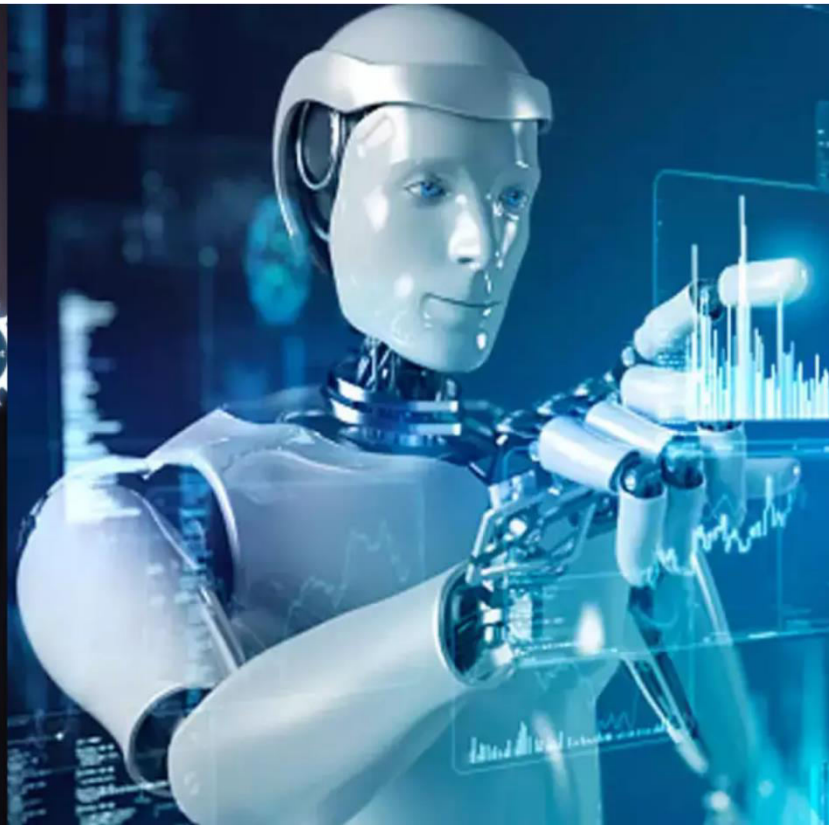
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 1, January 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501044|

Speech-Driven Lip Synchronization using AI

Prof. B.S.Patil, Apurva Auchare, Prachi Kakade, Vaibhavi Khalate, Pratik Shelar

Guide, Department of Artificial Intelligence & Machine Learning Department, AISSMS's Polytechnic, Pune, Maharashtra, India

Department of Artificial Intelligence & Machine Learning Department, AISSMS's Polytechnic, Pune, Maharashtra, India

ABSTARCT: Lip synchronization is a significant task in video synthesis, animation, and other digital applications, which demands perfect synchronization between lip motion and speech audio. The conventional lip synchronization methods are largely manual, time-consuming, and expert intensive, resulting in higher costs and efforts. Keeping these challenges in perspective, this project brings forth an Artificial Intelligence-based lip synchronization system for video synthesis. The primary aim of this project is to develop a system that automatically generates realistic synchronized lip motions for a given audio or text input. The suggested system applies deep learning algorithms to identify related speech features from audio, which are then correlated with related lip motion in video frames. The proposed system applies AI algorithms to generate realistic lip-synced videos that are less time-consuming, less manual, and more accurate compared to conventional lip synchronization methods. The experimental outcome validates that the developed system has been able to generate synchronized videos efficiently with considerable visual quality. The proposed project can be applied effectively in various applications like dubbing, animation, virtual models, e-learning, and digital content production.

I. INTRODUCTION

Lip synchronization is the technique used for synchronizing the lip movements of a person in the video with the input. Lip synchronization is widely used in video generation, animation, video production, dubbing, and content generation. Lip synchronization increases the authenticity as well as the quality of the video, making the video more realistic since synchronization is a part of multimedia communication efficiently, where synchronization is necessary for proper communication in multimedia applications.

Conventionally, lip synchronization has been done through traditional methods such as manual and rule-based animation systems. Manual lip synchronization is done through the use of skilled professionals, where each frame of the lip movement is manipulated frame by frame. Thus, the process of lip synchronization is not cost-effective and time efficient. On the other hand, although rule-based systems are partially automated, the process is not flexible and does not produce realistic results, particularly when handling languages, accents, and speech patterns. Thus, the traditional method is not applicable in. However, due to the rapid advancements in the field of Artificial Intelligence and Machine Learning, complex multimedia processing can now be done in an efficient automated manner. There have been substantial successes witnessed through the use of deep learning models Natural lip-synchronization and voice-over alignment capabilities have now been possible with this advancement in fields related to animation, virtual reality, and digital media production.

The demand for digital content creation, animation, and automated video generation is rising, and this has pointed out the need for efficacious and accurate lip synchronization solutions. Applications like movie dubbing, animated videos, virtual avatars, and educational content generation require perfect alignment between speech audio and lip movements to generate realistic videos. The manual lip-synch approach is cumbersome, laborious, and often full of inconsistencies in such applications, hence not suitable for large-scale video generation. Therefore, an automated approach for lip synchronization is essential to improve productivity along with visual quality.

This project holds a proposed Artificial Intelligence Based Lip Synchronization for Video Generation solution that is intended to automatically create synchronized lip actions in response to a specific audio or text input for a speech activity. The solution proposed for this use case leverages deep learning concepts to evaluate speech features extracted

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501044|

from either generated audio sources or provided audio sources and apply these features to corresponding synchronized lip actions in video frames.

The main aim of this project work is to produce an optimal, dependable, as well as friendly Artificial Intelligence-based solution to the problem of lip sync automation. The system developed has the capability to produce realistic lip-synced videos with minimal human involvement. The project work is applicable to real-life situations such as the dubbing of movies, animation studios, virtual characters, online educational video development, and digital media development. This highlights the key relevance of Artificial Intelligence in contemporary media

II. PROBLEM STATEMENT

Lip synchronization is a crucial aspect of video production, animation, and dubbing. In most cases, lip synchronization is carried out manually in synchronism with audio. This is a laborious process that involves highly trained staff and a considerable amount of time. This makes it significantly costly to produce a high volume of videos manually. Currently, rule-based techniques unable to produce realistic results pertaining to lip synchronization. Thus, there is a significant requirement for developing a system to automatically produce lip-synchronized videos with high accuracy.

OBJECTIVES OF THE PROJECT:

The main objectives of this project are as follows:

1. To design a system that can generate lip-synchronized videos using text or audio input.
2. To reduce manual effort involved in traditional lip synchronization methods.
3. To use Artificial Intelligence techniques for improving the accuracy of lip movements.
4. To generate consistent and visually acceptable lip-synced video output.
5. To develop a simple and user-friendly system for video generation.

II. LITERATURE SURVEY

Lip synchronization has been one of the oldest researched topics related to animation, multimedia, and video processing. Earlier, lip synchronization was mainly performed manually by animators, in which they used to modify lip movements with respect to a speech sound by observing every video frame. This method, although acceptable, required many hours of time and skilled labor. For this reason, manual lip synchronization was not suitable for big projects of video generation or frequent content creation.

Later on, rule-based lip synchronization techniques were developed to reduce manual work. Using pre-defined mouth shapes-also termed as visemes-the systems would map them with speech sounds. While the rule-based approaches helped in basic automation, it is not that flexible. These methods often result in unnatural lip movements, especially when there are different accents in speech, speed variations, or emotional expressions. Due to this, the resultant videos looked unnatural. With the advent of Machine Learning, researchers began working on data-driven approaches for lip synchronization. In these approaches, machine learning was trained with audio as well as video inputs in order to analyze the correlation between the audio speech and the lip motion. These techniques resulted in better performance compared to the traditional rule-based systems. But these early machine learning systems relied on human-selected features heavily, which resulted in weak generalization performance on some datasets.

In recent years, there has been noticeable advancement in lip synchronization using deep learning methods. Convolutional Neural Networks can learn complex patterns from audio inputs as well as video inputs. These models create more accurate and smooth lip movements compared to traditional methods, which require rules to create these movements. Some research has also attempted to enhance video quality, but certain issues related to variation or illumination changes still persist.

On analyzing existing research works, the effectiveness of AI-based techniques is proved to be better than conventional techniques. At the same time, there is a requirement for an easy and automated solution that can be used to generate lip-synced videos from text or audio inputs with very less human intervention. The importance of this research work lies in this aspect.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501044|

III. TECHNICAL ARCHITECTURE

"The technical architecture of the proposed system defines the overall structure and working procedure of the Artificial Intelligence-based lip synchronization system". This proposed system is designed in such a way that it can take text or audio as an input, and the output received is the lip-synchronized video. The designed architecture is divided in multiple modules, where each module is designed for performing a particular task in the generation of a video.

The process starts with the Input Module, where the user inputs text or audio. If the user inputs text, the process starts with converting the text to speech with the help of a text-to-speech module. If the user inputs audio, there is no need to perform any process on it, as it can directly go to the next process for processing. Here, all inputs are converted to audio format for analysis.

The resultant audio after conversion is then handled by the Audio Processing Module. In the module, key speech attributes such as phonemes along with timing information are extracted. This information proves essential in determining how to move the mouth when producing given speech sounds. The audio data after processing is ready to be fed to the deep.

Finally, the **Output Module** provides the generated video to the user in a standard video format. The output video can be viewed, stored, or used for further editing. This modular architecture makes the system easy to understand, maintain, and improve in the future.

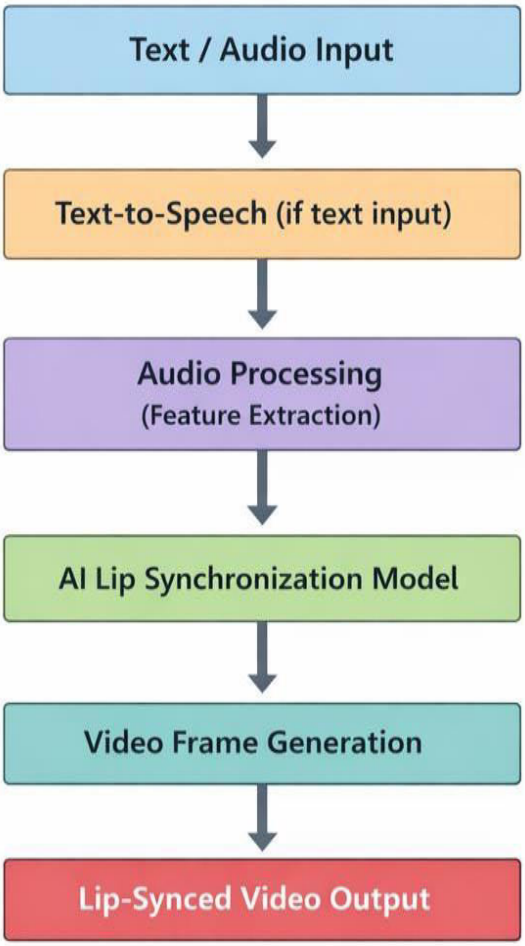


Figure X: Technical Architecture of AI-Based Lip Synchronization System.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501044|

IV. KEY FEATURES

1. **Text and Audio Input Support**
   The system allows users to provide input either in the form of text or pre-recorded audio. Text input is converted into speech before further processing.
2. **Automatic Lip Synchronization**
   Lip movements are generated automatically based on the given audio input, removing the need for manual editing or animation.
3. **AI-Based Processing**
   The project uses Artificial Intelligence and deep learning techniques to analyse speech patterns and generate matching lip movements in the video.
4. **Video Generation Output**
   The system produces a final lip-synchronized video as output, which can be used for further editing or direct viewing.
5. **Reduced Manual Effort**
   Traditional manual lip synchronization requires skilled work and time. This system reduces human involvement and speeds up the video creation process.
6. **Consistent and Accurate Results**
   The AI model provides consistent lip movements for similar speech inputs, improving accuracy compared to manual methods.
7. **User-Friendly Interface**
   The system is designed to be simple and easy to use, making it suitable for users with basic technical knowledge.
8. **Applicable for Multiple Use Cases**
   The generated videos can be used for animation, movie dubbing, educational content, virtual avatars, and digital media creation.

V. FUTURE SCOPE

Although the developed system successfully generates lip-synchronized videos using text or audio input, there are several areas where the project can be further improved in the future. One possible enhancement is improving the video quality and synchronization accuracy, which would make the system more effective for applications such as educational videos, animated content, or virtual presentations. Additionally, the system can be extended to support multiple languages and accents, allowing it to cater to a wider range of users.

Further optimization of the existing model could also reduce processing time and enhance overall performance.

In the future, the system can be extended to support multiple languages and accents. By training the model on diverse datasets, the lip synchronization quality can be improved for different regional languages and speech styles. This enhancement would make the system more flexible and suitable for a wider range of users.

Another improvement can be made by enhancing facial expressions and head movements along with lip synchronization. This would help in generating more natural and expressive videos. Integration with 3D avatars or animated characters can also be explored to expand the application of the system.

The quality of the generated videos can be further improved by using advanced deep learning models and higher resolution video processing. Additionally, the system can be integrated into content creation platforms or mobile applications to make it more accessible for general users.

Overall, the project has good potential for future development, and with further research and improvements, it can be transformed into a more advanced and versatile video generation system.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501044|

VI. CONCLUSION

This project presents an Artificial Intelligence based approach for lip synchronization and video generation using text or audio input. The proposed system aims to reduce the manual effort involved in traditional lip synchronization methods while improving accuracy and consistency. By using AI techniques, the system can automatically generate lip-synchronized videos in an efficient manner. The project highlights the importance of automation in modern video content creation. Overall, this base paper outlines the idea, working concept, and potential applications of the proposed system, which can be further developed and improved in future stages.

VII. ACKNOWLEDGEMENT

I take this opportunity to express my sincere appreciation for the cooperation given by Prof. S. G. GIRAM, Principal of AISSMS'S POLYTECHNIC, Pune and need a special mention for all the help extended by him, constant inspiration and encouragement to make my project a memorable experience. I am thankful to our H. O. D. of Artificial Intelligence & Machine Learning Department, Prof. B.S.Patil for his time-to-time support and valuable guidance. I am deeply indebted to my internal guide Prof.B.S.Patil, for completion of this project for which he has guided and helped me going out of the way. I am thankful to all teachers and professors of our department for sharing with me, valuable knowledge on their respective fields. I would also thank my fellow classmates and friends for their support and timely suggestions. I would also like to thank library staff and laboratory staff for providing me cordial support and necessary facilities, which were of great help for preparing the project report. Thanks to all!

REFERENCES

1) https://www.sciencedirect.com/science/article/pii/S2949719124000323
2) https://arxiv.org/abs/2008.10010?
3) S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: Learning lip sync from audio,” ACM Trans. Graph. (TOG), vol. 36, no. 4, p. 95, 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details