Speech2Face Generator Poster

Faculty of computer and information Science – Ain Shams University
Scientific Computing Department

Authors : Abdelrahman Mohamed, Abdelrhman Yasser, Ahmed Magdy, Ahmed Samy, Sara Mohamed, Nourhan Mahmoud

Acknowledgements

Primarily, we would thank God for being able to complete this project with success. Then We would like to express our thanks and gratitude to our supervisor Dr. Manal Tantawy for her guidance understanding and support throughout the project.
Also, we would like to show our deep appreciation to T.A. Mirna Al-Shetairy and T.A. Aya Nasser as we were able to successfully complete this project with their help, support and patience.
Finally, we want to thank all our dear friends and colleagues for their valuable suggestions and help.

Abstract

Speech2Face is a technology that generates realistic images of human faces based on speech signals. Generative Adversarial Networks (GANs) are a promising approach to developing Speech2Face generators. GANs consist of two neural networks, a generator and a discriminator, which are trained together in a competitive game until the generator produces realistic samples that can fool the discriminator. Recent studies have shown that GAN-based Speech2Face generators can produce highly realistic facial images that resemble the speaker's actual face and emotional state. However, challenges remain, such as the need for large-scale datasets and efficient feature extraction methods. Despite these challenges, the potential applications of Speech2Face generators using GANs are vast, and the technology is rapidly advancing.

Introduction

Speech2Face is a technology that generates realistic images of human faces based on speech signals. Generative Adversarial Networks (GANs) are a promising approach to developing Speech2Face generators. GANs consist of two neural networks, a generator and a discriminator, which are trained together in a competitive game until the generator produces realistic samples that can fool the discriminator. Recent studies have shown that GAN-based Speech2Face generators can produce highly realistic facial images that resemble the speaker's actual face and emotional state. However, challenges remain, such as the need for large-scale datasets and efficient feature extraction methods. Despite these challenges, the potential applications of Speech2Face generators using GANs are vast, and the technology is rapidly advancing.

Methodology

1. **Datasets**: The Dataset consists of the voice recordings that are from the Voxceleb [1] dataset and the face images that are from the manually filtered version of VGGFace [2] dataset. Both datasets have identity labels. We use an intersection of the two datasets with the common identities, leading to 149,354 voice recordings and 139,572 face images of 1,225 persons. We use the whole dataset for training and testing on any recorded voice.
2. **Preprocessing**:
   a) Audio Data Preprocessing: involves using a voice activity detector from the WebRTC project to isolate speech-bearing regions of the recordings. 64-dimensional log mel-spectrograms are extracted using an analysis window of 25ms, with a hop of 10ms between frames, and each mel-frequency bin is mean and variance normalized. For training, an audio clip is randomly cropped around 3 to 8 seconds, while the entire recording is used for testing. If the audio is less than 10 seconds, it is repeated until it reaches 10 seconds. This preprocessing step helps to ensure that the input audio is in a suitable format for the subsequent processing and generation steps.
   b) Face Data Preprocessing: The cropped RGB face images of size 64 × 64 × 3 are obtained by similarity transformation. Each pixel in the RGB images is normalized by subtracting 127.5 and then dividing by 127.5.
3. **Build GANs Model**:
   In the General Model architecture[3] as shown in figure 1, a voice recording is used as input, and the pre-trained Voice Embedding Network extracts the voice embedding vector. The Generator then generates an image based on the voice embedding, and the Discriminator evaluates whether the generated image is real or fake by comparing it to a real image. The Classifier learns to assign any real face image to its identity label according to the loss function. The architecture of the model consists of several key components, including the Voice Embedding Network, the Face Embedding Network, the Generator, and the Discriminator (Classifier). The Voice Embedding Network consists of 1D convolutional layers with a kernel size of 3, and each convolutional layer is followed by a batch normalization layer and Rectified Linear Units (ReLU). The output is a 64-dimensional embedding. The Face Embedding Network uses a 2D convolutional layer with a kernel size of 1 as the input layer, followed by another convolutional layer with a kernel size of 4. Each convolutional layer is followed by a Leaky ReLU layer, and the final output is a 64-dimensional embedding. The Generator uses 2D convolutional Transpose layers with a kernel size of 4 to generate the image, with each layer followed by a Batch Normalization 1D layer and ReLU layer. The last layer is a 2D convolutional Transpose layer with a kernel size of 4, followed by a Tanh layer. The final output is a 64 × 64 × 3 image. The Discriminator (Classifier) takes the generated image as input and processes it through the Face Embedding Network to obtain a 64-dimensional embedding. The discriminator then classifies the embedding as either real or fake using a Sigmoid, which outputs either 0 or 1. Overall, this architecture enables the generation of realistic images based on a voice recording input and the evaluation of the quality of the generated images.
4. **Training the GANs**: The next step is to train the GANs using the preprocessed data. The generator network takes the speech features as input and generates a corresponding facial image, while the discriminator network evaluates the generated image and provides feedback to the generator network to improve the quality of the output as shown in figure 1. The two networks are trained together in a competitive game until the generator produces realistic samples that can fool the discriminator.
5. **Evaluation**: The final step is to evaluate the performance of the Speech2Face generator. This typically involves measuring the quality of the generated images using metrics

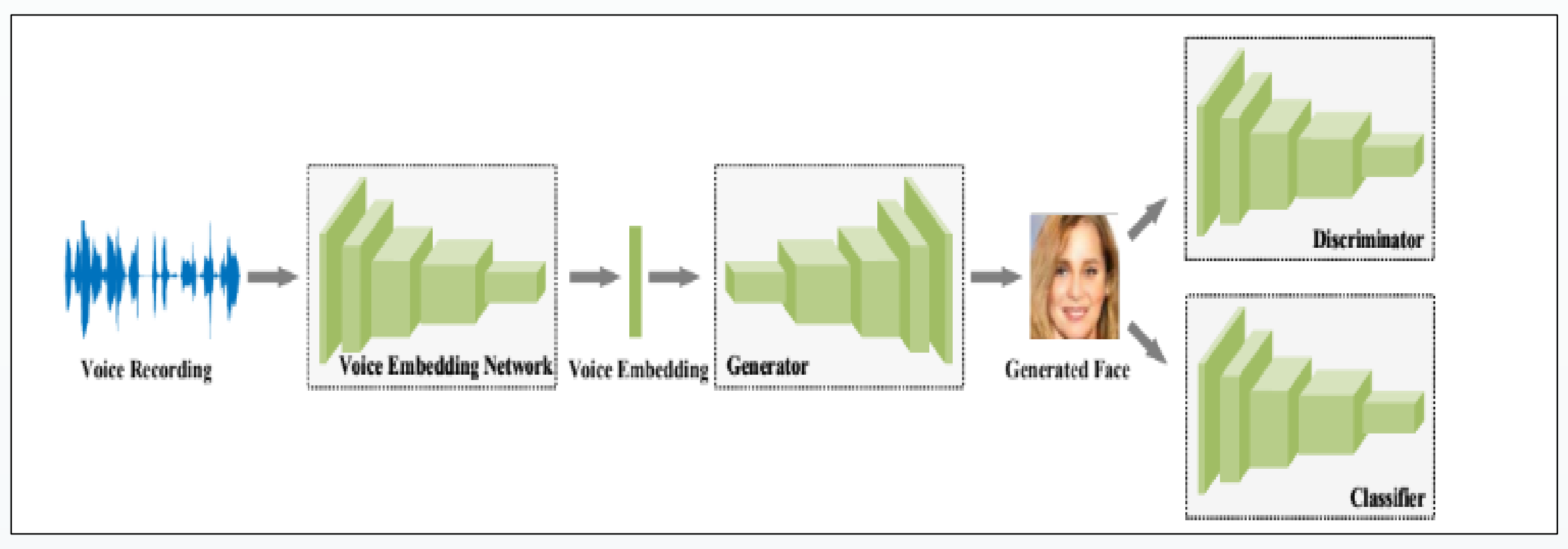

Figure 1. GANs General Architecture

Results

The study was conducted to evaluate the qualitative performance of different generative models in image generation tasks. Specifically, we aimed to compare the quality of the images generated by a general model, a female model, and a male model, in terms of their realism, diversity, coherence, and visual appeal as shown in the **Table 1**. **Qualitative Metric**.
For the quantitative results we evaluate our model with FID (Fréchet Inception Distance), L1 (mean absolute error), and Cosine Similarity as shown in **Table 2. Quantitative Metric**.

Table 1. **Qualitative Metric**

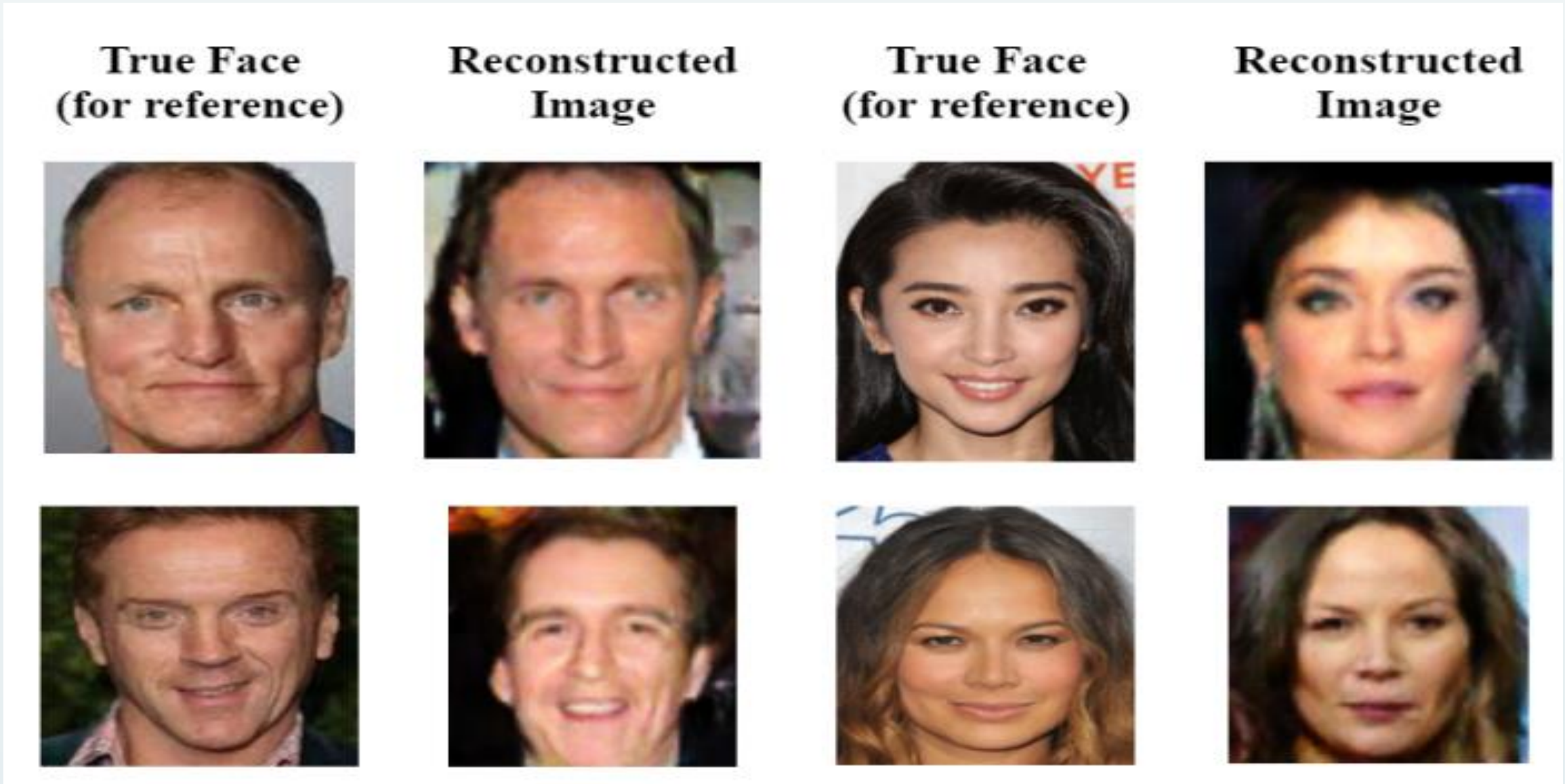

Table 2. **Quantitative Metric**

	FID	L1	Cos Similarity
General Model	**114.4445991**	61.9460271	0.244882
Females Model	126.365874	**29.41308**	**0.399265**
Males Model	129.303104	33.085962	0.2090025

Conclusion

In conclusion, Speech2Face technology using Generative Adversarial Networks (GANs) has shown great potential for generating lifelike images of human faces based on speech signals. GANs consist of two neural networks, a generator and a discriminator, which are trained together in a competitive game until the generator produces realistic samples that can fool the discriminator. Recent studies have shown that GAN-based Speech2Face generators can produce highly realistic facial images that closely resemble the speaker's actual face and emotional state, with applications in virtual assistants, speech therapy, and more. However, challenges remain, such as the need for larger datasets and efficient feature extraction methods. Despite these challenges, the technology is rapidly advancing and is poised to revolutionize the way we communicate and interact with each other. As more data becomes available and new techniques are developed, the quality and performance of Speech2Face generators using GANs are likely to continue improving, unlocking new possibilities for enhancing human communication and expression.

References

[1] Paul H Ptacek and Eric K Sander. Age recognition from voice. Journal of speech and hearing Research,9(2):273–277, 1966.

[2] Paul Mermelstein. Determination of the vocal-tract shape from measured formant frequencies. The Journal of the Acoustical Society of America, 41(5):1283–1294, 1967.

[3] Wen, Yandong, Bhiksha Raj, and Rita Singh. "Face reconstruction from voice using generative adversarial networks." Advances in neural information processing systems 32 (2019).