Visual Speech Generation (VSG)[12], [13], And the detailed survey is published in the document [14] by Sheng et al. Following are a few state-of-the-art models for visual speech theory that are relevant to the proposed model in this paper are briefly discussed.

Tae-Hyun Oh et al.[15]have presented research on reconstructing a face from an audio recording of a speaker. By employing millions of real-world recordings of people speaking, the proposed system learns to match the feature space of speech with that of a trained face decoder. The Paper has shown that the suggested model can predict believable faces with facial characteristics compatible with actual photographs. The proposed model is a computationally cheap solution for physical attribute retrieval and discussed the core architecture required, which will be useful in designing a more optimal pipeline.

A. Ephrat et al.[16] have isolated an individual's speech from a video containing multiple voices. This work proposes many audio pre-processing and images pre-processing techniques. This gave us the idea of connecting physical attributes with voices. D. R. Reddy [17] has discussed the approaches used for speech recognition and classification. The Proposed model is an audio-visual neural network-based model for single-channel, speaker-independent speech separation. The proposed approach performs effectively in difficult situations, such as multi-speaker blends with background noise. The author team built a new audio-visual dataset comprising thousands of video clips from the Web that featured clearly visible speakers and clean speech to train the model. The research presented state-of-the-art speech separation results with possible voice recognition and video captioning applications. R. A. Khalil [18]et al. discussed speech recognition and classification approaches and emphasized the datasets used. This paper thoroughly analyzes deep learning methods for speech emotion recognition. Based on the classification of several natural emotions, including pleasure, joy, sadness, neutral, surprise, boredom, disgust, fear, and rage, deep learning techniques like DBM, RNN, CNN [29] [30] and AE and their layer-wise architectures are briefly described. This paper explores deep learning approaches' limitations, including their extensive internal layer-wise design, lower efficiency for temporally variable input data, and over-learning during layer-wise information memorization. This research is a foundation for assessing the effectiveness and constraints of existing deep-learning approaches. This report also identifies several promising lines of research for speech-emotion recognition systems. Forrester Cole et al. [19]have developed a neural network that converts pictures of faces taken in the wild into front-facing neutral-expression pictures to capture a person's likeness. The suggested model is resistant to input changes like lighting, position, and expression that impair the performance of earlier face frontalization techniques. The suggested approach offers other downstream possibilities, such as automatically white-balancing photographs and producing unique 3-D avatars. Computer graphics has made substantial use of spline interpolation, but we also need to be aware of work that has employed interpolation as a differentiable module inside a network. Masi et al.[20]discusses the developments in face recognition techniques and emphasized extracting faces from the video to extract physical features. The main developments in deep face recognition and, more generally, in learning face representations for verification and identification are surveyed in this paper. The main state-of-the-art face recognition algorithms that have surfaced in the last five years in prestigious computer vision settings are presented in this work in a straightforward, organized manner.

Apart from all approaches, a few more VSR models are proposed using self-supervised[21], [22], cross-modal knowledge distillation[23], [24], and Graph Neural Network[25], [26] is available for study. A detailed performance evaluation of VSG techniques can be studied from [13].

After performing a detailed survey, it is found that Intra-class variations like Speech context and speaker differences and Inter-class variations like visual ambiguities are the main challenges in recognizing speech. In contrast, information coupling is the main challenge in generating speech.

3. Proposed Architecture

Here we have proposed the system architecture for reconstructing the feature that captures the features of the Indian Accent in this work. The architectural view is shown in Figure 1.

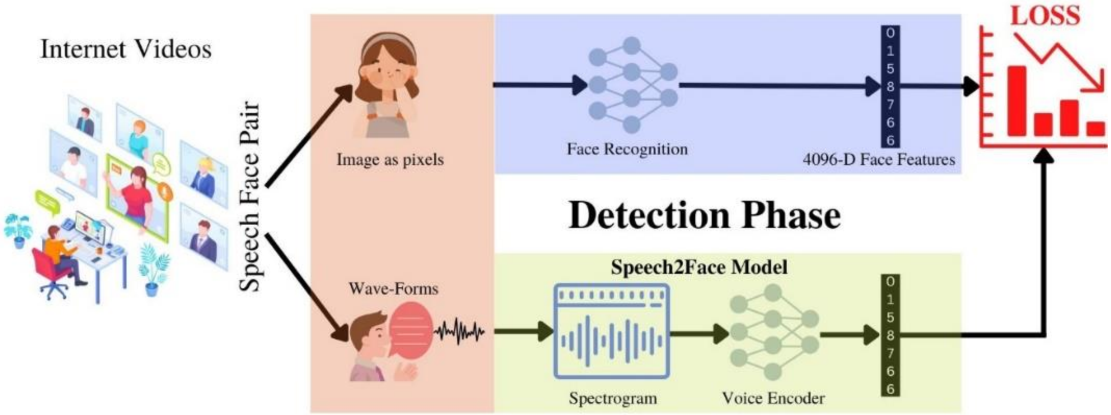


Fig. 1. Speech2Face model and training pipeline.

The major components of the above architecture are as follows-

- **Face Recognition-**The Dlib face detector was used to detect the face, and VGG Face was usedto extract 4096 image vectors from the image.
- **Voice Encoder-**The Speech2Face model takes a spectrogram as input and gives 4096 face vectors as input that are compared with the actual face embeddings for calculating the VGG perceptual loss.
- **Loss Function-**The loss function used here is the perceptual loss between actual and generated face features so as to reduce mathematical complexity. L$_2$ normalized function is added so as to reduce computational complexity and make the model convergence easier.

The Loss function is

$$L_{total} = L_1 + L_2 \quad (1)$$

And is given by

$$L_{total} = \lambda_1 || \frac{V_f}{||V_{f|}\ |} - \frac{V_s}{||V_s||} ||_2^2 + \lambda_2 \, L_{distill}\left(f_{VGG}(V_f), f_{VGG}(V_s)\right) (2)$$

Where

$$L_{distill}(a,b) = -\sum_i p_{(i)}(a) \log p_{(i)}(b) \text{ and } p_{(i)}(a) = \frac{\exp(\frac{a_i}{T})}{\sum_j \exp(\frac{a_i}{T})}$$

The process of the conversion of the speech to face feature conversion is shown in the Figure2.

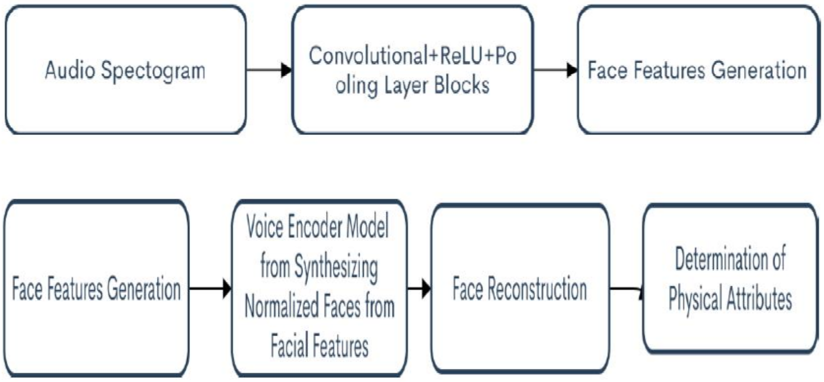

Audio Spectogram
Convolutional+ReLU+Pooling Layer Blocks
Face Features Generation
Face Features Generation
Voice Encoder Model from Synthesizing Normalized Faces from Facial Features
Face Reconstruction
Determination of Physical Attributes

Fig. 2. The Process of conversion of speech to face feature.

The audio spectrogram is given input the CNN layer block and the face features are generated. These Generated face features are further given to the voice encoder model from synthesizing normalized face from facial feature and the face is reconstructed. Forms the constructed face the physical attributes of the images are obtained.

The block diagram of the face reconstruction algorithm is proven in the discernFigure 3.

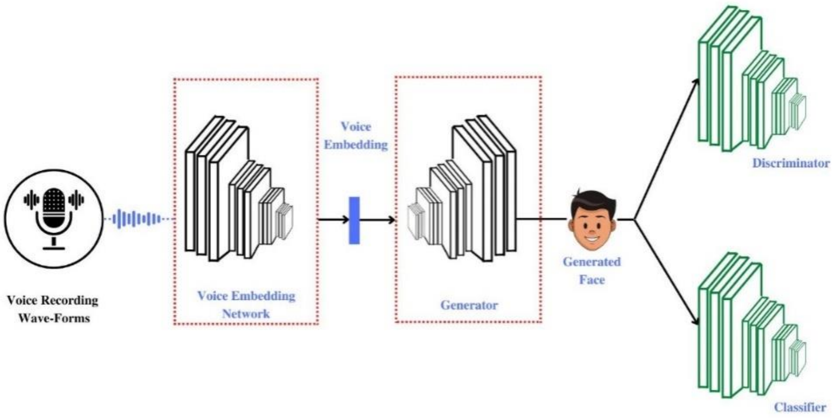


Fig. 3. Reconstruction Face Image Architecture

The above architecture contains four main components-

- **Voice Encoder**- It takes an audio spectrogram and converts it into a latent, hiddenvector for the face decoder network.
- **Generator**- It takes a latent vector and generates its face.
- **Discriminator-** It decides the veracity of the generated face and rates generator based on this so that the generator can increase its optimality.
- **Classifier-** It verifies whether the generated face corresponds to the face of the user whose audio is taken as input.

The above network is the major backbone of this work. The Pipeline was trained for about 72 hours using Google Colab on nearly 5000 videos of different speakers. The loss function is the perceptual loss to save the pixel-to-pixel loss because of the computational complexity.

The Face reconstruction algorithm used in this work is as follows.

Algorithm 1: Face Reconstruction structure

Input: a set of voice recordings with identification label (γ,γ_υ), a hard and fast of labeled face pix with identification label (F,γ_υ). A voice embedding network F_c (υ;θ_e)trained on γ with speaker popularity challenge. $\theta_e$is constant all through education. Randomly initialized θ_e is fixed in the course of the training. Randomly initialized θ_g,θ_d and θ_c

Output: The parameters θ_g

1: even as not converge, do

2: Randomly sample a minibatch of n voice recordings {υ_1, υ_2,……, υ_n} from γ

3: Randomly pattern a minibatch of m face pix {f_1, f_1 ,…….f_m} from F

4: update the discriminator F_d (f; θ_d) by using ascending the gradient (a[i] indicate

the i^{th} element of vector a)

$$\nabla\theta_d(\sum_{i=1}^{n}\log(1-F_d(\widehat{f_\iota})+\sum_{i=1}^{m}\log f_d(f_i)$$

(3)

5: Update the classifier F_c(f; θ_d) by ascending the gradient

$$\nabla\theta_d(\sum_{i=1}^{m}\log f_c(f_i)[y_i^f]$$

(4)

6: Update the generator $F_g(f;\theta_c)$ by ascending the gradient

$$\nabla\theta_g(\sum_{i=1}^{n}\log F_c\left(F_g\left(F_e(\upsilon_i)\right)\right)[y_i^\upsilon]+\sum_{i=1}^{m}\log F_d\left(F_g\left(F_e(\upsilon_i)\right)\right) \qquad (5)$$

7: end while

4.Experimental Details

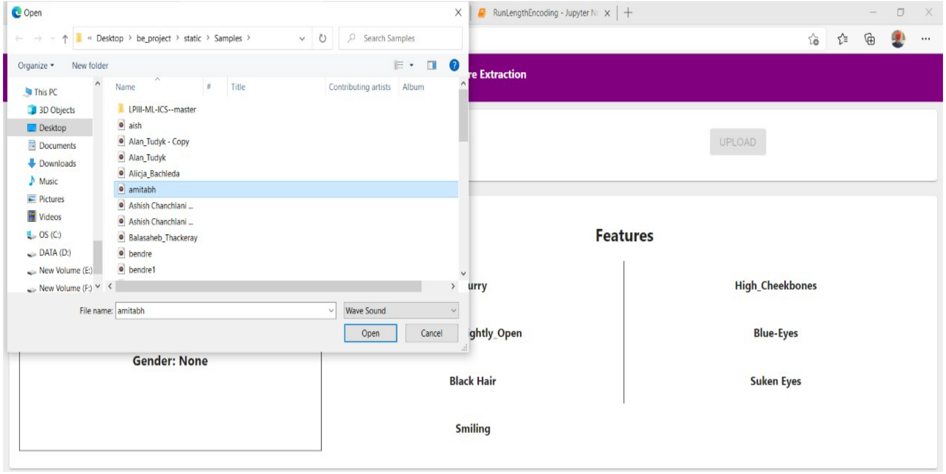
The Face reconstruction from the voice system requires a microphone device to record the audio. Pre-recorded audio from other devices can also be fed as input to the system. To run the ML model Graphics processing unit (GPU) with at least 2 GB is required. The system can be used on operating systems like Windows 10, Ubuntu, and other Linux distributions, provided the ML dependencies are properly installed. The ML model's weights are in Gigabytes, due to which it can't be hosted online. Since we have used Flutter for developing the front, Flutter, along with Android Studio and visual studio needs to be installed on the user's computer. The user should have at least 8GB RAM and an Intel i5 8th gen processor or its equivalent processor of AMD (2.5 GHz or faster). The processor should be quadcore. The recommended browser is Google Chrome, although Firefox can be used provided it is integrated with Flutter first. Various deep learning frameworks are used for developing pipelines which are PyTorch, Tensorow-

Keras,Caffe, ONNX, Dlib and OpenCV Face detector, VGGFace,Librosa, PIL Image etc.

Here we have used the base of the already available dataset that is theVoxCeleb[27]dataset. We have searched and bunched nearly 5000 YouTube videos containing different speakers. We have segregated the audio and the face from the YouTube video using ffmpeg library. The audio was converted into spectrogram using mfcc spectrum and Fourier Inverse Transform. The face images were detected and cropped using the Openface architecture and their face embeddings were stored in the local repository.

Also 40 attributes are taken into considerations which are Clock Shadow, Arched Eyebrows, attractive, bags under Eyes, Bald, Bangs, huge Lips, huge nostril, Black Hair, Blond Hair, Blurry, Brown Hair, furry Eyebrows, obese, Double Chin, Eyeglasses, Goatee, gray Hair, Heavy make-up, high Cheekbones, Mouth Male, slightly Open, Mustache, slender Eyes, No Beard, Oval Face, light pores and skin, Pointy nose, Receding Hairline, Rosy Cheeks, Sideburns,Smiling, immediately Hair, wearing jewelry, sporting Hat, carrying Lipstick, carrying Necklace, wea ring Necktie young etc.

Google Colab is a python Jupyter notebook tool that provides free GPU and uses local storage to run python applications. Colab notebooks were used to train the pipelines using the Free GPU from the Google colab and analyze the results. Atom is used to develop the frontend using React. The Graphical User interface is shown in Figure 4.



File name: amitabh

Gender: None

Fig. 4. Graphical User Interface

In the same GUI the Video clip is given as the input and by using the face reconstruction algorithms the individual face is created and by using the face retrieval algorithm the individual features have been extracted. The facts extraction flow chart is shown in the Figure 5

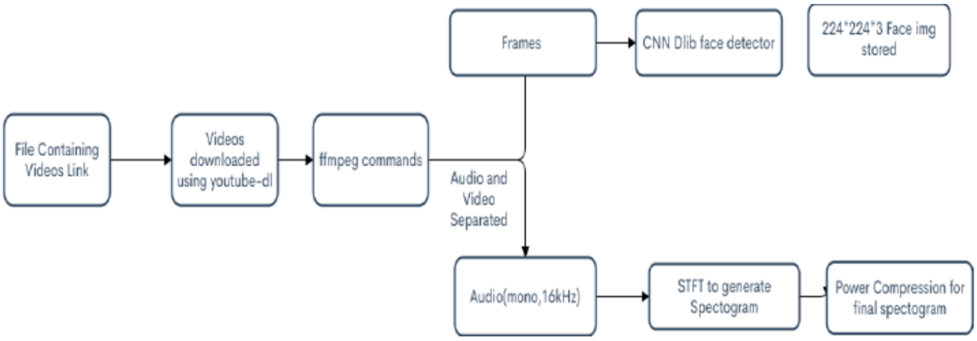

File Containing Videos Link
Videos downloaded using youtube-dl
ffmpeg commands
Frames
CNN Dlib face detector
224*224*3 Face img stored
Audio and Video Separated
Audio(mono,16kHz)
STFT to generate Spectogram
Power Compression for final spectogram

Fig. 5. The fact extraction flow chart

The file containing video link is taken and downloaded using the YouTube Id, then the ffmpeg command is applied to split the video into Audio file and video frame. By using the CNN Dlib face detector, the face image is stored in the image file and STFT is applied to the audio file to generate Spectrogram. Finally, the power compression is done for the final spectrogram.

5.Results and Discussions

The primary cause of this work is to

- Construct face from given audio input.
- Predict the speaker's physical attributes like gender, ethnicity, age, etc. from the audio input.

A variation autoencoder (VAE's) [28] is designed to take audio as an input which gives face embeddings as an output. The generated face embeddings were latent vector of 4096 units. The vector is compared with the stored embeddings from our dataset using the cosine distance and the top 10 embeddings with less distance are given as result.

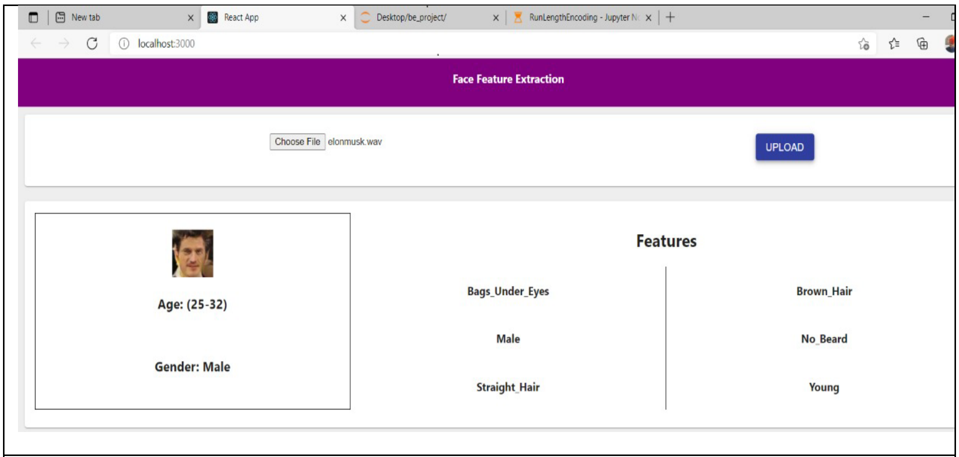
The conditional GAN architecture is designed which takes audio spectrogram asinput and gives reconstructed face image as output. The pipeline was trained on nearly 4800 different speakers. The reconstructed faces generated have shown nearly one to one correspondence with the actual images.

AgeNet pre-trained caffe model used for Age detection. GenderNet caffe model used for Gender detection. A trained lightened moon Mxnet model used for facial attribute extraction. The moon Mxnet model was trained in 40 attribute classes and finally, all the classes were predicted by taking reconstructed face image asinput. The model takes reconstructed face as input and gives the variable number of attributes from the 40 classes that define the face. The pre-trained Agenet and Gendernet model is used to predict age and gender respectively for the reconstructed face.

The complete flow of the system is as follows, firstly our system takes audio as an input, then audio is preprocessed using Fourier inverse transform and then audio is sent to the pipeline and pipeline generates the face as on output. The generated face is sent to mxnet model which predicts age, gender and attributes as output. The preprocessed audio input is sent to a similar face extractor pipeline which extracts the similar faces.

The input audio is captured, and face features are predicted for the sample. The generated face features are compared with the existing face features from the database.

The predicted outcome of the speaker's physical attributes like gender, ethnicity, age etc. from the audio input is as follows.



Face Feature Extraction

elonmusk.wav

Age: (25-32)

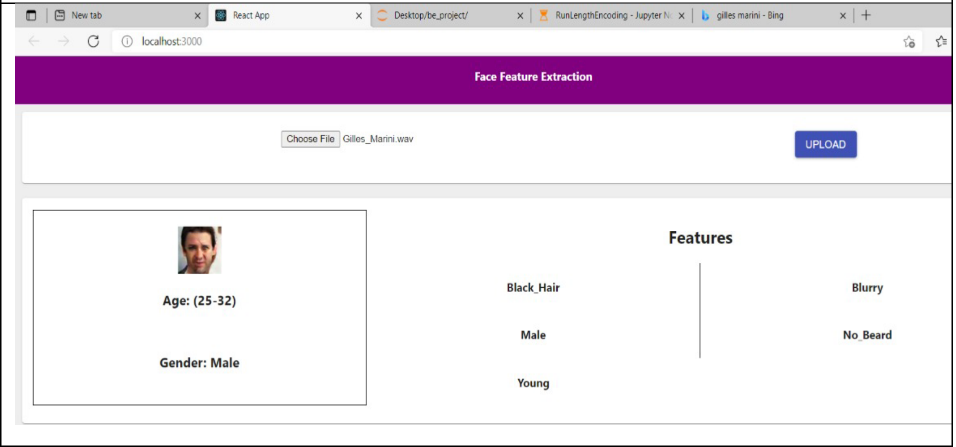
Gender: Male

Features

Bags_Under_Eyes
Male
Straight_Hair

Brown_Hair
No_Beard
Young

Fig. 6. Constructed Face and predicted attributes from Elon Musk's voice clip



Face Feature Extraction

Gilles_Marini.wav

Age: (25-32)

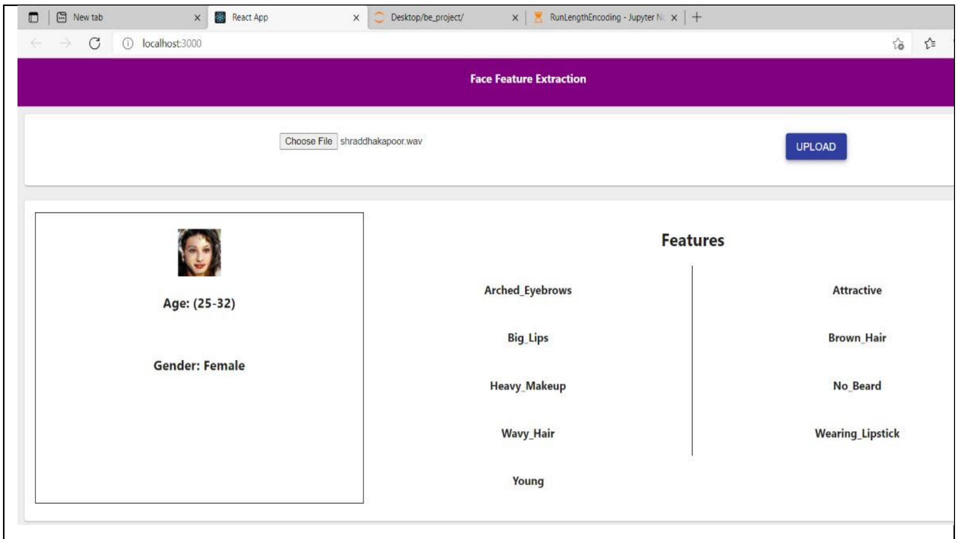
Gender: Male

Features

Black_Hair
Male
Young

Blurry
No_Beard

Fig. 7. Constructed Face and predicted attributes from Gilles Marini's voice clip



shraddhakapoor.wav

Age: (25-32)

Gender: Female

Features

Arched_Eyebrows

Attractive

Big_Lips

Brown_Hair

Heavy_Makeup

No_Beard

Wavy_Hair

Wearing_Lipstick

Young

Fig. 8. Constructed Face and predicted attributes from Shraddha Kapoor's voice clip

The proposed model results are given in Figures (6,7,8). Fig. 6gives the outcome of the face reconstruction algorithm and the outcome of the feature extraction algorithm from speech for Elon Musk's voice clip. This shows that the six features have been extracted for the same. Similarly, the outcome of the face reconstruction algorithm and the feature extraction algorithm for Gilles Marini's voice clip is given in figure (7). This shows that five individual features have been extracted. Fig. 8shows the face reconstruction algorithms and feature extraction algorithm's outcome from Shraddha Kapoor's voice clip. This show that nine individual features have been extracted for the Indian origin Video clip.

5. Conclusion

The principal intention of this work is to decide the quantity of records that may be extracted from someone's speech. Thus, we have proposed a neural network architecture inspired by GANs. This model detects equivalent faces matched with the individual's attributes. This network is able to extract features like age, gender, and ethnicity from a person's speech. The proposed neural network helped to recover maximum physical as well as ethnic attributes to determine the identity of an individual. The Proposed Voice Encoder keeps track of the vocal features, and the face decoder uses these vocal features to generate the face. Thus, we have successfully extracted the maximum number of features of the Indian video clip.

References

[1] H. Maniyar, S. v. Budihal, and S. v Siddamal, “Persons facial image synthesis from audio with Generative Adversarial Networks,” ECTI Transactions on Computer and Information Technology (ECTI-CIT), vol. 16, no. 2, pp. 135–141, May 2022, doi: 10.37936/ecti-cit.2022162.246995.

[2] O. P. Roy and V. Kumar, “A Survey on Voice over Internet Protocol (VoIP) Reliability Research,” IOP Conf Ser Mater Sci Eng, vol. 1020, no. 1, p. 12015, Jan. 2021, doi: 10.1088/1757-899X/1020/1/012015.

[3] B. Zope, S. Mishra, K. Shaw, D. R. Vora, K. Kotecha, and R. V. Bidwe, “Question Answer System: A State-of-Art Representation of Quantitative and Qualitative Analysis,” Big Data and Cognitive Computing, vol. 6, no. 4, 2022, doi: 10.3390/bdcc6040109.

[4] S. Shahriar, “GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network,” Displays, vol. 73, p. 102237, 2022, doi: https://doi.org/10.1016/j.displa.2022.102237.

[5] R. V. Bidwe et al., “Deep Learning Approaches for Video Compression: A Bibliometric Analysis,” Big Data and Cognitive Computing, vol. 6, no. 2, p. 44, Apr. 2022, doi: 10.3390/bdcc6020044.

[6] D. Mane, R. Bidwe, B. Zope, and N. Ranjan, “Traffic Density Classification for Multiclass Vehicles Using Customized Convolutional Neural Network for Smart City,” 2022, pp. 1015–1030. doi: 10.1007/978-981-19-2130-8_78.

[7] D. Mane, K. Shah, R. Solapure, R. Bidwe, and S. Shah, “Image-Based Plant Seedling Classification Using Ensemble Learning,” 2023, pp. 433–447. doi: 10.1007/978-981-19-2225-1_39.

[8] S. Bidwe, Dr. G. Kale, and R. Bidwe, “TRAFFIC MONITORING SYSTEM FOR SMART CITY BASED ON TRAFFIC DENSITY

ESTIMATION," Indian Journal of Computer Science and Engineering, vol. 13, no. 5, pp. 1388–1400, Oct. 2022, doi: 10.21817/indjcse/2022/v13i5/221305006.

[9] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep Learning-Based Automated Lip-Reading: A Survey," IEEE Access, vol. 9, pp. 121184–121205, 2021, doi: 10.1109/ACCESS.2021.3107946.

[10] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," Image Vis Comput, vol. 78, pp. 53–72, 2018, doi: https://doi.org/10.1016/j.imavis.2018.07.002.

[11] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," Image Vis Comput, vol. 32, no. 9, pp. 590–605, 2014, doi: https://doi.org/10.1016/j.imavis.2014.06.004.

[12] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," Speech Commun, vol. 66, pp. 182–217, 2015, doi: https://doi.org/10.1016/j.specom.2014.11.001.

[13] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What comprises a good talking-head video generation?," in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[14] C. Sheng et al., "Deep Learning for Visual Speech Analysis: A Survey," arXiv preprint arXiv:2205.10839, 2022.

[15] T.-H. Oh et al., "Speech2Face: Learning the Face Behind a Voice," CoRR, vol. abs/1905.09773, 2019, [Online]. Available: http://arxiv.org/abs/1905.09773

[16] Ephratet al., "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," CoRR, vol. abs/1804.03619, 2018, [Online]. Available: http://arxiv.org/abs/1804.03619

[17] D. R. Reddy, "Speech recognition by machine: A review," Proceedings of the IEEE, vol. 64, pp. 501–531, 1976.

[18] R. A. Khalil, E. Jones, M. Babar, T. Jan, M. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," IEEE Access, vol. PP, p. 1, Dec. 2019, doi: 10.1109/ACCESS.2019.2936124.

[19] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing Normalized Faces from Facial Identity Features," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 3386–3395. doi: 10.1109/CVPR.2017.361.

[20] M. Wang and W. Deng, "Deep Face Recognition: A Survey," CoRR, vol. abs/1804.06655, 2018, [Online]. Available: http://arxiv.org/abs/1804.06655

[21] C. Sheng, M. Pietikäinen, Q. Tian, and L. Liu, "Cross-Modal Self-Supervised Learning for Lip Reading: When Contrastive Learning Meets Adversarial Training," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2456–2464. doi: 10.1145/3474085.3475415.

[22] R. Arandjelovic and A. Zisserman, "Objects that Sound," in Proceedings of the European Conference on Computer Vision (ECCV), Sep. 2018.

[23] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards Practical Lipreading with Distilled and Efficient Models," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7608–7612. doi: 10.1109/ICASSP39728.2021.9415063.

[24] T. Afouras, J. S. Chung, and A. Zisserman, "ASR is All You Need: Cross-Modal Distillation for Lip Reading," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2143–2147. doi: 10.1109/ICASSP40776.2020.9054253.

[25] H. Liu, Z. Chen, and B. Yang, "Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion," in Proc. Interspeech 2020, 2020, pp. 3520–3524. doi: 10.21437/Interspeech.2020-3146.

[26] C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, "Adaptive Semantic-Spatio-Temporal Graph Convolutional Network for Lip Reading," IEEE Trans Multimedia, vol. 24, pp. 3545–3557, 2022, doi: 10.1109/TMM.2021.3102433.

[27] Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," CoRR, vol. abs/1706.08612, 2017, [Online]. Available: http://arxiv.org/abs/1706.08612

[28] Sarode, S., Thatte, R., Toshniwal, K., Warade, J., Bidwe, R. V., & Zope, B. (2023, March). A System for Language Translation using Sequence-to-sequence Learning based Encoder. In 2023

International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-5). IEEE.

[29] Mane, Deepak, et al. "An Improved Transfer Learning Approach for Classification of Types of Cancer." Traitement du Signal 39.6 (2022): 2095.

[30] Khetani, V., Gandhi, Y., Bhattacharya, S., Ajani, S. N., & Limkar, S. (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253–262.