ONE-SHOT STYLE-BASED AUDIO-DRIVEN TALKING-HEAD VIDEO GENERATION

Under the Guidance of Dr. Nataraj K S

Presented By: 20BCS059 - Himanshu Shekhar
20BCS125 - Somisetty Sai Praneeth
20BCS128 - Sree Deva Krupananda B Reddy

IIITDWD | 2024

OVERVIEW

- Abstract
- Introduction
- Problem
- Literary Review
- Theoretical
- Objectives
- Hypothesis
- Limitations
- Methodology
- Implementation
- Result
- Conclusion

ABSTRACT

This study pioneers one-shot style-based audio-driven talking-head video generation, aiming to produce authentic videos with accurate lip synchronization, natural head movements, and lifelike eye blinking using minimal data. We evaluate performance across different Indian languages, particularly in voice dubbing and creating suitable lip movements for dubbed audio in movies.

INTRODUCTION

- **The rising importance of bridging real and virtual worlds due to the COVID-19 drives demand for realistic digital twins for applications like virtual conferencing and AR/VR platforms.**
- **One-shot audio-driven talking head generation is emerging as a key technology for synthesizing talking head videos with a single reference image/video and arbitrary audio.**
- **Challenges persist in achieving realistic head poses and eye blinks, prompting recent developments in pose-controllable techniques with limitations in capturing essential facial attributes.**

PROBLEM

- We aim to advance one-shot style-based audio-driven talking-head video generation, evaluating performance across different languages for voice dubbing and creating suitable lip movements for dubbed audio in movies.

- This endeavor involves using minimal data to develop techniques to produce authentic talking-head videos with accurate lip synchronization, natural head movements, and lifelike eye blinking.

LITERARY REVIEW

1. **StyleTalk: This paper presents a pioneering one-shot, style-controllable audio-driven talking face generation framework with a universal style extractor and a dynamic transformer decoder, enabling diversely stylized talking head videos from a single target speaker image.**

2. **Audio2Head: This paper introduces a novel audio-driven talking-head generation framework, featuring a motion-aware recurrent neural network for realistic head movements synced with input audio, an image motion field generator ensuring spatial and temporal consistency, resulting in state-of-the-art photo-realistic videos.**

3. **EAMM: This paper introduces the Audio2Facial-Dynamics module for unsupervised motion representation, the Implicit Emotion Displacement Learner extracting face-related representations from emotion sources, and the Emotion-Aware Motion Model enabling one-shot talking head animations with emotion control.**

Table 1: **The quantitative results on VoxCeleb2.** Lower the better for LMD and LSE-D, and the higher the better for the other metrics. We color the best scores in red and second best scores in blue.

Model	SSIM↑	MS-SSIM↑	PSNR↑	LMD↓	LSE-D↓	LSE-C↑
Wav2Lip [25]	0.95	Nan	31.85	0.97	9.42	5.64
MakeItTalk [48]	0.53	0.48	17.27	1.93	10.77	3.52
Audio2Head [43]	0.42	0.28	12.84	2.71	9.04	5.31
PC-AVS [47]	0.64	0.68	19.63	2.23	7.03	8.02
StyleTalker (ours)	0.63	0.73	19.72	1.90	6.45	8.51

Table 2: **The quantitative results on VoxCeleb2 useen identities.** Details are same with Table 1.

Model	SSIM↑	MS-SSIM↑	PSNR↑	LMD↓	LSE-D↓	LSE-C↑
Wav2Lip [25]	0.95	Nan	32.47	1.02	10.89	4.60
MakeItTalk [48]	0.56	0.48	17.95	1.92	10.45	4.40
Audio2Head [43]	0.47	0.34	15.50	2.76	8.93	5.91
PC-AVS [47]	0.64	0.66	20.03	2.27	6.85	8.62
StyleTalker (ours)	0.62	0.66	19.73	1.91	6.07	9.19

(a) StyleTalker(Base Paper)

Method	MEAD					HDTF				
	SSIM↑	CPBD↑	F-LMD ↓	M-LMD ↓	$Sync_{conf}$↑	SSIM↑	CPBD↑	F-LMD ↓	M-LMD ↓	$Sync_{conf}$↑
MakeitTalk	0.725	0.106	3.969	5.324	2.104	0.593	0.248	5.084	4.447	2.563
Wav2Lip	0.795	**0.178**	2.718	4.052	**5.257**	0.618	0.299	4.544	3.630	3.072
PC-AVS	0.504	0.071	5.828	4.970	2.183	0.422	0.132	10.506	3.931	2.701
AVCT	0.832	0.139	2.923	5.520	2.525	0.755	0.233	2.733	3.610	3.147
GC-AVT	0.340	0.142	8.039	7.103	2.417	0.337	0.296	10.537	6.206	2.772
EAMM	0.397	0.084	6.698	6.478	1.405	0.387	0.144	7.031	6.857	1.799
Ground Truth	1	0.222	0	0	4.131	1	0.307	0	0	3.961
Ours	**0.837**	0.164	**2.122**	**3.249**	3.474	**0.812**	**0.302**	**1.941**	**2.412**	**3.165**

Table 1: The quantitative results on MEAD and HDTF.

(b) StyleTalk

THEORETICAL FRAMEWORK

- **Overview**

  Wave2Lip leads in one-shot style-based audio-driven talking-head video generation, specializing in precise lip movement synchronization for dubbed audio, with a focus on evaluating performance across diverse languages.

- **Proponents**

  Wave2Lip excels in achieving accurate lip movement synchronization for dubbed audio, particularly in evaluating performance across various languages.

OBJECTIVES

- **Objective 1**

  Enhance Wave2Lip's capability for precise lip movement synchronization across diverse languages, focusing on its application in dubbed audio production for movies.

- **Objective 2**

  Evaluate Wave2Lip's performance across various languages to ensure accurate lip movement synchronization in dubbed audio, advancing its applicability in the entertainment industry.

HYPOTHESIS

The Wave2Lip technique, specializing in voice dubbing and accurate lip movement synchronization for dubbed audio in movies and similar applications, can generate high-quality audio-driven talking head videos that synchronize accurately with the input audio, offering significant benefits over existing methods.

LIMITATIONS

- Existing methods achieve precise lip movements for familiar faces but falter with unfamiliar ones, causing lip sync issues.
- Challenges remain in achieving seamless lip sync with audio across varied video recordings.
- Wav2Lip developers overcome this hurdle by allowing the model to adjust lip shapes to match any input video's accompanying audio.

METHODOLOGY

- **Unlike past approaches, Wav2Lip employs a trained discriminator capable of accurately detecting errors in lip-sync videos.**
- **Further training on noise-generated videos enhances its ability to measure inaccuracies, improving lip shape quality.**
- **Wav2Lip integrates an image quality discriminator to enhance image quality in generated videos.**

IMPLEMENTATION

Phase 1

Collect video and audio recordings of talking human faces for training and testing purposes.

Phase 2

Train a Wave2Lip model on standard datasets to improve lip synchronization and facial expressions.

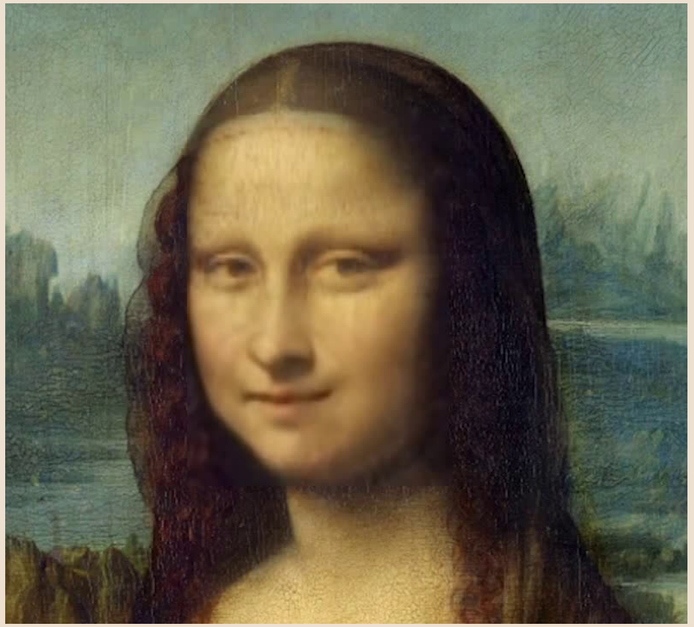
Phase 3

Test the model's ability to generate videos in various Indian languages, assessing its effectiveness.

Phase 4

Assess model performance across languages for voice dubbing and lip sync in dubbed movies.

RESULT



Realistic talking-head videos are generated using different images and videos but with the same audio file

RESULT

Source: Always Filmy

Realistic talking-head videos are generated using same image and video but with the different audio files

PERFORMANCE

	Image and Audio	Video and Audio
Telugu	53.068	10.441
Tamil	**21.243**	06.508
Kannada	28.812	**05.533**
Malayalam	28.194	10.312

LSE Scores obtained for the videos generated using different Languages

CONCLUSION

- **Training Wave2Lip models enhances talking-head video generation, enabling realistic talking videos with accurate lip sync.**
- **Achieving a discriminatory loss of 0.40 in around 100 epochs signifies significant progress in optimizing model performance for authentic video synthesis.**
- **Among different Indian Languages, our model is good in Tamil with Images and Kannada with Videos trained with respective audio.**

OTHER WORKS

- We also tried to pioneer the one-shot style-based audio-driven talking-head video generation field by incorporating MRI video and audio data.
- This ensures anatomical consistency, addressing image artifacts, preserving temporal coherence, and validating results for authenticity in MRI video synthesis using minimal input data.
- This advancement opens new avenues for speech synthesis and related applications.

Original Video

Created Video

OTHER WORKS

- Pre-trained Wave2Lip models facilitate MRI video generation, promising realistic video synthesis from limited input data like MRI scans.
- This approach is still under development, so we might need to experiment with different models and techniques to achieve optimal results for MRI video synthesis.
- However, the lack of large training datasets remains a hurdle. Generative models and medical image synthesis advancements could enable smaller datasets for future MRI video synthesis from audio.

FUTURE WORKS

- Refining model architecture, using advanced lip tracking, audio-visual sync techniques, and larger datasets can enhance performance.
- Integrating Wav2Lip into virtual assistants, chatbots, VR, and AR can improve user experience with lifelike avatars.
- Wav2Lip finds use in speech therapy, aiding communication for speech disorders, and telemedicine, improving articulation and pronunciation.

THANK YOU