Emotional-Controllable Talking Face Generation on Real-Time System

Van-Thien Phan[1], Hyung-Jeong Yang[1]*, Seung-Won Kim[1], Ji-Eun Shin[2], Soo-Hyung Kim[1]
[1]Dept. of Artificial Intelligence, Chonnam National University
[2]Dept. of Psychology, Chonnam National University
*Corresponding Author
phanthienbka@gmail.com, hjyang@jnu.ac.kr, seungwon.kim@jnu.ac.kr, jieunshin@jnu.ac.kr, shkim@jnu.ac.kr

Abstract

Recent progress in audio-driven talking face generation has focused on achieving more realistic and emotionally expressive lip movements, enhancing the quality of virtual avatars and animated characters for applications in entertainment, education, healthcare, and more. Despite these advances, challenges remain in creating natural and emotionally nuanced lip synchronization efficiently and accurately. To address these issues, we introduce a novel method for audio-driven lip-sync that offers precise control over emotional expressions, outperforming current techniques. Our method utilizes Conditional Deep Variational Autoencoder to produce lifelike lip movements that align seamlessly with audio inputs while dynamically adjusting for various emotional states. Experimental results highlight the advantages of our approach, showing significant improvements in emotional accuracy and the overall quality of the generated facial animations, video sequences on the Crema-D dataset [1].

Keywords: Audio-driven, emotional generation, lip-sync, variational autoencoder

1. Introduction

Audio-driven talking face generation has gained significant attention due to its potential applications in various fields such as entertainment, virtual assistants, language learning, and healthcare. The primary challenge in this domain is to create realistic and natural facial animations that accurately synchronize lip movements with the provided audio while also capturing a wide range of human emotions. Traditional methods often struggle to achieve both high-quality lip synchronization and dynamic emotional expressions, limiting their effectiveness in real-world applications.

Recent methods, such as those proposed by Vougioukas et al. [2], Eskimez et al. [3], and Magnusson et al. [4], have made notable advancements in improving lip-sync quality and facial animation realism. However, these methods still face limitations, including suboptimal emotional accuracy and higher Frechet Inception Distance (FID) scores, which indicate a gap in achieving the desired level of realism and emotional expressiveness.

To address these limitations, we propose a novel approach for audio-driven lip-sync generation with controllable emotions that aims to outperform existing methods in both accuracy and quality. Our method utilizes Conditional Deep Variational Autoencoder which mainly based on Unet architecture [5] to ensure a comprehensive representation of both visual and auditory cues, enhancing the overall realism of the generated outputs. Furthermore, applying the encoder method to extract features of audio and visual and to employ loss functions for lip-synchronization, emotional discriminator which referred at Eskimez et al. [3] and Prajwal et al. [6].

The experimental results demonstrate that our approach significantly outperforms previous state-of-the-art methods. Our method shows a notable reduction in FID to 6, reflecting image quality, and achieves improved lip-sync metrics (LSE-C and LSE-D), confirming its effectiveness in maintaining natural and dynamic expressions.

In summary, our proposed method offers a robust solution for generating emotionally expressive and high-quality talking face animations, setting a new benchmark in the field. This work contributes to the development of more interactive and realistic virtual avatars and characters, opening new possibilities for applications across various domains.

2. Proposed Method

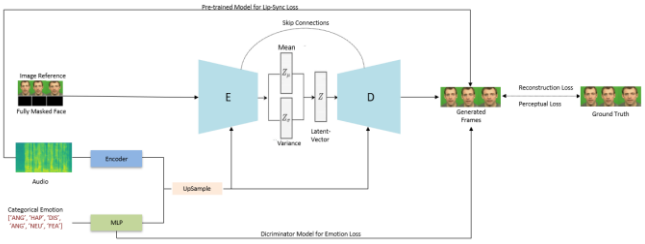

Figure 1. A proposed model for emotional-controllable talking face generation

Figure 1 illustrates a framework for generating emotion-controlled lip-sync videos using audio input, categorical emotions, and visual references. We applied various data augmentation techniques- including random brightness, contrast, gamma adjustment, channel shuffle, RGB shift, and Gaussian noise- to all input frames. This ensured visual consistency across frames and improved model generalization by increasing training data diversity and accommodating different background settings.

A fully masked face serves as a visual reference to guide the appearance of the generated frames, while the spectrogram derived from the audio input drives the lip-sync process. An emotion label is encoded via a Multi-Layer Perceptron (MLP) with Leaky ReLU activation and then concatenated with audio features after upsampling to align with the temporal resolution of visual features in both the encoder and decoder.

The visual reference is encoded and concatenated with the upsampled output from the combined emotion and audio features to extract latent features. This process results in a latent vector consisting of a mean and variance, forming a probabilistic distribution in the latent space. The latent vector, along with the upsampled resolution from the audio and emotion features, is passed through a decoder that reconstructs the video frames, preserving both the lip-sync and emotional attributes. Skip connections are employed to improve information flow from the encoder to the decoder, ensuring finer details are retained in the generated frames.

2.1 Loss Functions and Optimization

Reconstruction Loss: To minimize the difference between the generated frames $\hat{y}_i$ and the ground truth frames y_i. Loss MSE reconstruction used to ensure accurate reproduction.

$$Lrecon = \sum_{i=1}^{N} \left(y_i - \hat{y} \right)^2 \quad (1)$$

KL Divergence Loss: computed as the sum of the Kullback-Leibler divergence between the predicted probability distributions and the target distributions. This loss measures the difference between these distributions, encouraging the model to produce outputs that closely match the target distribution. And probability distributions between the predicted distribution P and the target distribution Q is given by:

$$KL(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (2)$$

Perceptual Loss: Intermediate features from both ground truth and generated videos are extracted using a pre-trained VGG-19 network [7]. The perceptual loss (PL) [8] is then calculated as the mean-squared error between these features.

Emotion Loss: The emotion discriminator is first pre-trained which similar to that in [3] for several epochs, we adjust the image resolution 64x64 pixels to suitable with our model, then fine-tuned with the generator. For final training, we use cross-entropy loss: the discriminator learns from ground truth labels, while the generator learns from the discriminator's predictions on generated frames.

$$L = -\sum_{i=1}^{C} y_i \cdot \log(pi) \quad (3)$$

where C is the number of classes, yi is a binary indicator, and pi is the predicted probability of class i.

Lip-Sync Loss: assesses the synchronization accuracy of the generated frames with the audio input. Cosine similarity with binary cross-entropy loss, is used for optimizing the model weights:

$$Psync = \frac{v \cdot s}{max(\|v\|_2 \cdot \|s\|_2, \varepsilon)} \quad (4)$$

To obtain P_{sync}, we compute a dot product between the ReLU-activated video and speech embeddings v s to get the probability of synchronization of an audio-video pair.

$$Esync = \frac{1}{N}\sum_{i=1}^{N} -\log\left(P^{i}_{sync}\right) \quad (5)$$

Combined Objective Function. To train the generator, a full objective function given by:

$$Lgen = L_{recon} + L_{KL} + \alpha L_{sync} + \beta L_{perc} + \gamma L_{emo} \quad (6)$$

where, α, β, ɣ are the weights for the respective loss components.

3. Result

Crema-D is an emotional multimodal actor data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female. Six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

The results presented in this section compare the performance of our proposed model against several existing methods, including Vougioukas et al. [2], Eskimez et al. [3], and Magnusson et al. [4], based on various evaluation metrics.

Table 2 shows the performance of the different methods on several metrics include Averaged Emo-Accuracy, FID (Frechet Inception Distance) [9], LSE-C and LSE-D (Lip-Sync Error - Content and Dynamics) for six different emotions generated [6]. The metric evaluates the accuracy of the emotional expression generated by the models. Our method achieves an accuracy of 87.28%, outperforming Vougioukas et al. with 55.26% and Eskimez et al. with 65.67%. Lower FID scores indicate better-quality images. Our method achieves a significantly lower FID score of 5.228 compared to Eskimez et al.(79.11), suggesting superior image quality. To measure the accuracy of lip synchronization with the audio content, our method shows a substantial improvement, with LSE-C of 9.540 and LSE-D of 5.165, outperforming Eskimez et al. which reported LSE-C of 10.244 and LSE-D of 3.256.

The qualitative results are visualized in Figure 2, comparing the output of different methods. The ground truth shows the original reference images. The second row displays the results of the Eskimez et al. method, where the generated images have moderate similarity to the ground truth in terms of facial expression and lip synchronization. The third row shows the outputs from Magnusson et al., which exhibit less accurate facial expressions and emotional representation. The fourth row demonstrates the outputs generated by our proposed method, which closely resemble the ground truth, particularly in terms of emotional expression, lip synchronization, and overall image quality.

Figure 2. Comparison of "happy" emotion frames generated by various methods.

Overall, our model demonstrates the performance in both quantitative and qualitative evaluations, providing high-quality image generation and accurate emotion depiction compared to previously methods.

4. Application System

Figure 3 illustrates a face application system designed to generate customized outputs by integrating three input modules: visual, audio, and categorical emotion. These inputs are processed by a feature extraction module, which uses algorithms to create a combined multimodal feature representation. The generator module then uses these features, along with data from a database to synthesize the final output. This output is presented to the user through the display module.

Table 2. Compares 4 indies include FID, averaged emotional accuracy, lib-sync than previous methods

	Averaged Emo-Accuracy ↑	FID ↓	LSE-D ↓	LSE-D ↑
Vougioukas et al. [2]	55.26	71.12	-	-
Eskimez et al. [3]	65.67	79.11	10.244	3.256
Our	**87.28**	**5.228**	**9.540**	**5.165**

Figure 4 integrated approach allows the system to create dynamic and personalized content based on diverse input types, making it suitable for applications in interactive multimedia,

virtual assistants, and emotion-aware systems.

The interface demonstrates the application of a multimodal system where users can upload audio, visual, and select an emotion to generate a customized output in Figure 4. The result displayed suggests that the system is capable of blending facial characteristics or modifying visual content based on the provided inputs. The generated output shows a modified image, implying the system uses both facial manipulation and possibly deep learning techniques to combine and generate content that reflects the selected emotion.

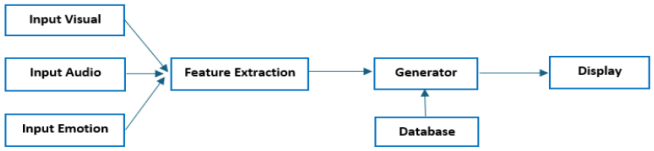

Figure 3. Block diagram of face generation system

Figure 4. A web-based application interface for demonstrating a user input and output process.

5. Conclusion

In this paper, our model outperforms existing methods as demonstrated by qualitative evaluations. This system shows promise for applications in interactive multimedia, virtual assistants, and emotion aware technologies. Future work will focus on experimenting other datasets and improving the model.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (RS-2023-00219107). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT). This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-RS-2022-00156287) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

Reference

[1] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4):377–390, 2014.

[2] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. CoRR, abs/1906.06337, 2019. URL http://arxiv.org/abs/1906.06337.

[3] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. IEEE Transactions on Multimedia, pages 1–1, 2021. doi: 10.1109/TMM.2021.3099900.

[4] Ian Magnusson, Aruna Sankaranarayanan, and Andrew Lippman. Invertable frowns: Video-to-video facial emotion translation. In Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection, pages 25–33, 2021.

[5] Child, R. (2020). Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. *ArXiv, abs/2011.10650*.

[6] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, pages 484–492, 2020.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision, pages 694–711. Springer, 2016.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.