International Journal of Innovative Research in Engineering
Volume 5, Issue 1(January-feburary2024), PP: 16-19.
https://www.doi.org/10.59256/ijire.20240501004
www.theijire.com
ISSN No: 2582-8746

ML-Driven Facial Synthesis from Spoken Words Using Conditional GANs

Vaishnavi Srivastava[1], Sakshi Srivastava[2], Sakshi Chauhan[3], Divyakshi Yadav[4]
[1,2,3,4] B.Tech IT 4th Year, Institute of Technology & Management Gida Gorakhpur, Uttar Pradesh, India.

How to cite this paper:
Vaishnavi Srivastava[1], Sakshi Srivastava[2], Sakshi Chauhan[3], Divyakshi Yadav[4] "ML-Driven Facial Synthesis from Spoken Words Using ConditionalGANs", IJIRE-V5I01-16-19.

Abstract: *A Human Brain may translate a person's voice to its corresponding face image even if never seen before. Training adeeplearning network to do the same can be used in detecting human faces based on their voice, which may be used in findinga criminal that we only have a voice recording for. The goal in this paper is to build a Conditional Generative Adversarial Network that produces face images from human speeches which can then be recognized by a face recognition model to identifythe owner of the speech. The model was trained, and the face recognition model gave an accuracy of 80.08% in training and 56.2% in testing. Compared to the basic GAN model, this model has improved the results by about 30%.*

Key Word: *Face image synthesis, Generative adversarial network, Face Recognition*

I.INTRODUCTION

Humans' voices are strongly related to their faces, you could hear a person's voice and your brain starts to draw an image for his face based on some characteristics such as the tone of his voice, and frequency. A greatnumber of research mentioned such methods including [1] which had made an experiment to investigate whether the faces and voices of a person have the same identity information. This experiment included givingthe participants, either a voice to hear and two faces or a face and two voices, the participants had to decide which combination had the same identity. In the Face Voice experiment the matching performance was 54%and, in the Voice-Face experiment the matching performance was 55% which proves that the face is related to the voice of a person. There are also a lot of person's characteristics that can affect his voice, such as the age of a person and his gender. The voice of a person is directly affected by his gender, proved that there's a significant difference between the median pitch of men and women. The median pitch of women ishigher than men;(189 Hz) and (111 Hz) respectively. The age of a personal so affects his voice, a young 12 year old child's voice will be different from his voice at the age of 20 and different from his voice at the age of 70Infact, once a person growing old, his face changes and his voice changes as well. had all experiments done on young and old healthy adults, they proved that the frequency and amplitude are changed according to a person's age. In this paper the problem of generating human faces based on their speech is addressed, some researches were done in this area and gave convenient results but the proposed model will not only be given the voice of a person but also certain characteristics of this person such as his gender and his age. One of the main objectives of the current research paper is to show how the human face is affected by his voice. In addition, the face structure is changed by changing the age and the gender characteristics. A Generative Adversarial Network (GAN) will be used as the main model because this network will be provided by the ageand gender characteristics it will be a Conditional Generative Adversarial Network (CGAN). The proposed CGAN is the default GAN architecture that takes the age and gender as conditions. The idea of a GAN networkis training two competing models; these models are the generator and the discriminator which are normally two Convolution Neural Networks (CNNs). Creswell etal.[4] referred to the generator as an art forger that creates forgeries and aims to make realistic images, while the discriminator is referred to as an art expert thatshould be able to detect there a images from the forgeries produced by the generator.

Defined conditional generative adversarial networks as an extended version of the generative adversarial networks. This is done by giving the generator and the discriminator an extra input that can be any additionalinformation such as class labels, as shown in Fig. 1 the generator and the discriminator are represented by twoCNN models, each model has its own inputs and outputs, the generator takes as input the vector (Z) which isnormally noise and the discriminator takes as input either the real image (X) or the fake image produced by the generator (X'), the output of the discriminator will be either "Real" or "Fake".

1. Problems

The problem of image synthesis from speech has been stated in a number of papers. All these papers had thesame goal which is to produce a face image of a person from his speech that is close to his real face.
Some of the problems are listed below:

1. Facial Expression Generative- Developing a system that can generate realistic facial expressions based on spoken

words, accurately capturing the nuances and emotions conveyed through speech.

2. Synchronization with Speech- Ensuring that the generated facial animations are synchronized with the corresponding speech, creating a seamless and natural interaction between the spoken words andthe facial expressions.

3. Real-Time Generation- Designing a system that can generate facial animations in real-time, allowing for immediate response and interaction with the user.

4. Quality & Realism- Enhancing the overall quality and realism of the generated facial expressions, making them indistinguishable from human expressions and improving the user experience.

5. Application in Virtual Characters or Avatars- Enabling the integration of this technology into virtual characters or avatars, enhancing their expressions & making them more engaging and lifelike in various applications such as virtual reality gaming, or animation.

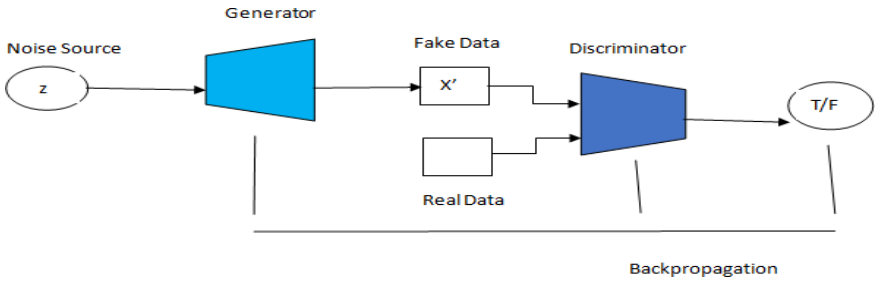

Fig. 1: The Generator (G) and the discriminator (D) models are trained in parallel to each other.

II.LITERATURE REVIEW

The problem of image synthesis from speech has been stated in a number of papers. All these papers had thesame goal which is to produce a face image of a person from his speech that is close to his real face. Wen et al. [6]used the voice recordings from Vox Celeb dataset [7], [8] , [9] , [10] and faces from the VGG Face dataset [11].They trained a very basic GAN network that takes voice recordings as input and produces face images based on these recordings. The features from speech were extracted using a voice embedding network. The authors calculated the model accuracy and it gave good results considering that there are not so many papers in the area of face reconstruction from speech but they used a very basic GAN structure that could have been improved. Duarte et al. collected videos from YouTube website that were uploaded by you tubers. They extracted the voices and faces of these you tubers from their videos and trained a GAN network that took a voice embedding produced from a speech encoder based on the structure of the discriminator in the Speech Enhancement Generative Adversarial Network (SEGAN) model stated in [13] the GAN network then produces images based on this voice embedding.Although the produced images were some how blurry but they had the advantage of designing an end to end network that takes speech and produces images. Oh et al. [14] used AV Speech dataset [15] and Vox Celeb datasetto train a network that extracts important face features from a pre trained VGG model [16] and compares these features to faces produced from their model which takes a speech and passes it by a speech encoder then a face decoder network. The authors had the advantage of getting the important features of the face from the VGG modeland also putting the produced images in a canonical form. Choietal.[17] designed two models; Cross Modal Identity Matching: An inference model used to detect if a given voice and speech are related and Cross Modal Generation: A generation model used to generate image from a given speech. AV Speech dataset was used to trainthe inference model while Vox Celeb and VGG Face datasets were used to train the generation model. The authorsused the inference model, gave it the speech, the original photo and the synthesized photo produced by the generation model. The inference model chose the synthesized photo over the original one by 76.65% but in somecases the synthesized images were very far fromthe ground truth image seven in some important features such as the age. Baietal.[18] applied some preprocessing on the Vox Celeb data set to increase the quality of images and named it High Quality Vox Celeb (HQ- Vox Celeb). They extracted features from the voice recordings by applyinga window over the speech, eachwindow is regarded as an individual speech segment. Features are extracted using voice encoder from each segment and der to produce images which will be then passed by the discriminator. Theauthors maintained the Vox Celeb dataset which gave better results but they didn't use an end to end network andthey had to extract the features from each window of the speech segment. Song et al. [19] used an audio viewer that is used to visualize audios using generative models. They used their model to change audios into motion pictures, they had the advantage of visualizing more information than lip motion. Fang et al. [20] considered taking into consideration the emotions of the voice in the voice to face generation. They had the advantage of training the discriminator to check if the produced image is true or fake not only by speech but also by the given emotion, which helps in making the generator produce better images.

M. H. H. L. K & V.-B. E. Kamachi studies that Speech perception provides compelling examples of a strong link betweenauditory and visual modalities. This link originates in the mechanics of speech production, which, in shaping the vocal tract, determine the movement of the face as well as the sound of the voice. In this paper, we present evidence that equivalent information about identity is available cross-modally from both the face and voice. Using a delayed matching to sample task, XAB, we show that people can match the video of an unfamiliar face, X, to an unfamiliar voice, A or B, and vice versa, but only when stimuli are moving and are played forward. The critical role of time-varying information is underlined by the ability to match faces to voices containing only the coarse spatial and temporal information provided by sine wave speech [5]. The effect of varying sentence content across modalities was small, showing that identity-

specific information isnot closely tied to particular utterances. We conclude that the physical constraints linking faces to voices result in bimodally available dynamic information, not only about what is being said, but also about who is saying it.

A.a.W.T.a.D.V.a.A.K.a.S.B.a.B.A.A.Creswell studies that Generative adversarial networks (GANs) pro-vide away to learndeep representations without extensively annotated training data. They achieve this through deriving back propagation signals through a competitive process involving a pair of networks. The representations that can be learned by GANs may be used in a variety of applications, including image synthesis, semantic image editing, style transfer image super-resolutionand classification. The aim of this review paper is to provide an overview of GANs for the signal processing community, drawing on familiar analogies and concepts where possible. In addition to identifying different methods for training and constructing GANs we also point to remaining challenges in their theory and application.

M. a. O. S. Mirza studies that Generative Adversarial Nets [8] were recently introduced as an ovel way to train generative models. In this work we introduce the conditional version of generative adversarial nets, which can be constructed by simplyfeeding the data, y, we wish to condition on to both the generator and discriminator. We show that this model can generate MNIST digits conditioned on class labels. We also illustrate how this model could be used to learn a multi-modal model, and provide preliminary examples of an application to image tagging in which we demonstrate how this approach can generate descriptive tags which are not part of training labels.

J. a. N. A. a. Z. A. Chung, studies that Speech Recognition builds a bridge between the multimedia streaming (audio- only, visual- only or audio-visual) and the corresponding text transcription. However, when training the specific model of new domain, it often gets stuck in the lack of new-domain utterances especially the labeled visual utterances. To break through this restriction, we attempt to achieve zero-shot modality transfer by maintaining the multi-modality alignment in phoneme space learned with unlabeled multimedia utterances in the high resource domain during the pre-training (Shietal.,2022) and propose a training system Open-modality Speech Recognition (Open SR) that enables the model strained on a single modality (e.g., audio-only) applicable to more modalities (e.g., visual-only and audio-visual).Furthermore, we employ a cluster-based prompt tuning strategy to handle the domain shift for the scenarios with only common words in the new domain utterances. We demonstrate that Open SR enables modality transfer from one to any in three different settings (zero-, few- and full- shot), and achieves highly competitive zero- shot performance comparedto the existing few- shot and full-shot lip-reading methods. To the best of our knowledge, Open SR achieves the state- of-the-art performance of word error rate in LRS2 on audio-visual speech recognition and lip-reading with 2.7% and 25.0%, respectively.

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, studies that Speech is a rich biometric signal that contains information about the identity, gender and emotional state of the speaker. In this work, weexplore its potential to generate face images of a speaker by conditioning a Generative Adversarial Network (GAN) withraw speech input. We propose a deep neural network that is trained from scratch in an end-to- end fashion, generating a face directly from the raw speech waveform without any additional identity information (e.g reference image or one-hot encoding). Our model is trained in a self-supervised approach by exploiting the audio and visual signals naturally aligned in videos. With the purpose of training from video data, we present a novel dataset collected for this work, with high-quality videos of you tubers with notable expressiveness in both the speech and visual signals.

S. a. B. A. a. S. J. Pascual studies that Traditional speech enhancement algorithms are only suitable for dealing withstationary noise, but the noise in the stage of flight is non stationary noise, so the traditional method is not suitable for dealing with the noise in the stage of flight. This paper proposes a speech enhancement algorithm based on a generative adversarial network: Deep Convolutional–Wasserstein Generative Adversarial Network (DWGAN).

Firstly, the model integrates the deep convolutional generative adversarial network and the Wasserstein distance based on the generative adversarial network. Secondly, it introduces a conditional model to improve the enhanced speech quality, and the spectral constraint layer is used to prevent the model from falling too fast and causing collapse. Finally, the L1 loss termis introduced into the loss function to reduce the number of training times and further improve the enhanced speech quality. The experimental results show that the intrusiveness of background noise and overall processed speech quality of DWGAN are improved by about 7.6 and 9.4%, respectively, comparedwith WGAN in the acoustic environment of simulated aircraft operation.

T.-H. a. D. T. a. K. C. a. M. I. a. F. W. T. a. R. M. a. M. W. Oh on computer vision and pattern recognition, studies that Scene text detection methods based on neural networks have emerged recently and have shown promising results. Previous methods trained with rigid word-level bounding boxes exhibit limitations in representing the text region inan arbitrary shape. In this paper, we propose a new scene text detection method to effectively detect text area by exploring each character and affinity between characters. To overcome the lack of individual character level annotations, our proposed framework exploits both the given character-level annotations for synthetic images and the estimated character-level ground-truths for real images acquired by the learned interim model. In order to estimate affinity between characters, the network is trained with the newly proposed representation for affinity. Extensive experiments on six benchmarks, including the Total Text and CTW-1500 datasets which contain highly curved texts in natural images, demonstrate that our character-level text detection significantly outperforms the state-of-the-art detectors. According to the results, our proposed method guarantees high flexibility in detecting complicated scene text images, such as arbitrarily-oriented, curved, or deformed texts.

O. M. a. V. A. a. Z. A. Parkhi, studies that Facial analysis systems are used in a variety of scenarios such as law enforcement, military, and daily life, which impact important aspects of our lives. With the onset of the deep learningera, neural networks are being widely used for the development of facial analysis systems. However, existing systems have been shown to yield disparate performance across different demographic subgroups. This has led to unfair outcomes for

certain members of society. With an aim to provide fair treatment in the face of diversity, it hasbecome imperative to study the biased behavior of systems. It is crucial that these systems do not discriminate basedon the gender, identity,skin tone, or ethnicity of individuals. In recent years, a section of the research community has started to focus on the fairness of such deep learning systems. In this work, we survey the research that has been done in the direction of analyzing fairness and the techniques used to mitigate bias. A taxonomy for the bias mitigation techniques is provided. We also discuss the databases proposed in the research community for studying bias and the relevant evaluation metrics. Lastly, we discuss the open challenges in the field of biased facial analysis.

H.-S. a. P. C. a. L. K. Choi studies that Face super-resolution (FSR), also known as face hallucination, which is aimed at enhancing the resolution of low-resolution (LR) face images to generate high-resolution face images, is a domain-specific image super-resolution problem. Recently, FSR has received considerable attention and witnessed dazzling advances with the development of deep learning techniques. To date, few summaries of the studies on the deep learning-based FSR are available. In this survey, we present a comprehensive review of deep learning-based FSR methods in a systematic manner. First, we summarize the problem formulation of FSR and introduce popularassessment metrics and loss functions. Second, we elaborate on the facial characteristics and popular datasets used in FSR. Third, we roughly categorize existing methods according to the utilization of facial characteristics. In each category, we start with a general description of design principles, present an overview of representative approaches, and then discuss the pros and cons among them. Fourth, we evaluate the performance of some state-of-the-art methods. Fifth, joint FSR and other tasks, and FSR-related applications are roughly introduced. Finally, we envision the prospects of further technological advancement in this field.

III.CONCLUSION

The experiments mentioned within the paper had used a dataset that was produced from the Vox Celeb dataset. The production was done through passing the videos of celebrities to the HOG face detection technique whichproduced a face image for each frame in the video. The generated dataset was used as an input to a Conditional GAN (CGAN) model. The conditions were the celebrity's age and gender. The images produced by the proposed model were passed by a face recognition model that gave an accuracy of 80.08% for training and 56.2% for testing compared to 76.63% for training and30.0% for testing in the basic GAN model that did not use the conditions part. The effect of changing the age and gender of a celebrity was also shown by giving theresults of face images for different celebrities at different age ranges and different genders. The findings of the paper proved that the age and gender of a person can affect his voice as well as his face.

References

1. *1.A.N.a.J.C.a.W.X. a.A. Zisserman, "'Vox celeb: Large-scale speaker verification in the wild" Computer Speech & Language, 2019. a. W. T. a. D. V. a. A. K. a. S. B. a. B. A. A. Creswell, "Generative adversarial networks: An overview" IEE signal processing magazine, vol. 35,pp.53—65,2018.*
2. *A.a.C.J.a.Z.A.Nagrani,"'Vox Celeb: alarge – scale speaker identification dataset',*
3. *C. L. a. T. M. a. G. M. J. a. T. P. Lortie, "Effects of age on the amplitude, frequency and perceived*
4. *voice,"Age, vol. 37, pp. 1--24, 2015.*
5. *M.a.O.S.Mirza,"Conditionalgenerativeadversarialnets,"2014.*

a. *M. H. H. L. K. &. V.-B. E. Kamachi, "Putting the face to the voice": matching identity*
6. *across modality.,"Biology, vol. 13, pp.1709--1714, 2003.*
7. *M. Biemans, "The effect of biological gender (sex) and social gender (gender identity) on three pitch*
8. *measures," Linguistics in the Netherlands, vol. 15,pp. 41--52, 1998.*