포트레이트토커: 텍스트 프롬프트 기반 음성 구동 3D 말하는 얼굴 생성*

DU XIAN[0,1], 유리[1,2,*]
아주대학교 인공지능학과[1], 아주대학교 소프트웨어학과[2]
duxian@ajou.ac.kr, riyu@ajou.ac.kr

PortraitTalker: Speech-Driven 3D Talking Head from Text Prompt

XIAN DU[0,1], Ri Yu[1,2,*]
Dept. of Artificial Intelligence, Ajou University[1], Dept. of Software and Computer Engineering, Ajou University[2]

Abstract

The reliance on reference images or 3D models presents a fundamental limitation for customizable digital avatar creation. We propose PortraitTalker, an end-to-end framework that generates photorealistic 3D avatars directly from text prompts and speech inputs without requiring manual rigging. Our system integrates a diffusion model with score distillation sampling for texture generation and a transformer-based audio encoder to drive FLAME-based facial animation, PortraitTalker achieves state-of-the-art performance on the HDTF dataset, improving lip synchronization (LSE-C: 7.230 , LSE-D: 7.712) and visual quality (FID: 21.997). This work advances automated avatar creation by removing conventional input constraints, enabling scalable applications in AR/VR and intelligent virtual agents.

1. Introduction

Digital avatars are playing an increasingly important role in immersive applications. However, current generation techniques are often constrained by their reliance on input images or predefined 3D templates. Although significant advancements have been made in text-to-3D synthesis [1] and speech-driven animation [2] independently, integrating both into a cohesive pipeline remains challenging, particularly in preserving temporal consistency and visual fidelity.

PortraitTalker addresses these challenges through a unified architecture composed of 3 key components:

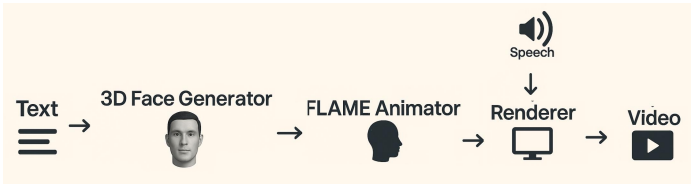

Figure 1: Pipeline of the PortraitTalker

(1) Text-to-3D Synthesis: SDS-optimized diffusion enables high-quality 3D texture generation from textual descriptions;
(2) Speech-Driven Animation: A transformer-based audio processor extracts FLAME-compatible parameters for expressive facial motion;
(3) Differentiable Rendering: A real-time renderer ensures temporal coherence and physical plausibility in the final output.

2. Methodology

2.1. Text-to-3D Synthesis

Our pipeline initiates avatar creation via SDS-optimized diffusion, which distills gradients from a pretrained text-to-image model. This generates a tri-grid representation that jointly encodes geometry and texture. Orthogonal feature planes facilitate efficient synthesis of animation-ready models with high visual fidelity.

2.2. Speech-Driven Animation

A transformer-based audio encoder is employed to predict frame-wise FLAME parameters [3] directly from raw speech input. These parameters capture both expression dynamics and head pose variations. The result is accurate, temporally aligned lip-sync and expressive motion patterns that reflect the speech signal.

* 구두발표논문
* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2025-RS-2023-00255968)

2.3. Differentiable Rendering

Our rendering pipeline employs a differentiable renderer to synthesize video frames. It composites the FLAME-based geometry with hierarchically structured tri-grid textures. The renderer inherently enforces spatiotemporal coherence and physically accurate shading, eliminating the need for post-processing stages.

3. Experiments

3.1. Qualitative Comparison

Our method demonstrates statistically significant improvements across all evaluation metrics on the HDTF dataset [6]. As quantified in Table 1, Portrait-Talker achieves a Lip Sync Error Confidence (LSE-C) score of 7.230, representing relative improvements of 43.2% and 48.4% over MakeItTalk [4] (5.051) and Wang et al. [5] (4.872) respectively. Concurrently, we reduce the Lip Sync Error Distance (LSE-D) by 22.9% (7.712 vs. 9.999/9.995), indicating superior temporal alignment accuracy. In terms of visual quality, our approach establishes a new state-of-the-art Frechet Inception Distance (FID) of 21.997, outperforming both baselines by significant margins.

Method	Lip Synchronization		Video Quality
	LSE - C↑	LSE - D↓	FID↓
MakeItTalk [4]	5.051	9.999	28.183
Wang et al. [5]	4.872	9.995	22.372
Ours	7.230	7.712	21.997

Table 1. Comparison with methods on HDTF [6] dataset

3.2. User Study

We conducted a comprehensive user evaluation with 20 participants assessing 50 generated video samples. As shown in Table 2, our method achieved dominant preference scores across four key perceptual metrics: lip-sync accuracy (68.13% preference), motion diversity (76.89%), video sharpness (74.06%), and overall naturalness (74.76%). Notably, 38% of participants explicitly identified our system as superior specifically for lip-sync quality.

Method	Lip Sync.	Motion Diversity	Video Sharpness	Overall Naturalness
MakeItTalk[4]	9.86%	7.04%	6.72%	9.41%
Wang et al.[5]	22.01%	16.07%	19.22%	15.83%
Ours	68.13%	76.89%	74.06%	74.76%

Table 2: User Study

3.3. Results

Figure 2 showcases five representative keyframes synthesized from the prompt “A casually dressed young adult European male”. The outputs show temporally coherent facial expressions, phoneme-level lip-sync, and consistent identity preservation throughout the animation sequence.

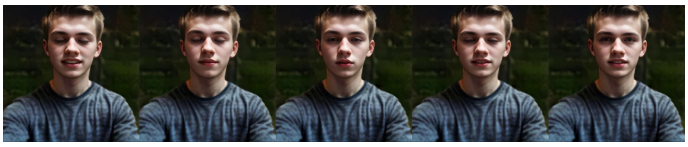

Figure 2: Result of the PortraitTalker.

4. Conclusion

We present PortraitTalker, a novel framework for generating photorealistic, speech-driven 3D avatars from text prompts. By eliminating the need for reference images and manual rigging, Our system enables the creation of high-quality, scalable avatars. Experimental results on both objective metrics and user studies demonstrate superior performance in lip-sync accuracy and video realism. Future work includes enhancing emotional expressiveness, incorporating neural shading for greater realism, and optimizing the model for lightweight real-time deployment.

Reference

[1] Wu Y, Xu H, Tang X, et al. Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior[J]. ACM Transactions on Graphics (TOG), 2024, 43(4): 1-12.

[2] Zhang W, Cun X, Wang X, et al. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 8652-8661.

[3] Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4D scans[J]. ACM Trans. Graph., 2017, 36(6): 194:1-194:17.

[4] Zhou Y, Han X, Shechtman E, et al. Makelttalk: speaker-aware talking-head animation[J]. ACM Transactions On Graphics (TOG), 2020, 39(6): 1-15.

[5] Wang S, Li L, Ding Y, et al. One-shot talking face generation from single-speaker audio-visual correlation learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 2531-2539.

[6] Zhang Z, Li L, Ding Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3661-3670.