Major Project Report

on

One-Shot Style-Based Audio-Driven Talking Head Video Generation

Submitted by

Team Members

20BCS059 – Himanshu Shekhar
20BCS125 – Somisetty Sai Praneeth
20BCS128 – Sreedeva Krupananda B Reddy

Under the guidance of

Dr. Nataraj K S
Assistant Professor
Department of Electronics and Communications Engineering
Indian Institute of Information Technology Dharwad
Karnataka, India-580009

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

22/03/2024

CERTIFICATE

It is to certify that the work contained in the project report titled ***"One-shot Style-based Audio-driven Talking Head Video Generation"*** by Himanshu Shekhar (20BCS059), Somisetty Sai Praneeth (20BCS125), and Sreedeva Krupananda B Reddy (20BCS128) has been submitted for the fulfilment of the requirement for the degree of ***"Bachelor of Technology in Computer Science & Engineering"***. Their work has been found satisfactory and hereby approved for submission.

Signature of the Supervisor

Dr. Nataraj K S

Assistant Professor

Department of Electronics and Communications Engineering

IIIT-Dharwad

DECLARATION

We declare that this written project submission represents our ideas and has not been taken from any other source. We have mentioned and cited the resources of any other publisher wherever needed in the report. We also declare that this work is done by following the principles of academic honesty and integrity and we have done this work with utmost sincerity towards our profession and the institute. We moreover understand the consequences of falsifying any information or misinterpretation of data and figures, hence we have tried to come up and present the best possible response and results.

Himanshu Shekhar – 20BCS059

Somisetty Sai Praneeth – 20BCS125

Sreedeva Krupananda B Reddy – 20BCS129

APPROVAL SHEET

The project entitled ***"One-shot Style-based Audio-driven Talking Head Video Generation"*** by Himanshu Shekhar (20BCS059), Somisetty Sai Praneeth (20BCS125), and Sreedeva Krupananda B Reddy (20BCS128) approved for the degree of Bachelor of Technology in Computer Science & Engineering.

Signature of the Supervisor

Dr. Nataraj K S

Assistant Professor

Department of Electronics and Communications Engineering

IIIT-Dharwad

Head of Department

Dr. Pavan Kumar C

Assistant Professor

Department of Computer Science and Engineering

IIIT-Dharwad

ACKNOWLEDGEMENT

We would like to express our sincere thanks to **Dr Nataraj K S**, Assistant Professor (Dept of Electronics and Communications Engineering) **Dr Sadhvi Manerikar**, Major Project Coordinator for their constant support and guidance throughout the project. It is their constant effort that helped us in completing the project well and on time.

We would also like to thank the **Department of Computer Science IIIT-Dharwad** for the motivation and encouragement, and sincere thanks to **Dr. Pramod Yelmewad,** Major Project Coordinator for providing us with the guidelines and resources to complete this report.

Last but not least a special thanks to the **Library, IIIT-Dharwad** for helping us identify any plagiarism.

CONTENTS

List of Figures

List of Tables

One-shot Style-based Audio-driven Talking Head Video Generation

Abstract

This research 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. Leveraging the Wav2Lip architecture, our approach focuses on evaluating performance across various languages, particularly in voice dubbing and creating suitable lip movements for dubbed audio in movies. We utilized LRS2 datasets comprising audio-visual recordings for training and several data sources from the internet in various Indian languages for testing. Our evaluation reveals exceptional performance, with the model achieving impressive Lip Sync Error (LSE) scores, particularly when trained with images for Tamil and videos for Kannada. The results demonstrate superior accuracy in synchronizing lip movements with the corresponding audio inputs, underscoring the model's potential for generating high-quality talking-head videos across diverse linguistic contexts. Overall, this research highlights the effectiveness of Wave2Lip in addressing the challenges of audio-driven video generation and its promising applications in voice dubbing and creating authentic lip movements for dubbed audio in movies.

Keywords: Wav2Lip, lip synchronization, voice dubbing, authentic videos, language evaluation

1 Introduction

The rapid evolution of virtual communication in response to the COVID-19 pandemic underscores the growing importance of bridging the gap between reality and virtuality. The era of digital twins, or virtual duplicates of actual people that can imitate speech and facial emotions with amazing realism, has begun as a result of this paradigm shift. In this context, audio-driven talking head generation emerges as a pivotal technology for creating lifelike digital avatars, essential for applications like virtual conferencing and VR/AR platforms.

Although earlier research has made strides in generating realistic digital twins, they frequently have no power over multiple identities and struggle to capture nuanced facial motions. One-shot audio-driven talking head generation methods offer a promising solution by synthesizing talking head videos from a single reference image and arbitrary audio.

Nevertheless, current methods are inadequate in accurately replicating facial expressions beyond lip movements. To overcome these constraints, our study explores novel techniques to enhance one-shot style-based audio-driven talking head video generation, with a focus on precise lip synchronization and natural head movements across various languages. By integrating state-of-the-art deep learning techniques and leveraging the power of minimal input data, our research endeavours to pioneer advancements in one-shot style-based audio-driven talking head video generation, paving the way to empower digital content creators with versatile tools for crafting compelling narratives and immersive storytelling experiences.

1.1 Contribution

This paper contributes to the advancement of one-shot style-based audio-driven talking-head video generation by addressing the need for precise lip synchronization across languages, particularly in the realm of dubbed audio production for movies. By introducing novel techniques tailored to Wav2Lip architecture, we aim to enhance its capability for generating high-quality talking-head videos with accurately synchronized lip movements offering significant benefits over existing methods. Additionally, our evaluation of Wav2Lip's performance across various languages contributes to its applicability in the entertainment industry, clearing the path for its extensive implementation in dubbed audio production and other related applications.

An outline of this paper's primary contributions is provided below:

1. Novel techniques to enhance Wave2Lip for precise lip synchronization across languages, focusing on dubbed audio production for movies and other entertainment purposes.

2. Evaluation of Wave2Lip's performance across languages, advancing its suitability for entertainment industry applications.

This is the format for the rest of the paper. Section 2 explores various techniques for synthesizing talking-head videos, outlining their methodologies and discussing their effectiveness in achieving realistic results. Section 3 addresses the problem statement, providing an overview of the challenges in audio-driven video generation and detailing the architecture of the Wave2Lip model. In Section 4, we present the outcomes of our experiments, showcasing the performance of Wave2Lip in generating talking-head videos with accurate lip synchronization and natural movements. We assess its efficacy in various Indian languages and discuss its potential applications in entertainment and communication. Finally, Section 5 concludes the research, emphasizing the importance of advancing audio-driven video generation techniques and highlighting the potential of Wave2Lip in various domains. We also discuss the broader implications of our work and avenues for future research in the field of talking head video synthesis.

2 Literature Survey

2.1 StyleTalker: Audio-driven Talking Head Generation

The proposed StyleTalker model introduces a groundbreaking approach to audio-driven talking head generation, offering a seamless synthesis of realistic videos from a single reference image. Leveraging pre-trained image generators and encoders, StyleTalker accurately synchronizes lip shapes, head poses, and eye blinks with input audio. It incorporates novel components, including a contrastive lip-sync discriminator and a conditional sequential variational autoencoder, to disentangle motion and lip movements while preserving identity. Additionally, an auto-regressive prior augmented with normalizing flow enables complex audio-to-motion modelling. StyleTalker surpasses existing baselines, demonstrating superior perceptual quality and accuracy in lip-syncing. By offering both motion-controllable and completely audio-driven synthesis, StyleTalker showcases its versatility and efficacy in generating natural and diverse talking head videos. Through extensive experiments and user studies, StyleTalker emerges as a state-of-the-art framework, opening new possibilities in audio-driven video synthesis.

2.2 StyleTalk: Talking Head Generation with Controllable Speaking Styles

The research addresses the limitation of existing one-shot talking head methods in generating diverse speaking styles. By introducing a one-shot style-controllable talking face generation framework, the study aims to achieve varied speaking styles in final videos. It proposes a method that extracts dynamic facial motion patterns from a style reference video and encodes them into a style code, which is then utilized to synthesize stylized facial animations from speech content and style. The framework integrates the reference speaking style into generated videos through a style-aware adaptive transformer, ensuring authentic visual effects. Comprehensive experiment results show that the suggested approach is effective in generating talking head videos with diverse speaking styles, achieving authentic visual effects, accurate lip-sync, and better identity preservation compared to existing techniques. This novel approach, termed StyleTalk, presents a significant advancement in one-shot audio-driven talking face generation, promising versatile tools for content creators in crafting compelling narratives and immersive storytelling experiences.

2.3 Audio2Head: Audio-driven Talking Head Generation with Natural Head Motion

The proposed method addresses two primary challenges in generating photo-realistic talking-head videos from a single reference image: producing natural head motions that align with speech prosody and stabilizing non-face regions during large head movements. By designing a motion-aware recurrent neural network (RNN) to predict head poses and a motion field generator to depict entire image motions from audio, head poses, and a reference image, the method achieves spatial and temporal consistency in the generated videos. Comprehensive tests show improved performance when compared to the state-of-the-art, with plausible head motions, synchronized facial expressions, and stable backgrounds. Although the method slightly sacrifices lip-sync accuracy for improved head motion and visual quality, user studies indicate an overall preference for this trade-off. Future work aims to enhance lip-sync accuracy and address limitations in capturing blink patterns and extreme poses or expressions. Despite potential misuse, the method's applications in video conferencing and movie dubbing offer significant positive implications, with code and models to be released to support progress in detecting synthetic videos and ensure responsible usage.

2.4 EAMM: Emotional Talking Face via Emotion-Aware Motion Model

Recent advancements in audio-driven talking face generation have yielded promising results, yet many methods overlook facial emotion or lack applicability to diverse subjects. Addressing this gap, the Emotion-Aware Motion Model (EAMM) is introduced in this paper. EAMM leverages an emotion source video to generate one-shot emotional talking faces, employing an Audio2Facial-Dynamics module to synthesize talking faces from audio-driven unsupervised motion. Additionally, an Implicit Emotion Displacement Learner is proposed to capture emotion-related facial dynamics. Experimental evaluations demonstrate the efficacy of the proposed method, yielding satisfactory results across arbitrary subjects with realistic emotion patterns. The generated emotional talking faces possess substantial application potential such as video conferencing and digital avatars.

However, the potential misuse of such technology for malicious purposes on social media raises ethical concerns. To address these concerns, efforts in deep fake detection are highlighted, emphasizing the importance of realistic and emotional portrait data for improving detection algorithms and promoting positive societal development. Through sharing generated emotional talking face results, this research aims to support the deepfake detection community and foster responsible use of the technology.

3 Methodology

3.1 Problem Statement

The main goal of this research paper is to enhance the effectiveness of one-shot style-based audio-driven talking-head video generation for voice dubbing and lip movement synchronization in movies and other entertainment purposes across various Indian languages. The research aims to develop techniques that utilize minimal data to produce authentic talking-head videos with precise lip synchronization, natural head movements, and realistic eye blinking. This endeavour involves addressing challenges such as accurately capturing diverse linguistic nuances, ensuring seamless lip movement synchronization, and maintaining the authenticity of facial expressions. By evaluating the performance of the proposed techniques across different languages, the research seeks to advance the applicability of audio-driven talking-head generation in entertainment industries while minimizing data requirements and enhancing overall video quality.

3.2 Proposed Method

A. Wav2Lip

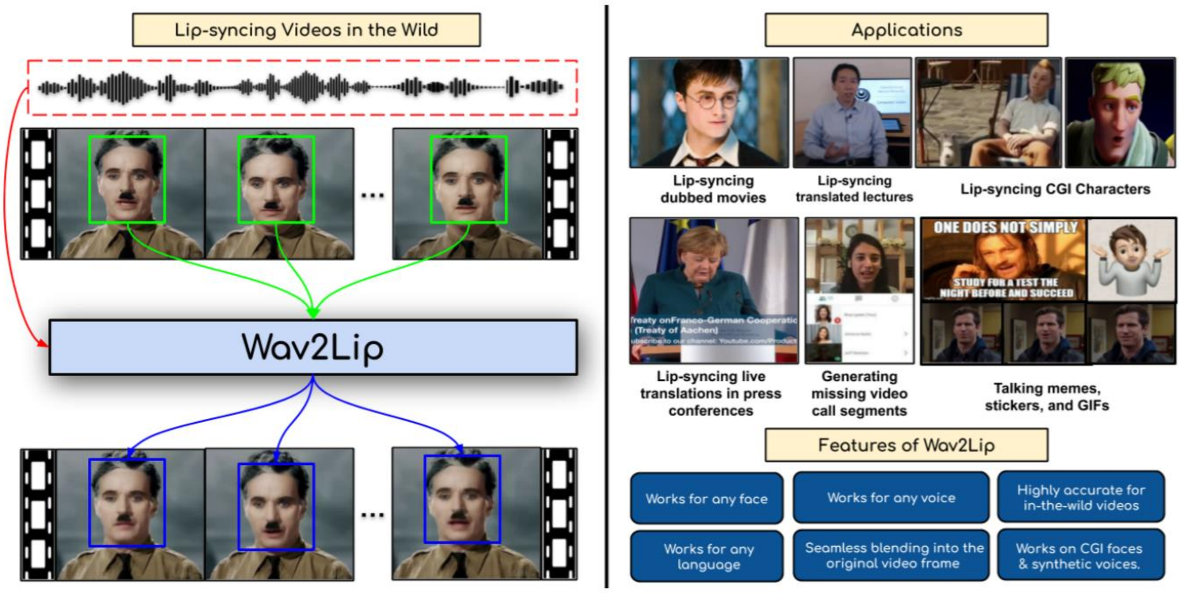

Lip-syncing Videos in the Wild
Wav2Lip
Applications
Lip-syncing dubbed movies
Lip-syncing translated lectures
Lip-syncing CGI Characters
Lip-syncing live translations in press conferences
Generating missing video call segments
Talking memes, stickers, and GIFs
ONE DOES NOT SIMPLY
STUDY FOR A TEST THE NIGHT BEFORE AND SUCCEED
Features of Wav2Lip
Works for any face
Works for any voice
Highly accurate for in-the-wild videos
Works for any language
Seamless blending into the original video frame
Works on CGI faces & synthetic voices.

Fig 3.2.1 Wav2Lip

Source: https://arxiv.org/pdf/2008.10010v1.pdf

The suggested method uses training in conjunction with a lip-sync model that has previously undergone training to synchronize the input video and audio recording. Earlier methods trained the discriminator in the GAN or merely used reconstruction loss for training. Wav2Lip makes use of a trained discriminator that can already identify lip sync video issues with accuracy. After that, noise-generated videos are used to train the discriminator. The discriminator's capacity to measure inaccuracies in the created videos improves with further training, which raises the overall grade of the produced lip forms in the video. Furthermore, Wav2Lip enhances the resulting videos' image quality by employing an image quality discriminator.

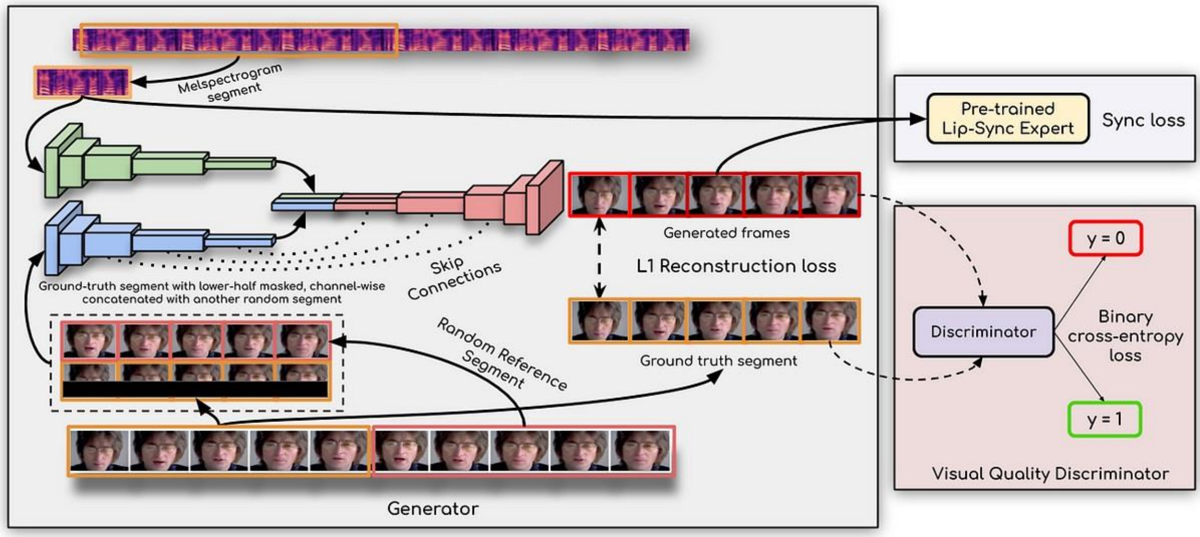

Melspectrogram segment
Pre-trained Lip-Sync Expert
Sync loss
Generated frames
L1 Reconstruction loss
Skip Connections
Ground-truth segment with lower-half masked, channel-wise concatenated with another random segment
Random Reference Segment
Ground truth segment
Generator
Discriminator
y = 0
Binary cross-entropy loss
y = 1
Visual Quality Discriminator

Fig 3.2.2 Model Architecture

Source: https://arxiv.org/pdf/2008.10010v1.pdf

An audio segment's Mel-spectrogram is sent into Wav2Lip along with joining the matching ground truth frames together (the bottom half of which is masked) and a randomly selected reference segment whose speaker verifies that of the ground truth segment. Convolutional layers are used to decrease this input and create a feature vector for the audio and frame inputs. After concatenating these feature representations, a sequence of transposed convolutional layers projects the resultant matrix onto a section of reconstructed frames. Between the levels of the identity encoder and face decoder, there are still skip connections.

Wav2Lip makes an effort to use their masked copies to fully recreate the ground truth frames. Between the reconstructed and ground truth frames, we calculate the L1 reconstruction loss. Next, the ground truth frames and the reconstructed frames are put through the Visual Quality Discriminator and the rebuilt frames are fed through a pre-trained “expert” lip-sync detector. The Visual Quality Discriminator works to improve the frame generator's visual generation quality by differentiating between ground truth and reconstructed frames.

B. Challenges

While past approaches struggled with accurately synchronizing lip movements in videos of individuals not seen during training, Wav2Lip introduces innovative solutions. Unlike previous methods, Wav2Lip employs a trained discriminator to detect errors in lip-sync videos effectively. Moreover, additional training on noise-generated videos enhances its ability to measure inaccuracies, thereby improving the quality of lip shapes.

To further improve the overall image quality of the produced films, Wav2Lip incorporates an image quality discriminator. These advancements address the limitations of existing approaches, allowing the model to adapt the shape of a person's lips to any input video recording based on the accompanying audio, thus improving synchronization and overall video quality.

3.3 Performance Metrics

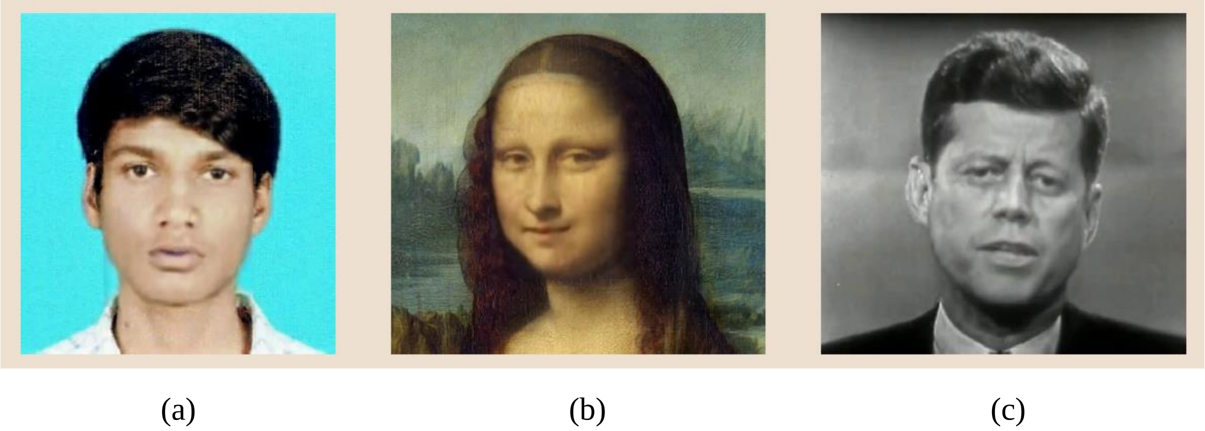
LSE(Lip-Sync Error): This refers to the discrepancy or error between the lip movements of a character in a video/audio clip and the corresponding audio content, especially when they are not in tune with each other. In the context of lip-syncing technology, LSE quantifies how accurately the lip movements match the spoken words or audio signals.

Lower LSE values indicate better synchronization between the lip movements and the audio, thereby enhancing the overall quality and realism of the lip-synced content. To quantify it, various metrics can be used to quantify LSE, including Mean Opinion Score (MOS), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), or other statistical measures. These measurements offer numerical values that show the degree of synchronization between the lip movements and the audio content.

4 Results and Discussion

After training the Wave2Lip model on the LRS2 dataset, which contains diverse video and audio recordings of human faces speaking, we observed significant improvements in lip synchronization accuracy. Throughout the training process, we closely monitored the loss function of the model, which continuously dropped across epochs, indicating effective learning in synchronizing lip movements with audio inputs. Ultimately, we achieved a minimum discriminatory loss of 0.40, demonstrating the success of our training approach.

Following training, we evaluated the model's performance on a subset of test samples from the LRS2 dataset and some other test samples from the internet.

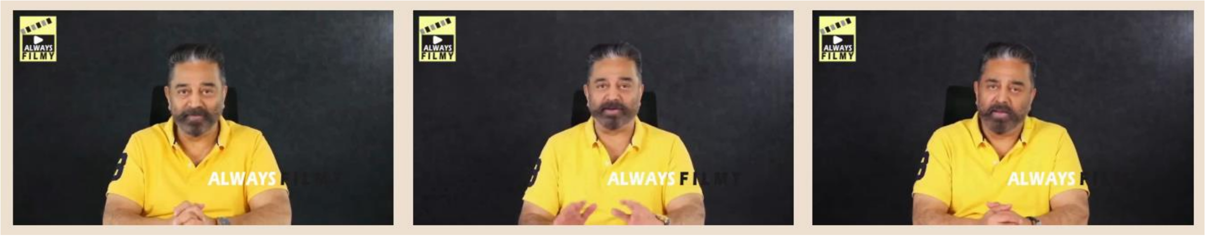


(a) (b) (c)

Fig 4.1 Realistic talking-head videos are generated using different images and videos but with the same audio file

The test samples given to the trained model are; (a) & (b) are the static images of one of our teammates and Monalisa respectively. (c) is the video of John F. Kennedy without audio. These were distributed with a single Hindi-language audio clip. The produced lip movements showed very positive effects closely matching the spoken words in the accompanying audio, indicating the model's proficiency in reproducing realistic lip movements.

Moreover, we extended our evaluation to diverse Indian languages, capturing videos of Kamal Hassan promoting his movie “Vikram” in Telugu, Kannada, Tamil, Hindi, and Malayalam to assess cross-lingual performance. We’ve separated the videos and audio for the respective languages and tested the model with different combinations, for example, an image of a video and any audio, any video without audio and an audio etc. We tried to achieve a bunch of such combinations with images, and videos without audio and different audio files.



(e) (f) (g)

Figure 4.2 Realistic talking-head videos are generated using the same image and video but with different audio files

Source: AlwaysFilmy Youtube Channel

The above input (e) is a static image and video is obtained with Tamil audio. (f) is a video without audio and output is obtained using Telugu audio. (g) is another video without audio and output obtained using Kannda audio. By performing the Lip Synchronization Error (LSE) metric, we compared the original recordings with the model-generated videos. At each time stamp, we compared the frames of the original video vs the generated video and calculated the error using several metrics to quantify LSE. Therefore, the lower the LSE score better the lip-syncing.

	Image and Audio	Video and Audio
Telugu	51.068	10.441
Tamil	**21.243**	06.508
Kannda	28.812	**05.533**
Malayalam	28.194	10.312

Tab 4.1 LSE Scores obtained for the videos generated using different Languages

By observing the values obtained from Tab 4.1, we could conclude that among different Indian languages, our model is good in Tamil with images and Kannada with videos trained with respective audio. The reason behind low scores with videos rather than with images is that we could obtain natural head movements, eye-blinkings and others from video input better than images because there is no such data in the images.

Our qualitative investigation confirmed these conclusions even more, with the generated videos exhibiting high visual fidelity and natural-looking lip movements, enhancing the overall viewing experience. In summary, our experiments with the Wave2Lip model underscore its potential as an effective tool for audio-driven talking-head video generation. With its demonstrated accuracy across languages and high-quality output, Wave2Lip has potential in many applications, from movie dubbing to virtual assistants.

5 Conclusion

Our research demonstrates the substantial advancements achieved in one-shot style-based audio-driven talking-head video generation through the training of Wav2Lip models. By leveraging this approach, we have successfully enhanced the generation of realistic talking videos characterized by accurate lip synchronization, natural head movements, and lifelike eye blinking. The attainment of a discriminatory loss to a minimum of 0.40 and an L1 loss to a minimum of 0.58 implies significant progress in optimizing the performance of our model for authentic video synthesis, thereby underscoring its efficacy in generating high-quality videos.

Furthermore, our findings reveal notable performance outcomes across different Indian languages, particularly highlighting the model's proficiency in Tamil with Images and Kannada with Videos trained with respective audio. These outcomes indicate not just how successful our strategy is but also underscore its potential applicability in diverse linguistic contexts, thereby paving the way for more developments in the video synthesis field and related applications.

Future research in Wav2Lip technology aims to refine model architecture, incorporating advanced techniques such as lip tracking and audio-visual synchronization, and leveraging larger datasets for enhanced performance. Integration into virtual assistants, chatbots, VR, and AR platforms promises to elevate user experiences by offering lifelike avatars with synchronized lip movements, enhancing immersion and engagement. Additionally, exploring applications in speech therapy, aiding individuals with speech disorders, and telemedicine can yield innovative solutions for improving articulation and communication skills. By providing visual feedback through synchronized lip movements, Wav2Lip holds the potential to revolutionize speech rehabilitation efforts and facilitate remote healthcare services, contributing to advancements in healthcare and communication technology.

References

[1] Min, D., Song, M., Ko, E., & Hwang, S. J. (2022). StyleTalker: One-shot Style-based Audio-driven Talking Head Video Generation. arXiv preprint arXiv:2208.10922.

[2] Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., & Yu, X. (2023). StyleTalk: One-shot Talking Head Generation with Controllable Speaking Styles. arXiv preprint arXiv:2301.01081.

[3] Wang, S., Li, L., Ding, Y., Fan, C., & Yu, X. (2021). Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. Retrieved from arXiv preprint arXiv:2107.09293.

[4] Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., & Cao, X. (2022). EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. arXiv preprint arXiv:2205.15278.

[5] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V., & Jawahar, C. V. (2020). A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. arXiv preprint arXiv:2008.10010.