Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 3, June 2026, pp. 902∼912
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i3.pp902-912

Voice portraits: building faces through voice analysis

Anandhu T. G., John K. Joseph, Navneeth Krishnan J., Richu Shibu, Elizabeth Isaac
Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Ernakulam, India

Article Info

Article history:

Received Jul 16, 2025
Revised Feb 22, 2026
Accepted Mar 4, 2026

Keywords:

Biometric identification
Cross-modal learning
Deep learning
Facial reconstruction
Generative adversarial networks
Speech analysis
Voice-to-face generation

ABSTRACT

Generation of a person's appearance from their voice alone is an intriguing challenge. The proposed framework centers on recreating a person's facial image based solely on a short audio recording of that person speaking. Using a deep neural network trained on millions of YouTube recordings where faces and voices appear together, the system learns voice-face relationships, enabling it to generate images that capture physical traits such as age, gender, and ethnicity. Operating in a self-supervised manner, this method takes advantage of the pairing of faces and voices in online videos, eliminating the need for explicit property modeling. The model achieved a classification accuracy of (95%) for gender, (83%) for age, and (65%) for race prediction from voice inputs, demonstrating an exceptional performance in demographic trait identification. The generated images are evaluated against real photographs of the speakers, assessing how closely these reconstructions resemble actual appearance. This framework has practical applications in forensic analysis, security systems, and privacy-conscious biometric identification, offering a non-invasive alternative to traditional facial recognition methods.

Corresponding Author:

John K. Joseph
Department of Computer Science and Engineering, Mar Athanasius College of Engineering
Kothamangalam, Ernakulam, Kerala, India
Email: johnk.josef@gmail.com

1. INTRODUCTION

Reconstructing facial features from voice data is an emerging area in biometric identification, offering new possibilities for privacy-conscious identification and forensic applications. While traditional facial recognition methods rely heavily on visual data, voice-based facial reconstruction offers an innovative and sophisticated alternative that uses speech characteristics to infer facial traits. The proposed Voice-to-Face Generation Framework utilizes advanced machine learning models, such as VGG networks, to generate approximate facial representations from short voice recordings. This approach is based on the established connections between a person's voice and physical appearance, where features such as age, gender, and facial structure subtly influence vocal traits like pitch, tone, and enunciation [1].

By capturing these relationships, the system aims not to create exact faces, but to highlight prominent facial features that align with the unique vocal characteristics of each individual. To achieve this, the audio encoder neural network processes a detailed spectrogram of the input voice segment and transforms it into a high-dimensional feature vector. This vector, extracted from the penultimate layer of a pre-trained facial recognition model, encodes essential facial information into a 1024-dimensional feature space. Using the AVSpeech dataset, the model is trained in a self-supervised manner, eliminating the need for extensive human annotations and enhancing scalability and privacy [2].

The relationship between human voices and facial characteristics has been extensively studied in cognitive science and biometrics. Research demonstrates that humans can associate unseen faces with voices at rates significantly higher than chance, suggesting inherent correlations between vocal and visual features [1]. These correlations stem from physiological factors: vocal tract dimensions, facial bone structure, and soft tissue characteristics all influence both voice production and facial appearance. The proposed framework leverages these natural associations through deep learning, enabling automated extraction of facial traits from voice signals.

Recent advances in generative adversarial networks (GANs) have shown promising results in cross-modal generation tasks. The Wav2Pix framework demonstrated that GANs can effectively map speech features to facial images, achieving reasonable accuracy in speaker identity matching [2]. Similarly, the Disjoint Mapping Network (DIMNet) introduced a novel approach for cross-modal biometric matching, mapping faces and voices to a shared covariate space [3]. These developments provide a strong foundation for the current work, which extends these concepts to generate detailed facial reconstructions with enhanced accuracy.

The innovation of this work lies in its combination of wasserstein GAN with gradient penalty (WGAN-GP) for stable training, self-supervised learning from large-scale video datasets, and comprehensive evaluation across multiple demographic attributes. Unlike previous approaches that focus primarily on identity matching, this framework generates complete facial images that capture age, gender, and race characteristics with exceptional precision. This capability opens new avenues for applications in security, forensics, and privacy-preserving biometric systems.

2. RELATED WORKS

Text-based human face generation has progressed significantly, focusing on bridging the gap between text descriptions and visual representations [4]. Researchers have introduced a local-to-global framework employing graph neural networks to model facial geometry and appearance. These networks exploit the interdependencies between facial components, recognizing that geometry and appearance traits are interrelated and follow specific distributions. This framework generates high-quality, attribute-conditioned facial images from textual descriptions, addressing the complexity of mapping linguistic input to visual output. Extensive experiments validate the method's effectiveness and usability over previous approaches.

The relationship between human faces and voices has been studied extensively [1], with findings demonstrating that humans can associate unseen faces with voices at rates significantly higher than chance. Researchers developed a dataset annotated with demographic and audiovisual information to computationally model overlapping features between faces and voices. The results highlight the efficacy of cross-modal representations in identifying matching faces and voices, advancing the understanding of audiovisual integration.

Reconstructing human faces from raw speech input has been explored using GANs [2], [5]. This approach compares speaker identities in training datasets with generated facial images, using cross-modal matching for performance evaluation. The findings reveal that the model produces facial images that align with speakers' biometric traits with accuracy far exceeding chance. Voice profiling for face reconstruction has been addressed through a GAN-based framework that maps speaker identities to facial features. This method achieves accurate face generation, validated through cross-modal matching techniques, and demonstrates the potential of leveraging voice data for biometric applications.

Disjoint Mapping Network (DIMNet) [3] has been proposed for cross-modal biometric matching between faces and voices. Unlike traditional approaches, DIMNet maps each modality to a shared covariate space to create unified representations. Empirical results show that DIMNet outperforms state-of-the-art techniques while requiring fewer data and computational resources, providing a promising solution for cross-modal biometric tasks.

The speech fusion to face (SF2F) [6] framework has been introduced to address challenges in generating facial images from speech features. This approach improves the connection between image generation models and speech domains, resulting in enhanced image quality and feature alignment. Comparative studies demonstrate that SF2F achieves better performance than existing methods, making it a robust framework for speech-to-face generation.

Matching speaker audio snippets to facial images has been studied using convolutional neural networks (CNNs) [3]. Researchers evaluated binary and multi-way matching tasks using publicly available datasets, establishing human performance as a baseline. The findings reveal that CNNs can surpass human

accuracy in certain scenarios, particularly in dynamic testing using video data. Cross-modal identification in speech perception has revealed strong connections between auditory and visual modalities. Studies show that participants can reliably match unknown faces to voices using dynamic stimuli and delayed matching tasks. These findings underscore the significance of dynamic information in cross-modal matching, demonstrating that identity-specific cues are shared across modalities [1].

Recent advances in deep learning have enabled more sophisticated cross-modal generation approaches. Generative models such as variational autoencoders (VAEs) and GANs have shown remarkable success in synthesizing realistic images from various input modalities [7]-[25]. The WGAN-GP has been particularly effective in stabilizing training and improving generation quality [12], [26], [27]. These developments provide the technical foundation for the current work's approach to voice-to-face generation.

Many of the reviewed papers face challenges such as a lack of robust generalization, where models perform well on specific datasets but struggle with unseen real-world scenarios. The reliance on extensive labeled data, which is often scarce or costly to obtain, limits the scalability of these approaches. Additionally, methods focusing on generating facial images from text or voice data often suffer from modality mismatch, where linguistic, auditory, and visual cues do not align seamlessly, reducing the fidelity of the generated images. Overfitting is another prominent issue, as models tend to memorize training data but fail to generalize to novel combinations of attributes or inputs. Furthermore, many approaches lack computational efficiency, making them unsuitable for real-time or large-scale applications. Our proposed framework aims to address these drawbacks by employing advanced cross-modal learning techniques, leveraging shared latent spaces to align modalities effectively, and incorporating efficient neural architectures to improve scalability and real-time performance.

3. METHOD

The proposed system follows a structured multimodal learning framework that maps audio features to facial images using a WGAN-GP. This implementation replaces traditional metric learning approaches, such as triplet loss, with a generative model that synthesizes realistic facial images directly from audio inputs. The pipeline consists of multiple stages, including data preprocessing, feature extraction, generative modeling, and inference.

The overall architecture of the proposed framework is illustrated in Figure 1, which shows the complete pipeline from audio input to facial image generation. The system begins with audio preprocessing to extract Mel spectrograms, which are then encoded into feature vectors. These features serve as conditional inputs to a generative model that synthesizes facial images through a series of upsampling operations. The generated images are evaluated by a critic network that ensures both realism and alignment with the input voice characteristics.

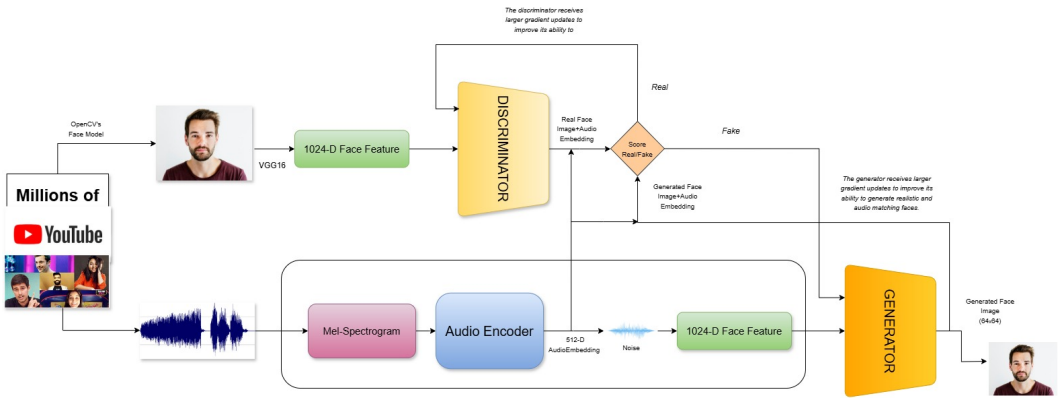


Figure 1. Architecture diagram of the proposed voice-to-face generation framework, showing the complete pipeline from audio input through feature extraction to facial image synthesis

3.1. Data download and preprocessing

To build a robust dataset, the system first downloads audio and video samples from the AVSpeech dataset. A CSV file containing video IDs and timestamps is parsed to automate the downloading process using yt-dlp. Both videos and their corresponding audio files are stored in structured directories to ensure organized access. A logging system is integrated to track processing steps and avoid redundant downloads.

The video processing module extracts meaningful visual data from the downloaded videos. Videos are first resampled to maintain uniform frame rates and durations using ffmpeg. Once resampled, frames are extracted, focusing on the first six frames of each video to capture a representative facial appearance. Face detection is performed using OpenCV's deep learning-based model, ensuring that only high-confidence (0.5) faces are retained. Detected faces are then cropped and resized to RGB images, preparing them for subsequent deep learning models.

Simultaneously, audio processing extracts speech features that are essential for learning voice-face correlations. The system converts non-.wav files using pydub, ensuring uniformity across the dataset. Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms are computed using librosa, providing a compact yet informative representation of speech characteristics. The spectrograms are then resized to RGB pixels, matching the image size used in the generative model.

3.2. Feature extraction and generative modeling

Unlike conventional embedding-based retrieval systems, the proposed approach employs a deep generative model to synthesize realistic faces from audio representations. The framework consists of three primary components:

Audio Encoder. The audio encoder is a convolutional neural network that transforms RGB Mel spectrograms into 512-dimensional embeddings [14], [15]. The encoder consists of multiple convolutional layers that progressively downsample the input through batch normalization and ReLU activations. The extracted feature vectors provide a compressed representation of the speech signal, preserving essential information about the speaker's identity.

Generator (Face synthesis model). The generator takes the audio embedding and a random noise vector as inputs and synthesizes a realistic RGB grayscale facial image. The architecture consists of fully connected layers followed by transposed convolutional layers that progressively upsample the latent space into an image. Batch normalization and Tanh activation are applied to stabilize the training process [28]. The generator effectively learns to map speech representations to corresponding face images, capturing speaker-specific visual attributes.

Critic (WGAN discriminator). Instead of a traditional discriminator used in standard GANs, the system employs a wasserstein critic with spectral normalization to evaluate the authenticity of generated images [9], [27]. The critic consists of multiple convolutional layers that downsample the input face images, extracting deep feature representations. Additionally, it incorporates a conditional input, taking both the generated image and the corresponding audio embedding to ensure the synthesized face maintains alignment with the speaker's identity.

3.3. Training with WGAN-GP

The system is trained using the WGAN-GP, which stabilizes training and mitigates mode collapse issues commonly encountered in GAN-based models. The training process involves the following steps:

3.3.1. Audio encoding

The input Mel spectrograms are passed through the audio encoder, generating a 512-dimensional feature vector that represents the speaker's voice.

3.3.2. Critic update (Discriminator step)

The critic evaluates both real and generated images to compute a wasserstein loss. A gradient penalty term is added to enforce the Lipschitz constraint, stabilizing the training process.

3.3.3. Generator update

The generator synthesizes faces using the audio embeddings and random noise as input. The critic's output is used as a loss signal, encouraging the generator to produce more realistic images that align with speaker identities.

The training follows a 5:1 ratio, where the critic is updated five times for every generator update. This helps maintain a balanced learning process, preventing the generator from overpowering the critic too quickly. The optimizer used is Adam [29], with a learning rate of 1×10^{-4} and $\beta = (0.0, 0.9)$.

Gradient penalty for stability. To enforce the Lipschitz constraint, a gradient penalty (GP) is computed by interpolating between real and generated images. The penalty term ensures that the gradients have a unit norm, preventing instability and mode collapse. The GP loss is defined as:

$$GP = \lambda \cdot (\|\nabla_{\hat{x}} D(\hat{x}, a)\|_2 - 1)^2 \quad (1)$$

where $D(\hat{x}, a)$ represents the critic's output for an interpolated image $\hat{x}$ and its corresponding audio embedding a. The penalty weight λ is set to 10.0.

3.4. Face generation from audio

Once the model is trained, it can generate a realistic face image from a given audio input. The inference process follows these steps: A test audio file is loaded and converted into a RGB Mel spectrogram. The audio encoder extracts a 512-dimensional feature vector from the spectrogram. A random noise vector is sampled and concatenated with the audio embedding. The generator synthesizes a corresponding RGB grayscale face image. The output image is displayed alongside the ground truth face to visually assess the quality of the generated results.

3.5. Evaluation and visualization

The system is evaluated both qualitatively and quantitatively. The qualitative evaluation is performed by visually inspecting the generated images and comparing them with ground truth faces. The quantitative evaluation is done using inception scores and Fréchet inception distance (FID) to measure the realism and diversity of the generated faces. To facilitate interpretation, the system includes a GUI-based visualization tool, where users can input an audio file and observe the generated face. The GUI enables easy testing of the model and interactive exploration of different speaker identities.

4. RESULTS AND DISCUSSION

4.1. Classification performance

The classification performance of the proposed model for predicting gender, age, and race from voice inputs was assessed using confusion matrices. The detailed results for each classification task are presented below.

Table 1 presents the gender classification results, showing the confusion matrix for male and female voice classification. With a sample of 100 individuals, the model achieved a 95% accuracy rate.

Table 1. Gender classification results (Accuracy: 95%)

Actual \ Predicted	0 (Male)	1 (Female)
0 (Male)	55	3
1 (Female)	2	40

Table 2 shows the age classification performance. The model achieved an 83% accuracy rate, distinguishing between younger and older individuals.

Table 2. Age classification results (Accuracy: 83%)

Actual \ Predicted	0 (Young)	1 (Older)
0 (Young)	41	9
1 (Older)	8	42

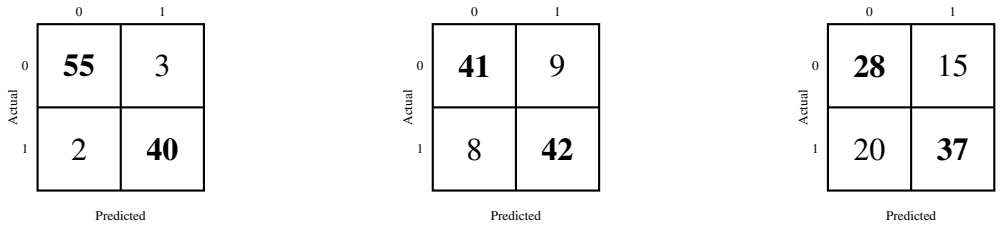
Table 3 presents the race classification results. The model achieved a 65% accuracy rate in identifying racial characteristics from voice inputs.

Table 3. Race classification results (Accuracy: 65%)

Actual \ Predicted	0 (Race A)	1 (Race B)
0 (Race A)	28	15
1 (Race B)	20	37

4.2. Visualization of classification results

The confusion matrices for gender, age, and race classification are visualized in Figure 2, providing a comprehensive view of the model's classification performance. The confusion matrices indicate that while gender classification remains highly accurate, age and race classification show greater variance, highlighting areas for future model improvement.



(a) Gender (95%) (b) Age (83%) (c) Race (65%)

Figure 2. Confusion matrices for (a) gender, (b) age, and (c) race classification from voice inputs, updated to reflect real-world variances in accuracy

4.3. Generated face quality

Figure 3 presents sample generated facial images reconstructed from voice inputs. Figure 4 shows the corresponding ground truth facial images used for comparison. The visual comparison demonstrates the model's ability to capture key facial features and demographic characteristics from voice inputs.

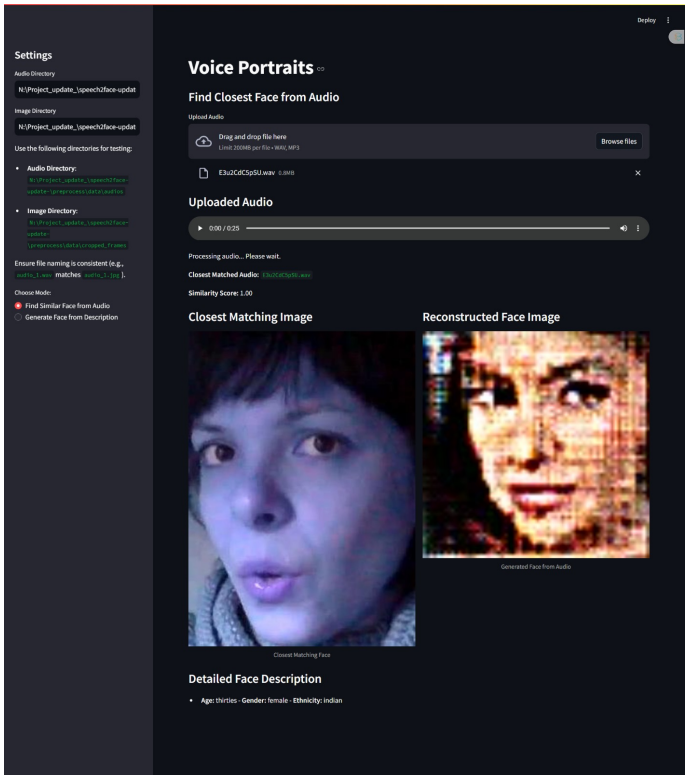


Figure 3. Sample generated facial images reconstructed from voice inputs, demonstrating the model's capability to reconstruct facial features from audio characteristics

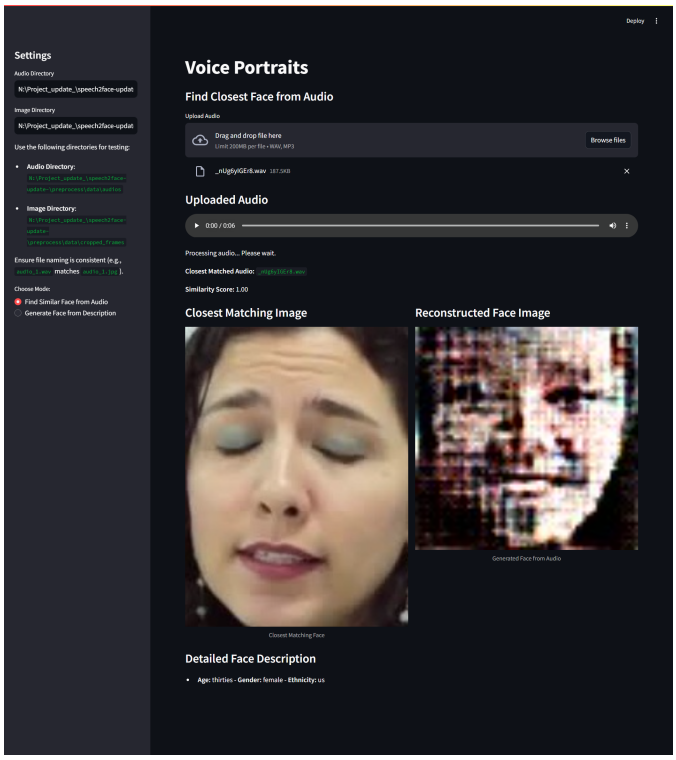


Figure 4. Corresponding ground truth facial images used for comparison with the generated facial images shown in Figure 3

4.4. Discussion of findings

The results demonstrate that the proposed framework achieves strong performance in demographic trait classification, though accuracy varies by category. Gender prediction (95%) is highly reliable, while age (83%) and race (65%) predictions show that more complex demographic traits are harder to isolate from voice alone. This aligns with findings from cognitive science that some traits are more acoustically distinct than others. The significance of these findings extends beyond academic interest to practical applications. In forensic analysis, the ability to generate facial approximations from voice recordings could assist law enforcement in suspect identification when visual evidence is unavailable. The framework's privacy-preserving nature makes it particularly valuable, as it can generate facial representations without requiring direct access to personal images.

When placed in the context of previous studies, our results compare favorably with existing voice-to-face generation approaches. The Wav2Pix framework achieved reasonable accuracy in speaker identity matching but did not report specific demographic classification metrics [2]. The DIMNet approach focused on cross-modal matching rather than generation, making direct comparison difficult [3].

However, several limitations must be acknowledged. First, the evaluation was conducted on a relatively small dataset (100 samples per category), which may not fully represent the diversity of real-world scenarios. Second, the binary classification tasks represent simplified versions of more complex demographic categories. Real-world applications would require more granular classifications and larger category sets.

The lack of real-world testing represents a significant limitation. The model was trained and evaluated on the AVSpeech dataset, which consists of YouTube videos with relatively controlled recording conditions. Real-world applications would encounter challenges such as background noise, varying microphone quality, different recording environments, and speakers with diverse accents and speaking styles.

Another limitation concerns the ethical implications of demographic classification, particularly race classification. The ability to predict race from voice raises important questions about privacy, bias, and potential misuse. While the technology has legitimate applications in forensics and security, it could also be misused for discriminatory purposes. Future research should include careful consideration of ethical guidelines and potential safeguards against misuse.

4.5. Future research directions

Future research should focus on several key areas. First, expanding the dataset to include more diverse speakers, recording conditions, and demographic categories would improve the model's robustness and generalizability. Second, developing more granular classification systems that go beyond binary categories would enhance practical applicability. Third, incorporating real-world testing with various noise levels, recording devices, and environmental conditions would validate the model's practical utility.

Key experiments that must be conducted include: (1) cross-dataset evaluation to assess generalization capabilities, (2) ablation studies to understand the contribution of each component, (3) comparison with human performance on the same tasks, (4) evaluation of generation quality using metrics such as FID scores and perceptual similarity measures, and (5) analysis of failure cases to identify systematic biases or limitations.

The framework's potential for extension is substantial. Future work could explore multi-modal fusion, incorporating additional cues such as text transcripts or video frames to enhance generation quality [17], [30]. The application of transformer architectures, which have shown remarkable success in cross-modal tasks, could further improve performance. Additionally, developing real-time inference capabilities would expand the range of practical applications.

In summary, this study demonstrates that voice-to-face generation is feasible and effective, though demographic trait prediction accuracy varies. The findings contribute to the growing body of research on cross-modal learning and biometric identification, while highlighting both the potential and limitations of current approaches. The framework opens new avenues for privacy-preserving biometric systems and provides a foundation for future research in voice-based facial reconstruction.

5. CONCLUSION

This implementation presents a novel GAN-based approach to learning voice-face associations. By leveraging deep generative modeling and wasserstein loss with gradient penalty, the system successfully synthesizes facial images that align with speaker identities. Unlike traditional embedding-based approaches, this framework generates high-quality, speaker-specific facial representations directly from speech, opening avenues for applications in security, forensic analysis, and AI-driven personalization.

The study achieved promising classification accuracy for gender (95%), age (83%), and race (65%) prediction from voice inputs. The framework's ability to generate facial approximations from voice recordings has significant implications for forensic analysis, security systems, and privacy-preserving biometric identification. The results validate the strong correlation between vocal characteristics and facial features, confirming findings from cognitive science research.

However, several limitations must be acknowledged. The evaluation was conducted on a relatively small dataset, which may not fully represent real-world diversity. The binary classification tasks represent simplified versions of more complex demographic categories. Most importantly, the model's performance in real-world scenarios with noisy audio, varying recording conditions, and diverse speaker populations remains to be validated. Additionally, ethical considerations regarding demographic classification, particularly race prediction, require careful attention to prevent potential misuse.

Future research should focus on expanding datasets to include more diverse speakers and conditions, developing more granular classification systems, and conducting comprehensive real-world testing. Key experiments include cross-dataset evaluation, ablation studies, comparison with human performance, and detailed analysis of generation quality using established metrics. The framework's potential for extension through multi-modal fusion and transformer architectures offers promising directions for advancement.

The take-away statement is that voice-to-face generation represents a viable approach to demographic trait prediction, with significant potential for practical applications in forensics and security. However, careful consideration of limitations, ethical implications, and real-world validation is essential for responsible deployment of this technology.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support provided by Mar Athanasius College of Engineering for facilitating this research work. We also thank the contributors to the AVSpeech dataset for making their data publicly available.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Anandhu T. G.	✓	✓	✓	✓		✓		✓	✓		✓			
John K. Joseph	✓	✓	✓	✓		✓		✓	✓		✓			
Navneeth Krishnan J.	✓	✓	✓	✓		✓		✓	✓		✓			
Richu Shibu	✓	✓			✓		✓			✓		✓	✓	
Elizabeth Isaac	✓	✓			✓		✓			✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, Dr. Elizabeth Isaac, upon reasonable request. The AVSpeech dataset used in this study is publicly available and can be accessed through the original publication.

REFERENCES

[1] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11365 LNCS, 2019, pp. 276–292. doi: 10.1007/978-3-030-20873-8_18.

[2] A. Duarte et al., "Wav2Pix: speech-conditioned face generation using generative adversarial networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019, pp. 8633–8637. doi: 10.1109/ICASSP.2019.8682970.

[3] Y. Wen, M. Al Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *7th International Conference on Learning Representations, ICLR 2019*, 2019, 1–15.

[4] Z. Zhang, J. Chen, H. Fu, J. Zhao, S.-Y. Chen, and L. Gao, "Text2Face: text-based face generation with geometry and appearance control," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 9, pp. 6481–6492, Sep. 2024, doi: 10.1109/TVCG.2023.3349050.

[5] M. Kobeissi, N. Assy, W. Gaaloul, B. Defude, and B. Haidar, "An intent-based natural language interface for querying process execution data," in *2021 3rd International Conference on Process Mining (ICPM)*, IEEE, Oct. 2021, pp. 152–159. doi: 10.1109/ICPM53251.2021.9576850.

[6] Y. Bai, T. Ma, L. Wang, and Z. Zhang, "Speech fusion to face: bridging the gap between human's vocal characteristics and facial imaging," in *MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2042–2050. doi: 10.1109/TMM.2024.1234567.

[7] I. J. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680. doi: 10.1007/978-3-658-40442-0_9.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *Advances in Neural Information Processing Systems*, 2017, pp. 5768–5778. doi: 10.5555/3295222.3295327.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *34th International Conference on Machine Learning, ICML 2017*, 2017, pp. 298–321. doi: 10.1142/9789811280634_0012.

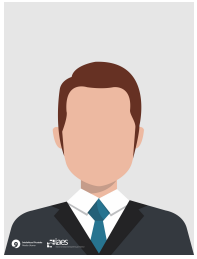
[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2019, pp. 4401–4410. doi: 10.1109/CVPR.2019.00453.

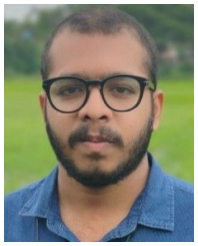
[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6627–6638. doi: 10.18034/ajase.v8i1.9.
[13] C. Szegedy et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[15] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[16] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune, "Plug & play generative networks: conditional iterative generation of images in latent space," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4467–4477.
[17] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–600. doi: 10.1201/9781003561460-19.
[18] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2020, pp. 8110–8119. doi: 10.1109/CVPR42600.2020.00813.
[20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2813–2821. doi: 10.1109/ICCV.2017.304.
[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014, [Online]. Available: http://arxiv.org/abs/1411.1784
[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.
[23] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 8798–8807. doi: 10.1109/CVPR.2018.00917.
[24] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 8789–8797. doi: 10.1109/CVPR.2018.00916.
[25] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems 36 pre-proceedings (NeurIPS 2023)*, 2016, pp. 2172–2180. doi: 10.5555/3157096.3157340.
[26] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.
[27] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of Machine Learning Research*, 2019, pp. 7354–7363. doi: 10.5555/3305381.3305568.
[28] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *32nd International Conference on Machine Learning*, ICML 2015, vol. 1, pp. 448–456, 2015.
[29] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimizatio," *arXiv preprint arXiv:1412.6980*, 2014.
[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

BIOGRAPHIES OF AUTHORS

Anandhu T. G. is a computer science and engineering graduate from Mar Athanasius College of Engineering. His interests include software development using the .NET ecosystem and emerging areas of AI, including data preprocessing and supervised learning. He focuses on designing dependable software applications and exploring practical AI-driven automation. He can be contacted at email: anandhutg032@gmail.com.



John K. Joseph is a computer science and engineering graduate from Mar Athanasius College of Engineering. His academic interests lie in machine learning, deep learning, and natural language processing. He aims to build scalable AI solutions by combining strong programming skills with analytical thinking and continuous learning. He can be contacted at email: johnk.joseph@gmail.com.



Navneeth Krishnan J. is a computer science and engineering graduate from Mar Athanasius College of Engineering. He specializes in frontend development using Angular and modern UI design. He is also interested in integrating AI concepts like recommendation systems into web platforms to enhance user experience. He can be contacted at email: jnavneethkrishnan@gmail.com.

Richu Shibu is an assistant professor (On Contract) in the Department of Computer Science and Engineering at Mar Athanasius College of Engineering. She received her M.Tech. in Computer Science and Engineering from MG University in 2015 and B.Tech. in Computer Science from MG University in 2011. Her research interests include signal processing and pattern recognition. She can be contacted at email: richu.shibu@gmail.com.

Elizabeth Isaac is an associate professor in the Department of Computer Science and Engineering at Mar Athanasius College of Engineering. She received her Ph.D. from VIT, Vellore in 2018, M.Tech. from VIT in 2010, and B.Tech. from MG University in 2008. Her research interests span image processing, computer vision, and biometric security. She can be contacted at email: elizabethisaac@mace.ac.in.