International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 02 | Feb – 2025 DOI: 10.55041/ISJEM02266
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Deep Face Gen: Speech-Driven Face Image Synthesis

P. Kamakshi Thai [1], P. Manisha [2], L. Abhigyna Reddy [3] and M. Nagendhar Reddy [4]

[1] *Assistant Professor of Department of CSE(AI&ML) of ACE Engineering College.*

[2,3,4] *Students of Department CSE(AI&ML) of ACE Engineering College.*

Abstract:

A framework based on Generative Adversarial Networks (GANs) is proposed to synthesize facial images from audio inputs. The system aims to automatically translate large volumes of audio into understandable facial images without human intervention. By using a GAN architecture, the model generates image features from audio waveforms to reconstruct facial images. It is trained on a dataset of labeled examples, producing facial images corresponding to the identities of the speakers. The method achieves an accuracy of 96.88% for ungrouped data and 93.91% for grouped data. This approach demonstrates its capability to generate accurate facial representations from audio, offering an automated solution for converting speech into intelligible visual data.

Keywords: Generative Adversarial Networks, facial image synthesis, audio-to-image, speech-to-visual, automated reconstruction.

Introduction:

Audio profiling aims to determine a person's physical traits like age, gender, and health conditions from their voice. The challenge is to generate a realistic face image from an unknown individual's audio. This technology has various applications, such as enhancing audio in recordings, enabling video chats in quiet or noisy environments, and improving media content. A major issue in speech generation is the loss of pitch information, reducing clarity. The proposed model overcomes this by using a learning-based approach that captures detailed audio features more effectively than traditional methods, ensuring better pitch and sound quality.

Literature Survey:

1. Speech2Face:

This study explores how much a person’s facial features can be inferred from their voice. A deep neural network is trained on YouTube videos to learn the relationship between speech and facial attributes like age, gender, and ethnicity. The model generates face images from audio without explicitly modeling these traits.

2. Seeing Voices & Hearing Faces:

This research investigates whether a person’s face can be identified from just their voice. Using deep learning on datasets like VGGFace (faces) and VoxCeleb (voices), the model successfully matches voices to faces, performing better than humans in complex cases (e.g., same age and gender).

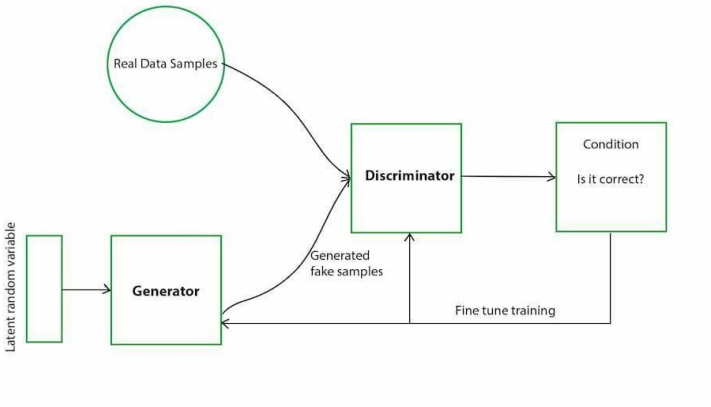
3. Learnable Pins:

This study creates a shared representation for faces and voices, allowing retrieval of a person’s face from their voice and vice versa. The model learns without labeled data and improves using a method called "hard negative mining." It is tested on unseen identities and applied to TV character identification.

4. Conditional C-GAN for Face Generation:

This model improves face image generation by using an attribute-guided approach. Given a low- resolution face and attribute details (e.g., gender, makeup), it generates a high-quality image. It supports applications like face super-resolution, face swapping, and frontal face generation while maintaining identity.

GAN Architecture:



Generative Adversarial Network (GAN) Architecture

Generative Adversarial Networks **(GAN)** is a deep learning framework that is used to generate random, plausible examples based on our needs. It contains two essential parts that are always competing against each other in a repetitive process (as adversaries). These two essential parts are:

Generator Network:

It is the neural network responsible for creating (or generating) new data. They can be in the form of an image, text, video, sound, etc., as per the data they are trained on.

Discriminator Network:

It's work is to distinguish between real and fake data from the dataset and data generated by the generator.

The responsibility of the generator model is to create new data real enough to "fool" the discriminator so that it cannot distinguish between real and fake (generated) data, whereas the role of the discriminator is to be able to identify if the data is generated or real data.

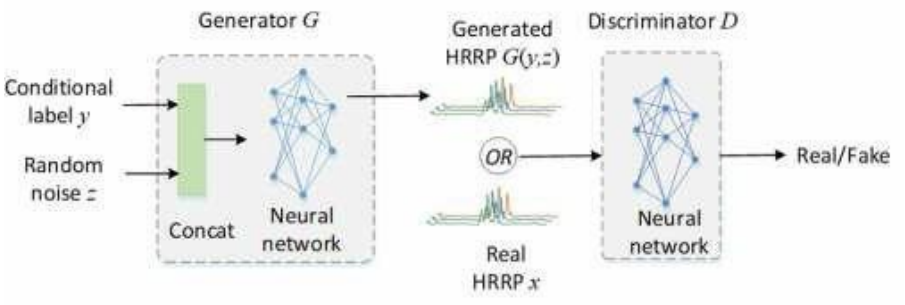
How GAN Works:

- The **Generator** takes random noise and produces synthetic samples.
- The **Discriminator** compares real and generated samples, learning to classify them correctly.
- Both networks compete: The **Generator** improves at creating realistic data, while the **Discriminator** gets better at detecting fakes.
- Over time, the **Generator** produces highly realistic outputs that can fool the **Discriminator**.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 02 | Feb – 2025 DOI: 10.55041/ISJEM02266
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Proposed Methodology:

CGAN Architecture:



Conditional Generative Adversarial Network (CGAN) Architecture

GANs can be extended to a conditional model by providing additional information (denoted as y) to both the generator and discriminator.

This additional information (y) can be any kind of auxiliary information, such as class labels or data from other modalities.

In the generator, the prior input noise (z) and y are combined in a joint hidden representation.

Generator Architecture:

The generator takes both the prior input noise (z) and the additional information (y) as inputs.

These inputs are combined in a joint hidden representation, and the generator produces synthetic samples. The

adversarial training framework allows flexibility in how this hidden representation is composed.

Discriminator Architecture:

The discriminator takes both real data (x) and the additional information (y) as inputs.

The discriminator's task is to distinguish between real data and synthetic data generated by the generator conditioned on y.

Conclusion:

In this Study, 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 VGG face 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," June 2019.

[2] V. B. Suneeta, P. Purushottam, K. Prashantkumar, S. Sachin and M. Supreet, Facial Expression Recognition Using Supervised Learning, 2020.

[3] S. Pavaskar and S. Budihal, Real-Time Vehicle Type Categorization and Character Extraction from the License Plates, 2019.

[4] A. Nagrani, S. Albanie and A. Zisserman, Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching, 2018.

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