ISSN: 2249-7196
IJMRR/Sep. 2024/ Volume 14/Issue 7/92-99
Mrs. B. Vasantha / International Journal of Management Research & Review

THE USE OF CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS FOR THE SYNTHESIS OF FACE IMAGES FROM SPEECH

Mrs. B. Vasantha [1], Cheryala Rashmitha [2], Boppudi Sharvani Sharma [3], Erram Manisha [4]

[1] Assistant Professor, Department of Information and Technology, Malla Reddy Engineering College for Women, Autonomous, Hyderabad

[2],[3],[4] Student, Department of Information and Technology, Malla Reddy Engineering College for Women, Autonomous, Hyderabad, rashmithach399@gmail.com, sharvaniboppudi77@gmail.com, manishaerram2828@gmail.com

ABSTRACT: *To create an artificial portrait of a person's face using audio alone, this research suggests using Generative Adversarial Networks (GANs). Two beings primarily share information via visual and auditory means. A huge quantity of audio has to be automatically translated into a comprehensible picture format in some data-intensive applications, with no human intervention required. This work presents a comprehensive methodology for generating understandable images from audio signals. The model's generative adversarial network (GAN) architecture synthesizes images from audio waveforms. The goal of building this model was to use the training dataset to generate a synthesised picture of the speakers' faces from audio recordings of their identity. The approach achieved a 96.88% accuracy rate for ungrouped data and a 93.91% accuracy rate for grouped data by using excitation signals to produce pictures of tagged humans.*

INTRODUCTION

One of the issues being discussed is the use of audio processing to infer bio-physical factors such as gender, age, and other health problems from audio recordings. A number of challenges arise from this work, including whether or not it is feasible to acquire knowledge an individual's somatic attributes and whether or not it is possible to synthesize the whole face from audio alone. This question is answered constructively by audio prototyping using our framework. The aim is to create a facial picture that has many matches with the speaker's identity provided an unheard audio waves from a source that is not known. As an example, it enables video-visiting in

peaceful areas like libraries in addition to boisterous settings, and it improves the sound quality of existing records in groups where someone is talking, such as recordings of blogs or news bondage. Discourse models lose pitch data and replicate speech with poor coherence, according to the investigation's suggested test of discourse age. The suggested approach takes into account the sound aging process and its high-dimensional sound-related spectrogram using a comprehensive learning-based model. The frequency and reverberation data of sound are quantified at each time step by a 128-dimensional element, in contrast to the current models LPC, LSP, Customary Spectrogram, and sound-related spectrogram jells. If the audio is 8KHz, the sound-related spectrogram will take into account a lot of factors. The system also has a hard time becoming intimately acquainted with the sound-related spectrogram because of how strongly linked these spectrograms are.

RELATED WORK

"Speech2Face: Learning the Face Behind a Voice,"

To what extent can one extrapolate an individual's appearance from their speech patterns? Reconstructing a person's face from a brief audio clip of them speaking is the subject of our investigation in this article. To do this, we use millions of user-generated recordings of people talking on YouTube and the Internet to build and train a neural network that is deep. In order to generate pictures that capture speakers' age, gender, and ethnicity, among other physical aspects, our model learns voice-face connections during training. Without explicitly modeling traits, this is accomplished in a self-supervised way by capitalizing on the presumed co-occurrence of faces and words in Internet videos. We mathematically measure the degree to which our Speech2Face reconstructions, derived from audio, match the speakers' actual facial pictures and assess their quality.

"Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching,"

Determine which of two facial photos represents the speaker from a brief audio clip is an apparently insurmountable challenge that we present here. How much can we deduce about the face from the speech, and how can the voice affect the face? This and other similar cross-modal tasks are investigated in this work. We investigate this problem "in the wild" using the publicly accessible datasets for speaker identification using audio (VoxCeleb) and face recognition using

static photos (VGGFace). You may use them to train your model for static and dynamic cross-modal matching tests. First, we present convolutional neural network (CNN) architectures for binary along with multi-way cross-modal face or audio matching. Second, we contrast static testing with dynamic testing, where only one still picture is available but the audio doesn't come from the same video. Finally, we calibrate the complexity of the assignment using human testing as a baseline. In both static and dynamic settings, we demonstrate that a convolutional neural network (CNN) can be trained to do this job, and it even outperforms chance on 10-way face-from-voice classification. The CNN's performance is on par with human levels for simple instances (such as faces of various genders), but it surpasses human levels for more complex examples (such as faces of the same gender, age, and country).

"Learnable pins: Cross-modal embeddings for person identity,"

We suggest and study a face-voice hybrid embedding that is sensitive to user identification. With this embedding, speech-to-face and face-to-voice cross-modal retrieval are both made possible. Our four contributions are as follows: first, we demonstrate that embedding can be learned from talking face videos without identity labels using a cross-modal self-supervision method; second, we create a curriculum discovering schedule to feed hard negative mining that is critical for this task's successful learning; third, we show how to assess cross-modal retrieval for unseen and unheard identities during training across multiple scenarios and set a benchmark for this new task; and at last, we indicate an instance for applying the joint embedding to feed automatically locating and labeling characters in TV dramas.

"Conditional C-GAN for attribute guided face image generation,"

Our research focuses on attribute-guided face creation, whereby a high-resolution (resolution) face image (resolution) is created from a low-resolution (resolution) input picture (input image) by using an attribute vector retrieved from a high-resolution (resolution) image. We circumstance the CycleGAN and suggest conditional CycleGAN to solve this issue. It can 1) handle uncoupled training data, which is common given that the training low/high-res as well as high-res attribute images aren't always in sync, and 2) make it easy to control the generated face's appearance using the input attributes. We showcase top-notch outcomes on the attribute-guided conditionally CycleGAN, which is capable of creating lifelike face pictures using user-supplied factors (such as gender, makeup, hair color, and eyeglasses) to effortlessly manage look.

Transforming the attribute-guided network into the identity-guided conditional CycleGAN—which yields intriguing and high-quality findings on identity transfer—involves using the attribute image as an identity to generate the associated conditional vector and integrating a face verification network. Using identity guided conditional CycleGAN, we showcase three applications: face swapping, frontal face creation, and identity-preserving face super resolution. These use cases consistently highlight the benefits of our innovative technique.

"Age estimation from telephone speech using i-vectors,"

via use of i-vectors,

This research presents a novel method for age estimate from telephone voice patterns that is based on i-vectors. The recommendation is driven by the efficiency of i-vectors in speaker identification. Every word is represented by its i-vector in this approach. After that, we use SVR (Support Vector Re to guess how old the speakers are. The Speaker Recognition Evaluations datasets maintained by the National Institute of Standards and Technology (NIST) in 2010 and 2008 were used to train and test the suggested approach. When it comes to estimating the age of speakers, the evaluation findings reveal that the suggested technique works better than several traditional methods.

METHODOLOGY

We have developed the following modules to carry out this project.

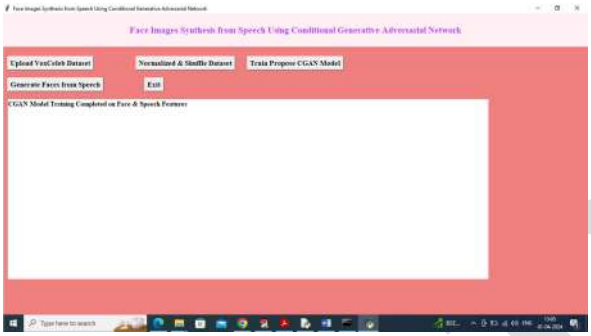
1) VoxCeleb Dataset Upload: This module allows you to upload a dataset that includes both faces and audio.

2) Standardized and Shuffled Dataset: This section will take facial and voice characteristics, normalize them, subsequently shuffle and reorder them.

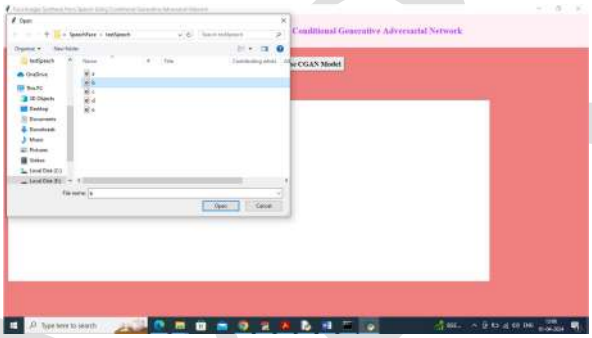
3. Train the CGAN Model: The CGAN technique will be used to train a model using the processed features.

4) Faces from voice: This module allows us to upload test voice files, extract speech characteristics, and then feed them into a convolutional neural network (CGN) algorithm to create a face. Then, we feed that face into another algorithm to identify a person, and finally, we show the accuracy of the face recognition.

RESULT AND DISCUSSION

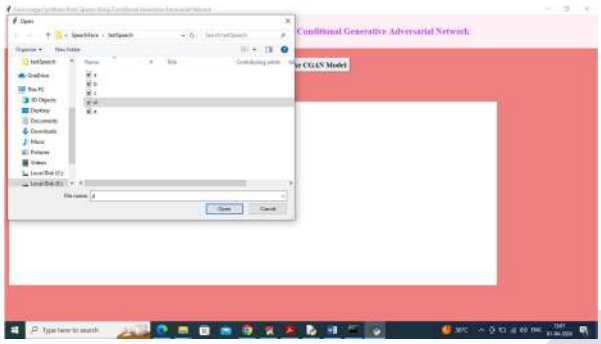


In above result CGAN model training completed and now click on 'Generate Faces from Speech' button to upload test speech and generate faces



In above result selecting and uploading test audio speech file and then click on 'Open' button to get below output

In above result can see generated face from test speech file and can see recognized person id along with recognition accuracy. Similarly you can upload and test other speech files



In above result uploading another speech file

In above result can see generated face along with recognition

CONCLUSION

Employing a GAN architecture that automatically acquires key visual cues, our study examined the possibility of face reconstruction from audio. We used the voxceleb2 comprising VGGface datasets, which include 1,06,584 face pictures from 924 unique individuals, and 113,322 voice segments, to train the network. To ensure accuracy and performance, we trained and tested the algorithm with additional speakers to cross-check its performance. In the final layers, we tweaked the activation function to verify it. The model's ability to 96.88% accurately synthesize the faces and their accompanying audio signals is the most significant outcome of the article.

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," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7531-7540, June 2019.

[2] T. Dekel, C. Kim, I. Mosseri, W. T. Freeman and M. Rubinstein, Speech2face: Learning the face behind a voice, IEEE conference on computer vision and pattern recognition, 2019.

[3] S. Pavaskar and S. Budihal, Real-Time Vehicle Type Categorization and Character Extraction from the License Plates, Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol. 768, 2019.

[4] A. Nagrani, S. Albanie and A. Zisserman, Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8427-8436, 2018.

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

[6] Y. Lu, Y.-W. Tai and C.-K. Tang, Conditional C-GAN for attribute guided face image generation, ,2017.

[7] Mohamad Hasan Bahari, ML McLaren, DA Van Leeuwen, et al. Age estimation from telephone speech using i-vectors, 2012.

[8] Y. Wen, M. A. Ismail, W. Liu, B. Raj and R. Singh, Disjoint mapping network for cross modal matching of voices and faces, International Conference on Learning Representations ICLR 2019, pp. 1-17, 2019.

[9] J. Bao, D. Chen, F. Wen, H. Li and G. Hua, Towards Open-Set Identity Preserving Face Synthesis, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6713-6722, 2018.

[10] V. B. Suneeta, P. Purushottam, K. Prashantkumar, S. Sachin and M. Supreet, Facial Expression Recognition Using Supervised Learning, Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol. 1108, pp. 275-285, 2020.

[11] A. Jamaludin, J. S. Chung and A. Zisserman, You said that?: Synthesising talking faces from audio:, International Journal of Computer Vision, pages 01-13, 2019.

[12] R. Singh, Reconstruction of the human persona in 3D and its reverse, In Proling Humans from their Voice, chapter 10. springer nature Press, 2020.

[13] R. Huang, S. Zhang, T. Li and R. He, Beyond face rotation: Global and local perception GAN for photo realistic and identity preserving frontal view synthesis, In Proceedings of the IEEE International Conference on Computer Vision, pp. 2439-2448, 2017.
[14] J. Thies, M. Elgharib, A. Tewari, C. Theobalt and M. Nieÿner, Neural Voice Puppetry: Audio Driven Facial Reenactment, Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 12361, 2020.
[15] R. Zhang, P. Isola and A. A. Efros, Colorful image colorization, In European conference on computer vision, pp. 649-666, 2016.

Mrs. B. Vasantha

"Mrs B Vasantha received B.Tech Degree(INFORMATION TECHNOLOGY)in SRI VASAVI Engineering in 2011. she received MASTER OF ENGINEERING in (COMPUTER SCIENCE AND ENGINEERING) Degree from SASI INSTITUTE OF TECHNOLOGY & Engineering in 2021. She is having Academic Experience of more than 2 years. She is associated with MRECW Her current area of research includes Machine learning. She is having 2 papers in reputed International Journals. Attended Various Workshops and Faculty Development Programs
"