랜드마크 시퀀스를 기반으로 한 개별 오디오 구동 화자 생성

Son Thanh-Hoang Vo, Quang-Vinh Nguyen, 양형정, 신지은, 김승원, 김수형*
전남대학교 인공지능융합대학
{hoangsonvothanh, vinhbn28, hjyang, jieunshin, seungwon, shkim}@jnu.ac.kr

Individual Audio-Driven Talking Head Generation based on Sequence of Landmark

Son Thanh-Hoang Vo, Quang-Vinh Nguyen, Hyung-Jeong Yang,
Jieun Shin, Seungwon, Soo-Huyng Kim*
College of Artificial Intelligence Convergence, Chonnam National University

Abtract

Talking Head Generation is a highly practical task that is closely tied to current technology and has a wide range of applications in everyday life. This technology will be of great help in the fields of photography, online conversation as well as in education and medicine. In this paper, the authors proposed a novel approach for Individual Audio-Driven Talking Head Generation by leveraging a sequence of landmarks and employing a diffusion model for image reconstruction. Building upon previous landmark-based methods and advancements in generative models, the authors introduce an optimized noise addition technique designed to enhance the model's ability to learn temporal information from input data. The proposed method outperforms recent approaches in metrics such as Landmark Distance (LD) and Structural Similarity Index Measure (SSIM), demonstrating the effectiveness of the diffusion model in this domain. However, there are still challenges in optimization. The paper conducts ablation studies to identify these issues and outlines directions for future development.

1. Introduction

Talking head generation is the task of generating talking video segments, which has highly practical applications in industries such as filmmaking, online communication, the metaverse, and especially in education and healthcare [1]. Audio-driven talking head generation is an approach to this problem that uses the speaker's audio as the primary input along with an identity image of the person to be generated. The goal of this task is to generate a video of the person talking that is synchronized with the given audio.

This is a topic that has attracted significant interest from researchers, and there have been certain achievements. Some studies approach the problem by removing the moving areas such as the mouth and eyes from the identity image and attempting to reconstruct them from the audio [2-4]. This approach is intriguing, but the results are unnatural because the head posture remains static. This method proves to be effective in handling the task of controlling emotion in videos. The landmark-based method has yielded remarkable results, with the output being able to mimic real-life movements. However, these methods often rely on models such as GANs and VAEs for generation, which usually result in image quality that is not truly optimal. Additionally, another approach that is gaining more attention from researchers is constructing a sequence of landmarks from audio and utilizing this sequence to generate the face [5-8]. Furthermore, with the development of current generative models, some studies have applied Diffusion models to this task, as demonstrated by [9-10] with very positive results.

Building on the strengths of previous methods, in this paper, we aim to develop a model to solve the Audio-Driven Talking Head generation task through a Sequence of Landmarks and image reconstruction using Diffusion. Additionally, we propose a highly effective noise addition method tailored for this task, and we will use a Diffusion model with video as input to easily learn temporal information from the input data.

In the following sections, the authors will present the method of generating a Sequence of Landmarks from audio, a method we have researched previously. The Proposed Method section will detail our proposed approach. Then, experiments

* 교신저자 (Corresponding Author)

will be conducted to demonstrate the effectiveness of our approach. Finally, in the Conclusion section, we will summarize the paper, highlighting its advantages and limitations.

2. Related Work

To implement the proposed method, the input we require is a sequence of landmarks generated from audio. Son et al. [11] proposed a method that utilizes Dual-Domain information, merged through the KAN model [13]. This model uses two separate domains, each learning different aspects of the audio data. The Global branch is responsible for extracting information from raw audio, as raw audio signals can easily represent volume variations, thus conveying information about mouth movements. The Content branch learns information based on audio vectors extracted by the Wav2vec [12] model, carrying emotional information, which provides a good representation of emotions. These two sources of information are combined through the KAN model, resulting in outstanding performance.

3. Proposed Method

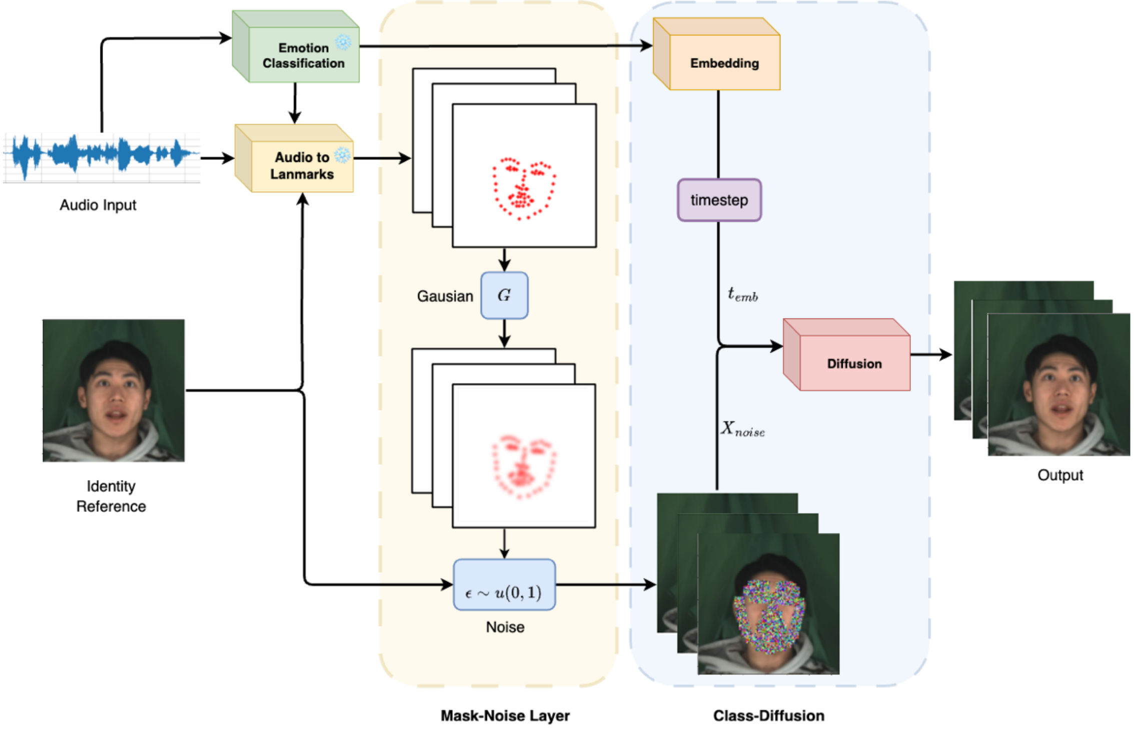

Figure 1. Overview of the method architecture

Overview: According to Figure 1, the proposed model inputs a speech audio clip and a reference identity frame of the actor. The model's predicted output is a talking video of the actor, where the mouth movements and head posture are synchronized with the input audio. The authors use two pre-trained models: Audio Emotion Classification (ConformerXL) and Audio to Landmarks (Sec. 2). Audio to Landmarks is used to generate a sequence of landmarks from the audio. Audio Emotion Classification is used to classify the audio into one of the seven emotions in the dataset.

Mask-Noise Layer: The authors have proposed a method for adding noise to diffusion models, which is an optimized approach for the problem of Talking Head Generation based on landmark sequences. This method first uses a Gaussian function to create a mask layer based on the landmarks. Each landmark position is transformed into a region with a kernel size of k and value distribution according to the Gaussian distribution. The resulting outcome can be understood using the following formula:

$$G(x,y) = \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

where x, y are the coordinates of the landmark point, σ is the standard deviation, and e is the exponential function. The result obtained from the previous step has a shape in the form of $X \in \mathbb{R}^{T\times W\times H\times C}$. Next, a mask layer is created with a threshold t using the formula.

$$M_{t,w,h,c} = \begin{cases} 1, if X_{t,w,h,c} > t \\ 0, if X_{t,w,h,c} < t \end{cases} \tag{2}$$

Afterward, a noise value is generated based on the mask M, with noise added at each position where $m = 1$, using the function $\epsilon \sim \mathcal{U}(0,1)$. Finally, these noisy positions are used to replace the corresponding positions in the original image X.

Class-Diffusion: In the subsequent layer, the author used the class prediction values from the Emotion Classification model and the sequence of noise images obtained as input to the model. To enable the model to accurately reproduce the corresponding emotion, an embedding layer is constructed to transform the class prediction values into a vector compatible with the noise image input. This emotion vector w is then combined with the time step s taken from the schedule and integrated with the noise image before being fed into the Diffusion model for training. The formula at this stage can be represented as follows formula (3).

For the Diffusion model, we use a 3D Basic UNet with 4 down-sampling layers corresponding to feature sizes of 32, 64, 128, and 256. The number of class embeddings is 8, which corresponds to the 8 emotion class labels in the dataset.

4. Experiments and Results

Dataset: In this paper, we use the MEAD dataset — A Large-scale Audio-visual Dataset for Emotional Talking-face Generation [14]. This dataset is designed for the task of Audio-Driven Talking Head Generation. The dataset consists of 60 individuals, each with 8 different types of emotions. The total video duration of this dataset is up to 40 hours. However, in this paper, we will first use a single identity to demonstrate the model's performance.

Experiments setup: To evaluate the experimental results, we trained the model and evaluated it on ID M030, splitting the data with a 90-10% ratio corresponding to the training and test sets. We use the metrics LD - Landmark Distance on whole facial landmarks and mouth specific landmarks [11] to assess the accuracy of facial motion and head pose. Additionally, we use the PSNR - Peak Signal-to-Noise Ratio and SSIM - Structural Similarity Index Measure to evaluate the quality of the output video.

For the environment, the algorithms and models are implemented in Python using the PyTorch library. Our hardware configuration includes an NVIDIA RTX 4090 GPU with 24GB of memory. For the experiments, the model is trained with the MSE loss function and optimized using the Adam optimizer with a learning rate of $1e-4$. The learning rate schedule used is Cosine Annealing, with a parameter $\mathrm{t}_{\mathrm{max}}$ equal to the number of epochs, which is 500. The diffusion scheduler used is DDPM with the number of timesteps set to 1000; however, during inference, we only use 300 timesteps.

Table 1. Quantitative comparison with related methods

Method	**PSNR**	**SSIM**	**M-LD**	**F-LD**
MEAD [14]	23.88	0.53	3.23	3.35
MakeItTalk [15]	26.78	0.56	3.23	3.65
Wav2Lip [16]	**27.69**	0.59	2.19	2.33
EVP [5]	27.63	0.71	1.38	2.88
Proposed	24.69	**0.85**	**1.22**	**1.33**

Results: The experimental results demonstrate the effectiveness of the proposed method compared to recent approaches. As shown in Table 1, our method achieves very high performance, particularly in terms of the LD metric for both the face and the entire head. This indicates that the noise addition technique we propose effectively preserves information from the sequence of landmarks. Correspondingly, the SSIM metric also shows the highest performance, proving that using the Video Diffusion model is an effective approach for the Talking Head Generation task.

However, for the PSNR metric, our results are not yet optimal. While our model delivers fairly good results, the color contrast is not as sharp as in the Ground Truth images. We believe this could be due to suboptimal post-processing or the effect of the number of timesteps used when adding noise. To investigate this, we conducted ablation studies to identify the cause.

$$\begin{aligned} w &= Embedding(w) \\ t_{emb} &= Embedding(s) \\ t_{emb} &= concat(t_{emb}, w) \end{aligned} \tag{3}$$

Ablation studies were designed to demonstrate the impact of the Diffuser Scheduler on the quality of the output results. As shown in **Table 2,** when the number of timesteps (t) increases, the evaluation metrics tend to worsen, with the best results achieved at an intermediate range of 200-300 timesteps. This finding indicates that while a larger number of timesteps allows the model to remove more noise, setting t too high leads to **'oversmoothing'**, whereas setting it too low results in remaining noise. Through this ablation study, we can propose some potential directions for further development of this task, such as designing a new Diffuser Scheduler that is better suited for this model.

Table 2. Ablation study on the quality of results depending on the number of time-steps in the Diffuser Scheduler

t	100	200	300	400	500	1000
PSNR	19.27	21.13	24.69	20.26	17.12	12.62
SSIM	0.728	0.8023	0.854	0.817	0.748	0.061

5. Conclusion

In this study, the authors have proposed a new method for Audio-Driven Talking Head Generation using a diffusion model that incorporates a sequence of landmarks as input. Our proposed noise addition technique effectively retains essential information from the landmark sequences, resulting in highly accurate facial motion and head pose synchronization, as demonstrated by our superior performance in LD and SSIM metrics. However, the model's performance in PSNR indicates that further refinement is necessary, particularly in the post-processing stage and the selection of timesteps for noise addition. The ablation studies revealed the trade-offs between timestep selection and output quality, suggesting that an optimal range of 200-300 timesteps provides the best balance. In the future, we plan to further develop this model and use a more diverse dataset to address the multi-individual task, enhancing the model's flexibility and broader applicability. Future work will also focus on designing a more suitable diffusion scheduler and additional methods to improve video quality, ensuring both robustness and realism in the generated talking head.

Acknowledgment

This work was supported by the 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), the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437718) supervised by IITP, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219107).

References

[1] P. Pataranutaporn et al., “AI-generated characters for supporting personalized learning and well-being,” Nat Mach Intell, vol. 3, no. 12, pp. 1013–1022, Dec. 2021, doi: 10.1038/s42256-021-00417-9.

[2] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li, “Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert,” in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 14653–14662. doi: 10.1109/CVPR52729.2023.01408.

[3] B. Zhang, X. Zhang, N. Cheng, J. Yu, J. Xiao, and J. Wang, “EmoTalker: Emotionally Editable Talking Face Generation via Diffusion Model,” Jan. 15, 2024, arXiv: arXiv:2401.08049. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2401.08049

[4] K. Cheng et al., “VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild,” Nov. 27, 2022, arXiv: arXiv:2211.14758. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2211.14758

[5] X. Ji et al., “Audio-Driven Emotional Video Portraits,” May 19, 2021, arXiv: arXiv:2104.07452. Accessed: Jun. 07, 2024. [Online]. Available: http://arxiv.org/abs/2104.07452

[6] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, “Hierarchical Cross-Modal Talking Face Generationwith Dynamic Pixel-Wise Loss,” May 09, 2019, arXiv: arXiv:1905.03820. Accessed: Jun. 21, 2024. [Online]. Available: http://arxiv.org/abs/1905.03820

[7] S. Tan, B. Ji, and Y. Pan, “EMMN: Emotional Motion Memory Network for Audio-driven Emotional Talking Face Generation,” in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023, pp. 22089–22099. doi: 10.1109/ICCV51070.2023.02024.

[8] J. Wang, Y. Zhao, L. Liu, T. Xu, Q. Li, and S. Li, “Emotional Talking Head Generation based on Memory-Sharing and Attention-Augmented Networks,” Jun. 06, 2023, arXiv: arXiv:2306.03594. Accessed: Apr. 04, 2024. [Online]. Available: http://arxiv.org/abs/2306.03594

[9] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic, “Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation,” in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA: IEEE, Jan. 2024, pp. 5089–5098. doi: 10.1109/WACV57701.2024.00502.

[10] S. Shen et al., “DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1982–1991. Accessed: Jun. 17, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Shen_DiffTalk_Crafting_Diffusion_Models_for_Generalized_Audio-Driven_Portraits_Animation_CVPR_2023_paper.html

[11] H.-S. Vo-Thanh, Q.-V. Nguyen, and S.-H. Kim, “KAN-Based Fusion of Dual-Domain for Audio-Driven Facial Landmarks Generation,” Sep. 09, 2024, arXiv: arXiv:2409.05330. Accessed: Sep. 10, 2024. [Online]. Available: http://arxiv.org/abs/2409.05330

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” arXiv.org. Accessed: Sep. 10, 2024. [Online]. Available: https://arxiv.org/abs/2006.11477v3

[13] Z. Liu et al., “KAN: Kolmogorov-Arnold Networks,” arXiv.org. Accessed: Sep. 10, 2024. [Online]. Available: https://arxiv.org/abs/2404.19756v4

[14] K. Wang et al., “MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation,” in Computer Vision – ECCV 2020, vol. 12366, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., in Lecture Notes in Computer Science, vol. 12366. , Cham: Springer International Publishing, 2020, pp. 700–717. doi: 10.1007/978-3-030-58589-1_42.

[15] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, “MakeltTalk: speaker-aware talking-head animation,” ACM Trans. Graph., vol. 39, no. 6, pp. 1–15, Dec. 2020, doi: 10.1145/3414685.3417774.

[16] “A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild | Proceedings of the 28th ACM International Conference on Multimedia.” Accessed: Sep. 10, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413532