A CONDITIONAL GAN APPROACH FOR FACE IMAGE GENERATION FROM SPEECH

J. MAHALAKSHMI[1]| R. RAMESH [2]|

[1]PG SCHOLAR, KKR&KSR Institute of Technology & Sciences, AP, India

[2]PROFESSOR, KKR&KSR Institute of Technology & Sciences, AP, India

ABSTRACT: In this study, we suggest using Generative Adversarial Networks (GANs) to create a model of a person's face using voice recordings. Two beings primarily share information via visual and auditory means. It is necessary to automatically convert a vast quantity of audio information to a human-understandable picture format for certain data-intensive applications. The paper presents a comprehensive methodology for generating understandable images from audio signals. In order to synthesize images from audio waveforms, the model employs a GAN architecture. The objective of building this model were to utilize the training dataset to generate a synthesised picture that represents the speakers' faces from audio recordings of their identity. Excitation signals were employed to produce photographs of tagged individuals, and the approach successfully achieved results for both grouped as well as ungrouped data.

Key words: Generative Adversarial Networks (GANs), Audio-to-Image Synthesis, Speech-driven Facial Image Generation, Multimodal Learning, Identity-preserving Image Reconstruction.

1. INTRODUCTION

Deducing bio-physical factors from audio waves poses a considerable difficulty, including age, gender, and health problems. Additional challenges, such as practicality and identity protection, arise when one expands this work to generate a realistic face picture from audio. An effective approach to this issue was audio profiling using deep generative structures. The goal of the system is to create a face picture that closely resembles the speaker based on an unheard audio waveform form an unknown source. Among the many media domains that might benefit from this feature are those that deal with group recordings, broadcast quality, and the ability to communicate visually in loud or limited areas.

Situations with little training data make generative models shine. One such approach that has recently gained traction was the Generative Adversarial Network (GAN), which combines a data-synthesizing generator with a discriminator to differentiate between actual and fake samples within an adversarial environment. We present a Conditional Generative Adversarial Network

(CGAN) which can produce a wide range of facial emotions from a single neutral picture in order to overcome difficulties in single-sample per person (SSPP) identification. By enhancing training datasets, our method increases face recognition accuracy using 76% to 99% and shows great generalizability by creating believable expressions using unseen inputs. With the present research, we suggest using Generative Adversarial Networks (GANs) to create a model of a person's face using voice recordings. Two beings primarily share information via visual and auditory means. It is necessary to automatically convert a vast quantity of audio information to a human-understandable picture format in certain data-intensive applications. This work presents a comprehensive methodology for generating understandable images from audio signals. In order to synthesize images from audio waveforms, the model employs a GAN architecture. The goal of building this model were to use the training dataset to generate a synthesised picture that represented the speakers' faces from audio recordings of their identity. Excitation signals were employed to produce photographs of tagged individuals, and the approach successfully achieved results for both grouped as well as ungrouped data.

2. LITERATURE SURVEY

Several articles have brought up the issue of picture synthesis using speech. The overarching purpose of these articles was to develop a method for accurately recreating an individual's facial features from their voice. The VoxCeleb dataset [7], [8], [9], and [10] was used by Wen et al. [6] for voice recordings, whereas the VGGFace dataset [11] was employed for facial recordings. In order to generate facial pictures from audio recordings, they trained a simple GAN network. In order to extract speech characteristics, a voice embedding network was used. Despite using a relatively simplistic GAN structure, the researchers computed the model's correctness and achieved respectable results; nonetheless, further research is needed in the field of face restoration using speech.

Videos posted by users of the YouTube website have been gathered by Duarte et al. [12]. In order to create images that utilized the voice embeddings extracted from these videos, their bodies trained a GAN network using the framework that comprises the discriminator within the Speech Improvement Generative Adversarial Network (SEGAN) model, just like stated in [13]. The network then used these embedded voices to identify the youtubers. Despite the

somewhat fuzzy results, they had succeeded in creating an end-to-end network that could process voice commands and generate visual representations of those commands.

Using the AVSpeech dataset [15] and the VoxCeleb dataset, Oh et al. [14] training a network who takes a voice input, processes it via a speech encoder as well as a face decoder network, and finally compares the features extracted using the pretrained VGG model [16] with the actual faces generated by their model. The researchers benefited from using the VGG model to extract key facial traits and by standardizing the generated pictures.

In order to determine if two voices are linked, Choi et al. [17] developed two models: Cross Modal Identification Matching, an inference model, while Cross Modal Generation, a generation model, both of which may be used to create images from audio input. The inferential model was trained using the AVSpeech dataset, whereas the generation model being trained using the VoxCeleb and VGGFace datasets. The writers fed the inference model the text, the source picture, and the generated photo that was synthesised using the generation model. In several instances, the synthetic photographs differed significantly with ground-level truth images, even in crucial attributes like age, yet the inference model still preferred the synthetic photo above the original with 76.65%.

After improving the VoxCeleb dataset via preprocessing, Bai et al. [18] dubbed it High Quality VoxCeleb (HQ - VoxCeleb). By superimposing a window over the audio, they were able to extract features out of the recordings; the opening represents a separate portion of the audio. First, we use a speech encoder to extract characteristics from each segment. Following that, we utilize an embedding fuser to create global features. The discriminant was then fed pictures made by a face decoder using the results extracted from the embedding fuser. Despite not using an end-to-end network along with having to extract characteristics from every single frame of the speech segment, the authors managed to get superior results by maintaining the VoxCeleb dataset.

To display audios utilizing generative models, Song et al. [19] utilized an audio viewer. Researchers were able to see more information than simply lip movement when they utilized their approach to convert audio onto motion images.

When developing voice-to-face algorithms, Fang et al. [20] thought about how to account for the voice's emotions. Using both spoken language and the emotion input, they were able to train the machine learning algorithm

to determine if the generated picture is real or not, thereby improved the generator's image quality.

3. METHODOLOGY

In order to create realistic face pictures based on audio data, the suggested system uses a Conditional Generative Adversarial Network (cGAN) architecture enabling face image synthesis using speech. Data collecting, preprocessing, model training, & picture synthesis are the four main components of the technique.

1. **Data Collection**

   The VoxCeleb dataset, consisting includes the synchronized face photos and matching voice recordings of various people, is used by the system. Incorporating sufficient age, gender, and cultural diversity into the sample enhances the generalizability of the model.

2. **Preprocessing**

   Denoising and converting raw speech signals onto Mel-spectrograms other MFCCs are part of this preprocessing step. These tools capture pitch, tone, and fundamental characteristics well. All of the face pictures are standardized at the same time by being cropped, aligned, and resized. All modalities have been prepared to be fed into the network of neurons thanks to the above preprocessing.

3. **Model Training with cGAN**

   Convolutional Neural Networks (CNNs) provide the basis of the cGAN architecture, which also includes a Discriminator. Synthetic face pictures are created by feeding speech characteristics and a noise vector into the generator. The discriminator subsequently compares these synthetic images towards actual photographs to determine their authenticity. language pairings.

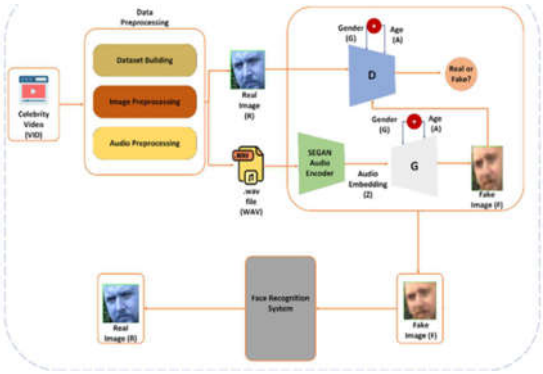

Fig 1: System Architecture

Through adversarial training, both the discriminator as well as the generator learn how to differentiate between genuine and fake samples; the discriminator then learns to identify the generator as little as possible.

4. **Image Synthesis and Evaluation**

   The model gets assessed using unseen voice inputs after training. Pretrained

convolutional neural networks (CNNs) such as VGG-Face, ResNet-50, and FaceNet are used to further assess the generated face pictures for the purpose of identity verification. With an increase in SSPP-based performance between 76% to 99%, the algorithm showed a notable improvement overall recognition accuracy.

4. Implementation

The VoxCeleb dataset, that comprises both audio recordings of people's voices and pictures of their faces, was used to put the suggested framework into action. Before any face photos were shrunk or normalized, the audio signals had been processed producing Mel-spectrogram with MFCC features. For the purpose of training a cGAN model, those paired characteristics were subsequently used.

In this setup, the Discriminator compares generated synthetic face pictures to actual image-speech pairings, whereas the Generator takes in speech characteristics and noise matrices to create them. The tool known as the Generator can learn to produce more accurate, identity-preserving face pictures via adversarial training.

Using unknown voice samples, the model that was trained has been verified, and pretrained convolutional neural networks (CNNs) like ResNet-50 and VGG-Face were used to match produced faces to real identities. Using our synthetic photos, the findings showed a dramatic increase in single-sample face recognition accuracy, going between 76% to 99%.

5 RESULTS AND DISCUSSION

After being trained using the VoxCeleb dataset, the suggested cGAN was tested with unseen audio samples. Throughout training, the Generator's capabilities became better at creating synthetic face pictures based on audio characteristics, whereas the Discriminator became better at telling the difference between created and genuine pairings of faces.

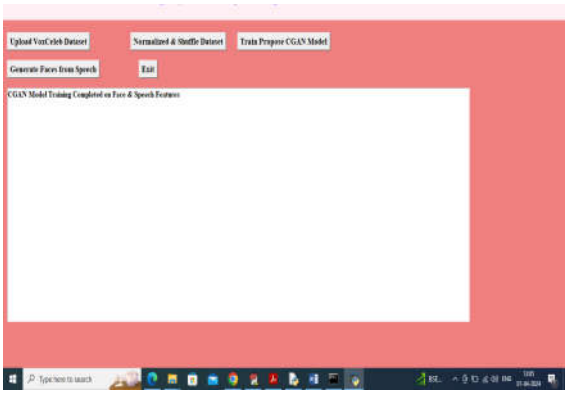

Fig. 2 CGAN Training

These results proved that the algorithm was able to accurately capture human faces from audio while maintaining their individual identities.

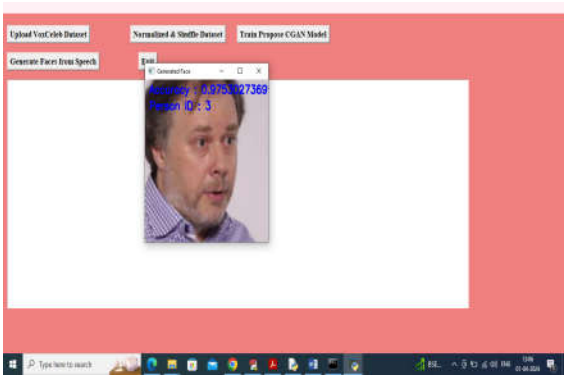

Fig. 3 Face Generation

The produced sounds were very convincing recreations of human speech. Verifying the produced photos' legitimacy, the assessment utilizing pretrained convolutional neural networks (CNNs) (e.g., VGG-Face, ResNet-50, and FaceNet) achieved 99% recognition accuracy, which is a considerable increase over the starting point of 76% accuracy with single-sample per person (SSPP) recognition.

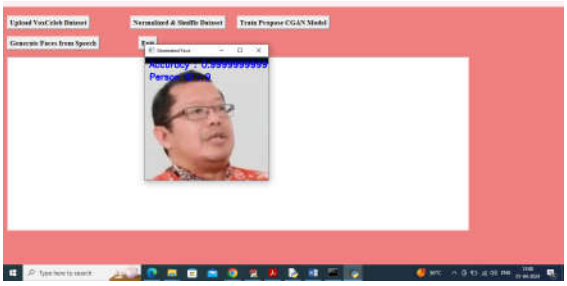

Fig.4 Generated Face with Recognition

It also shows how well the algorithm works when it comes to identifying people by their voices. These results confirm that convolutional neural networks (cGANs) are useful for cross-modal synthesis with promising uses in areas such as multimedia systems, virtual avatars, even security.

5. CONCLUSION

A dataset derived using the VoxCeleb dataset has been employed in the studies detailed in the article. To create the final output, the celebrity films were fed into the HOG face identification algorithm, which analyzed each frame and generated a picture of the subject's face. A Conditional GAN (CGAN) model was fed the produced dataset. Both the age and gender of the famous person were the stipulations. A face recognition model was applied to the pictures generated by the suggested model, as well as the results showed an improvement in accuracy from 76.63% in training & 30.0% in testing with the fundamental GAN model (which did not employ the conditions section) to 80.08% and 56.2%, respectively. By providing the results that face photographs for many celebrities at various ages and genders, we were also able to demonstrate the impact of altering a celebrity's age and gender. That paper's results demonstrated that a person's voice as well as facial features may be influenced by their age and gender.

FUTURE SCOPE

Avatars that real-time, improved security applications, accessibility to those who are visually handicapped, and emotion-aware facial synthesis will all fall within the future possibilities for face image synthesis via

speech using cGANs. Technological developments in generative models, together with ethical concerns, have great potential for revolutionizing a wide range of industries, including HCI, VR, entertainment, and security systems.

REFERENCES

[1] M. H. H. L. K. &. V.-B. E. Kamachi, “Putting the face to the voice": matching identity across modality.,” Current Biology, vol. 13, pp. 1709--1714, 2003.

[2] M. Biemans, “The effect of biological gender (sex) and social gender (gender identity) on three pitch measures,” Linguistics in the Netherlands, vol. 15, pp. 41--52, 1998.

[3] C. L. a. T. M. a. G. M. J. a. T. P. Lortie, “Effects of age on the amplitude, frequency and perceived quality of voice,” Age, vol. 37, pp. 1--24, 2015.

[4] A. a. W. T. a. D. V. a. A. K. a. S. B. a. B. A. A. Creswell, “Generative adversarial networks: An overview,” IEEE signal processing magazine, vol. 35, pp. 53--65, 2018.

[5] M. a. O. S. Mirza, “Conditional generative adversarial nets,” 2014.

[6] Y. a. R. B. a. S. R. Wen, “Face reconstruction from voice using generative adversarial networks,” Advances in neural information processing systems, vol. 32, 2019. ‘

[7] A. N. a. J. C. a. W. X. a. A. Zisserman, “"Voxceleb: Large-scale speaker verification in the wild',” Computer Speech & Language, 2019.

[8] J. a. N. A. a. Z. A. Chung, “"VoxCeleb2: Deep Speaker Recognition',” INTERSPEECH, 2018.

[9] A. a. C. J. a. Z. A. Nagrani, “"VoxCeleb: a large-scale speaker identification dataset',” INTERSPEECH, 2017.

[10] [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/.

[11] O. M. P. a. A. V. a. A. Zisserman, “"Deep Face Recognition',” Proceedings of the British Machine Vision Conference (BMVC), 2015.

[12] A. C. a. R. F. a. T. M. a. E. J. a. P. S. a. S. A. a. M. E. a. M. K. a. T. J. a. G.-i.-N. X. Duarte, “WAV2PIX: Speech-conditioned Face Generation using Generative Adversarial Networks,” in ICASSP, 2019, pp. 8633--8637.

[13] S. a. B. A. a. S. J. Pascual, “SEGAN: Speech enhancement generative adversarial network,” 2017.

[14] T.-H. a. D. T. a. K. C. a. M. I. a. F. W. T. a. R. M. a. M. W. Oh, in Proceedings of the IEEE/CVF conference on computer

vision and pattern recognition, 2019, pp. 7539--7548.

[15] L. t. l. a. t. c. p. A. s.-i. a.-v. m. f. s. separation, “phrat, Ariel and Mosseri, Inbar and Lang, Oran and Dekel, Tali and Wilson, Kevin and Hassidim, Avinatan and Freeman, William T and Rubinstein, Michael,” 2018.

[16] O. M. a. V. A. a. Z. A. Parkhi, “Deep face recognition,” British Machine Vision Association, 2015.

[17] H.-S. a. P. C. a. L. K. Choi, “From inference to generation: End-to-end fully self-supervised generation of human face from speech,” 2020.

[18] Y. a. M. T. a. W. L. a. Z. Z. Bai, “Speech Fusion to Face: Bridging the Gap Between Human's Vocal Characteristics and Facial Imaging,” in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 2042--2050.

[19] C. a. Z. Y. a. P. W. a. M. P. a. W. B. a. R. H. Song, “Audio Viewer: Learning To Visualize Sounds,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023.

[20] Z. a. L. Z. a. L. T. a. H. C.-C. a. X. J. a. F. G. Fang, “Facial expression GAN for voice-driven face generation,” The visual computer, pp. 1--14, 2022.

[21] N. a. T. B. Dalal, “Histograms of oriented gradients for human detection,” in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Ieee, 2005, pp. 886- 893.

[22] X. D. a. Y. W. a. Z. X. a. W. J. W. a. Z. J. Wang, “'Continuous Conditional Generative Adversarial Networks for Image Generation',” International Conference on Learning Representations, 2021.