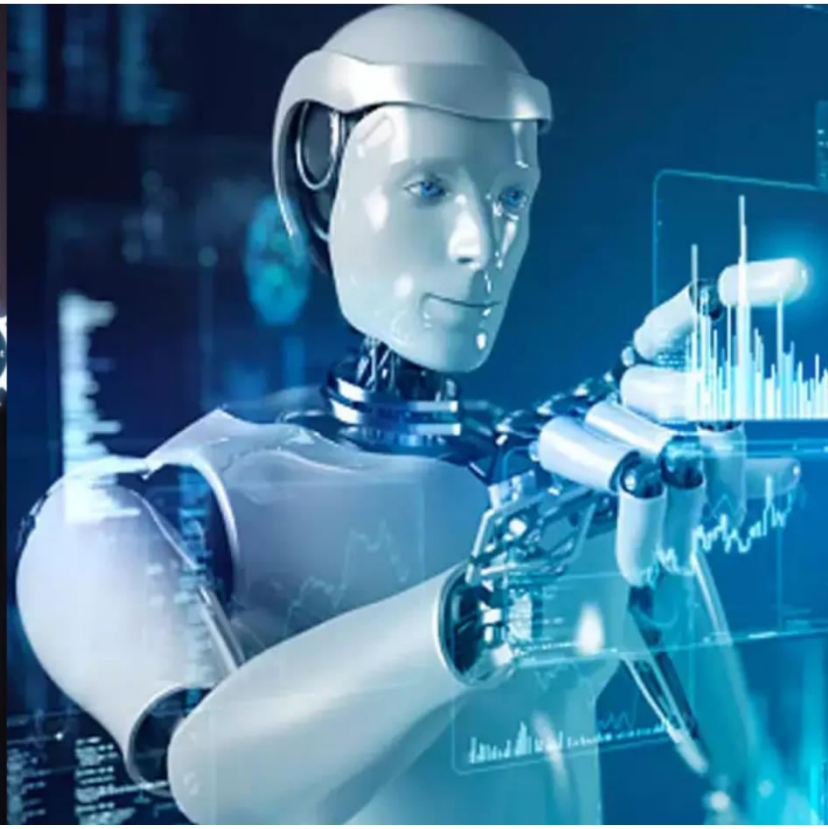
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 3, March 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 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|

TALKSYNC: Speech-Driven Lip Synchronization using AI

Prof. Basipetal, Apurva Auchare, Prachi Kakade, Vaibhavi Khalate, Pratik Shelar

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

Student, Diploma in Artificial Intelligence & Machine Learning, AISSMS's Polytechnic, Pune, Maharashtra, India

ABSTRACT: Lip synchronization plays an important role in animation, video production, and digital media, as the lip movements of a character must align accurately with spoken audio. Traditionally, this process is done manually by animators, which requires considerable time, effort, and professional expertise. To address this challenge, this project introduces TALKSYNC, an artificial intelligence–based system that automatically generates lip-synchronized animated videos using speech input.

The system enables users to choose animated characters from a predefined gallery or generate new characters based on their preferences. To promote responsible use of artificial intelligence and prevent misuse of real human images, the system works exclusively with animated characters rather than real human faces. Users can provide speech input either as text or audio. The system processes this input and synchronizes it with the selected animated character to produce natural and realistic lip movements.

The implementation combines a web-based interface with the SadTalker deep learning model, which generates facial animations from a static image and speech input. The resulting lip-synchronized video is then displayed directly to the user. Overall, the proposed system simplifies the process of creating animated speaking characters and can be useful in applications such as educational content creation, digital storytelling, virtual avatars, and animation production.

I. INTRODUCTION

Lip synchronization refers to the process of aligning a character's lip movements with spoken audio so that the speech appears natural and realistic. It is commonly used in animation, films, video games, digital storytelling, and virtual assistants. When lip movements match the audio accurately, it improves the overall visual experience and makes the communication more engaging for viewers.

In traditional animation workflows, lip synchronization is usually done manually by animators. This involves adjusting the character's mouth movements frame by frame according to the sounds in the dialogue. Although this method can produce high-quality results, it is a time-consuming process and requires a high level of skill and effort. Because of this, manual lip synchronization becomes difficult and inefficient when producing a large amount of content.

With recent developments in Artificial Intelligence (AI) and deep learning, automated lip synchronization systems have started to emerge. These systems can analys speech input and automatically generate matching lip movements, which significantly reduces manual work and speeds up the production process. However, many existing solutions rely on real human images or videos, which can lead to ethical concerns such as misuse of personal images or the creation of misleading media.

To overcome these concerns, this project introduces TALKSYNC: Speech-Driven Lip Synchronization using AI, a system designed to generate lip-synchronized animated videos using artificial intelligence techniques. The system provides a gallery of animated characters from which users can choose, and it also allows users to generate custom animated avatars according to their preferences. By using animated characters instead of real human faces, the system promotes responsible use of AI while still allowing creative video generation. Users can provide speech input either as text or audio, and the system processes this input to synchronize the speech with the selected animated character, producing a lip-synchronized animated video.

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 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|

II. PROBLEM STATEMENT

Lip synchronization is an important part of animation and video production. In traditional methods, the process is usually done manually, where animators carefully edit and adjust lip movements to match the spoken audio. This requires detailed animation work and significant effort, making the process both time-consuming and expensive. Skilled professionals are often needed to ensure that the lip movements accurately match the speech.

Another important issue in modern AI-based video generation systems is the potential misuse of real human images. Many existing systems allow users to manipulate real faces to create talking videos, which can raise serious ethical and privacy concerns. Such technologies may sometimes be used to produce misleading or harmful content. Because of these challenges, there is a need for a system that can automatically generate lip-synchronized videos while also ensuring the responsible use of artificial intelligence. The proposed system addresses this problem by using animated characters instead of real human images. In addition, it automates the lip synchronization process using AI-based techniques, making the system more efficient while reducing the risk of misuse.

III. OBJECTIVES OF THE PROJECT

The main objectives of the TALKSYNC system are as follows:

- To develop an AI-based system that can automatically synchronize speech with animated characters.
- To provide a built-in gallery of animated characters so that users can easily select a character for generating lip-synchronized videos.
- To allow users to create custom animated characters based on their preferences if a suitable character is not available in the gallery.
- To support both text and audio input so that users can generate speech for lip synchronization in different ways.
- To reduce the manual work involved in traditional lip synchronization methods used in animation.
- To generate realistic lip-synchronized animated videos through a simple and user-friendly web interface.
- To promote the ethical use of artificial intelligence by avoiding the use of real human images and instead using animated avatars.

IV. LITERATURE REVIEW

Recent developments in Artificial Intelligence (AI), deep learning, and computer vision have significantly improved the generation of realistic talking face animations. Speech-driven lip synchronization has become an active area of research due to its applications in animation, digital media, virtual avatars, gaming, and online education. Researchers have proposed various techniques to automatically synchronize lip movements with speech audio while maintaining natural facial expressions.

Early research in this field focused on generating talking face animations from static images using audio input. These systems attempted to learn the relationship between speech signals and corresponding lip movements. Machine learning models were trained on datasets of speaking faces to predict mouth shapes based on phonemes present in speech. Although these approaches showed promising results, they often lacked realism and sometimes produced unnatural facial movements due to limited modelling capabilities.

With the advancement of deep learning techniques, more sophisticated models were developed to improve lip synchronization accuracy. One well-known approach is Wav2Lip, a deep learning–based model designed to generate lip movements synchronized with speech audio. The system analyses both the audio signal and visual features of the input face to produce accurate mouth movements that match the spoken words. Wav2Lip has demonstrated significant improvement in lip synchronization accuracy compared to earlier techniques and has been widely used in research related to talking face generation.

During the development phase of this project, the Wav2Lip model was also tested to evaluate its performance. Although the model produced synchronized lip movements, certain visual quality issues were observed in the generated videos. In some cases, the output showed visual artifacts, reduced clarity in facial regions, and unnatural blending around the mouth area, which affected the overall realism of the generated video.

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 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|

Another important advancement in this field is SadTalker, a model capable of generating talking face animations from a single image and speech input. Unlike earlier approaches that mainly focus on mouth movements, SadTalker extracts motion-related information such as lip movements, facial expressions, and head pose from speech signals. This motion information is applied to the input image to generate a more expressive and natural talking animation. Due to its ability to generate stable and realistic facial animations, SadTalker was selected as the core model for the proposed TALKSYNC system. By combining this AI model with a web-based interface and animated character gallery, the system enables users to generate lip-synchronized animated videos while reducing ethical risks associated with the manipulation of real human images.

V. PROPOSED SYSTEM

The proposed system, TALKSYNC: Speech-Driven Lip Synchronization using AI**,** is designed to automatically generate lip-synchronized animated videos using speech input. It provides a web-based platform that enables users to create talking animated characters easily, without requiring professional animation skills.

The platform includes a built-in gallery of animated characters that users can select for generating a video. If a suitable character is not available, users can also create a custom animated character through a chatbot interface. This ensures that the system only uses animated characters, avoiding real human images and minimizing ethical concerns

After selecting or creating a character, users can provide speech input either as text or audio. Text input is converted into speech using a text-to-speech process. The speech input is then processed by the SadTalker deep learning model, which generates accurate lip movements and facial expressions for the selected character.

The generated lip-synchronized video is displayed directly on the same page beneath the input section, allowing users to view the result view the result within a few minutes. Users can also download the video if desired. By automating the lip synchronization process and integrating it within a single web interface, the system reduces manual animation effort and allows quick creation of talking animated videos in a simple, user-friendly environment.

VI. SYSTEM ARCHITECTURE

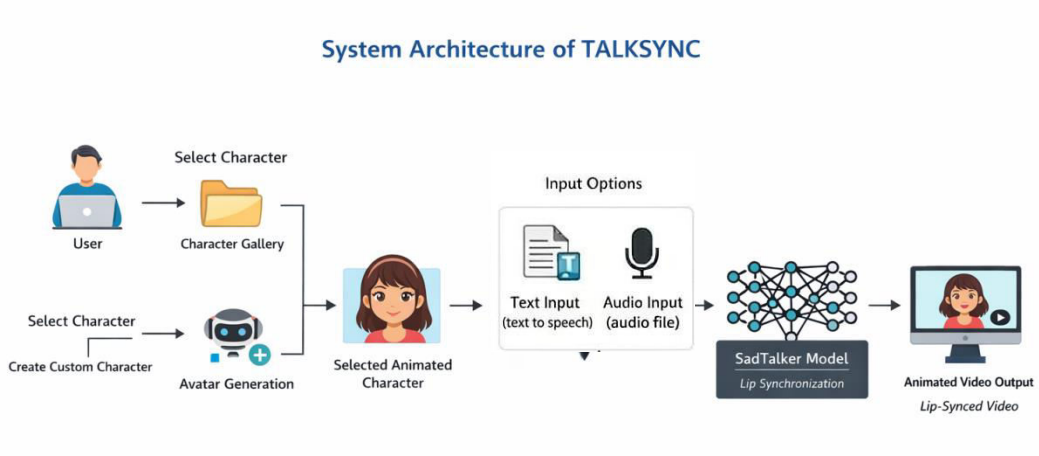


Figure 1: System Architecture of TALKSYNC

Description:
The system architecture of TALKSYNC explains how the data flows through the system to generate the final lip-synchronized video. The process begins with the user interface, where the user selects an animated character from the available character gallery. If the user does not find a suitable character, a new animated avatar can be created using the avatar generation feature.

Once the character is selected, the user provides the speech input. The system supports both text input and audio input. When text is provided, it is first converted into speech using a Text-to-Speech (TTS) module, while audio input can be processed directly by the 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 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|

The generated or uploaded speech is then passed to the SadTalker deep learning model. This model analyzes the speech signal and generates the corresponding lip movements and facial expressions for the selected animated character. By learning the relationship between speech patterns and mouth movements, the model produces synchronized lip animations.

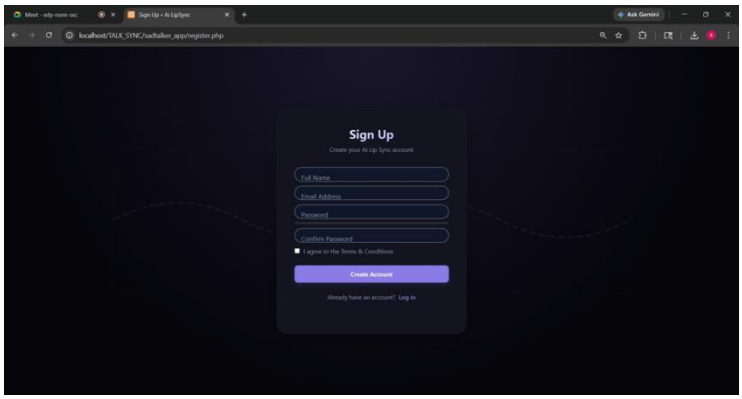
Finally, the generated frames are combined to create the lip-synchronized animated video, which is displayed on the output section of the interface. This architecture allows different modules of the system, such as the user interface, speech processing module, and AI model, to work together smoothly to generate realistic talking animations.

VII. METHODOLOGY

The methodology of the TALKSYNC system describes the step-by-step process followed to generate a lip-synchronized animated video from speech input. The system integrates a web interface, speech processing modules, and the SadTalker deep learning model to automate the lip synchronization process.

Step 1: Character Selection
The process begins with the user selecting an animated character from the built-in character gallery available on the platform. If the user does not find a suitable character, a new animated avatar can be generated using the chatbot-based avatar generation feature.
Step 2: Input Selection
After selecting the character, the user chooses the type of input for speech generation. The system supports two input options: text input and audio input.
Step 3: Text-to-Speech Conversion
If the user provides text input, the system converts the text into speech using a Text-to-Speech (TTS) module. This ensures that both text and audio inputs are converted into a compatible speech format for further processing.
Step 4: Speech Processing
The generated or uploaded speech is processed to extract speech features such as phonemes, timing, and speech patterns. These features help the AI model understand the speech structure and generate appropriate lip movements.
Step 5: Lip Synchronization (SadTalker)
The extracted speech features are fed into the SadTalker deep learning model, which generates accurate lip movements for the selected animated character. SadTalker ensures that the mouth shapes and timing are synchronized precisely with the speech input.
Step 6: Video Generation
After lip synchronization, the system compiles the animated character frames with the processed audio to generate a full video. This includes rendering the character's facial expressions, movements, and other animations to produce a smooth and realistic output.
Step 7: Output Display
Finally, the generated video is displayed on the web interface. Users can view, download, or share the video directly from the platform. The system ensures fast processing so that results can be viewed within a few minutes.

VIII. RESULT



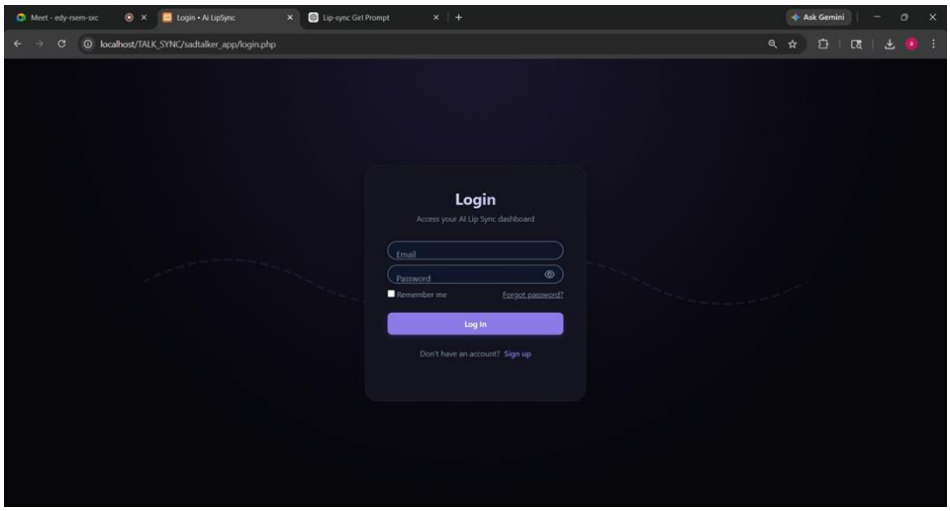
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

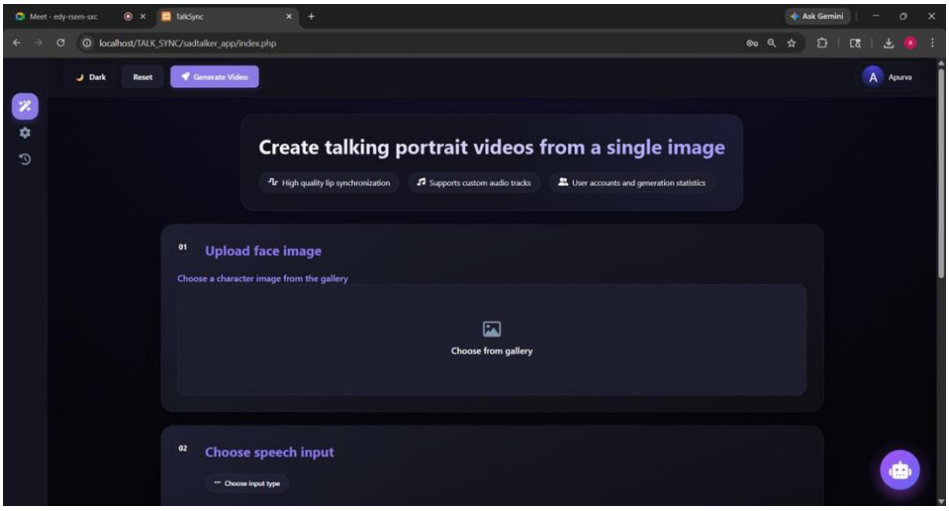
Impact Factor 8.699

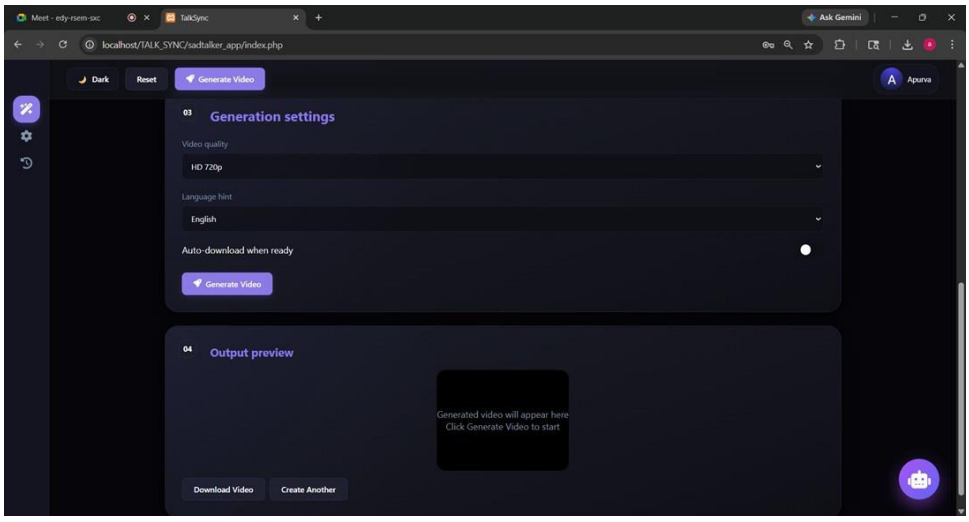
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|







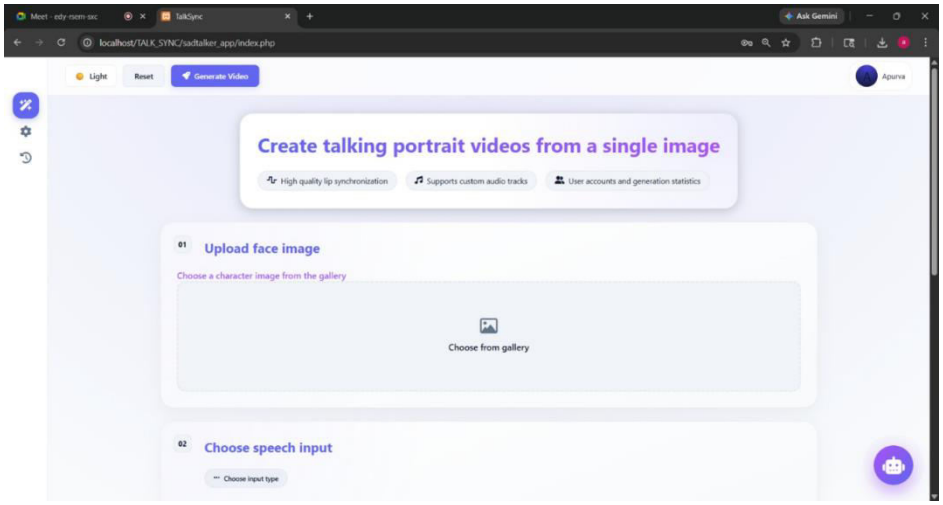
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

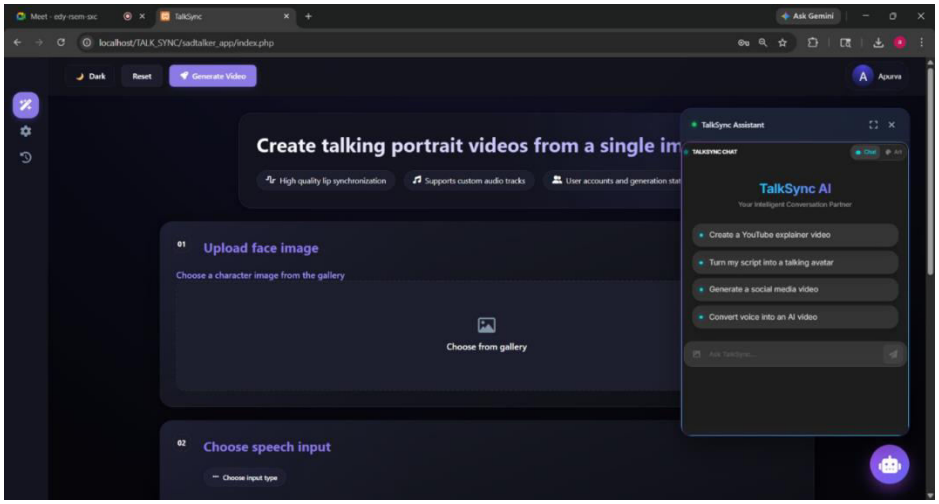
Impact Factor 8.699

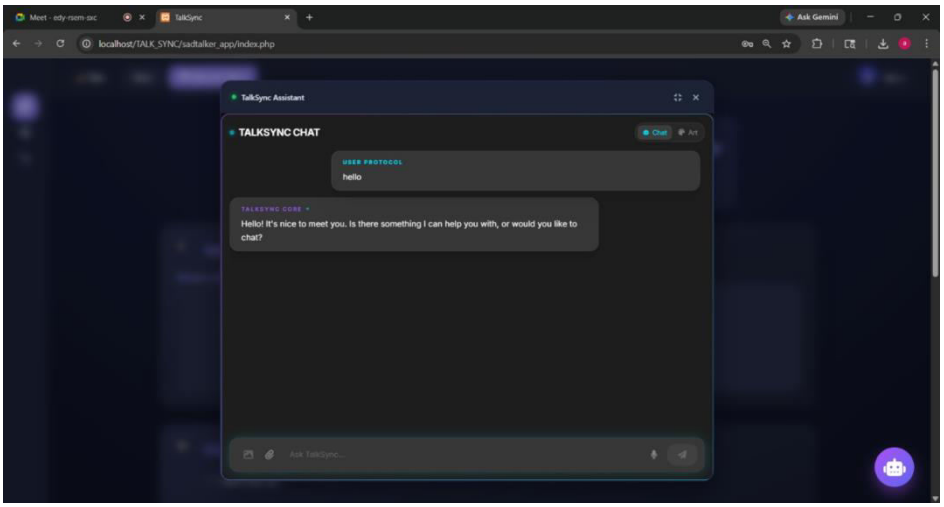
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|







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 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503208|

IX. CONCLUSION

The TALKSYNC system demonstrates how Artificial Intelligence can be used to automate the process of lip synchronization for animated characters. The system allows users to easily generate lip-synchronized animated videos by providing speech input in the form of text or audio. By integrating a simple web interface with the SadTalker deep learning model, the platform simplifies the process of creating talking animations without requiring professional animation skills.

One of the important features of the system is the use of a built-in gallery of animated characters. This helps prevent the misuse of real human images and ensures that the generated content is created using animated avatars only. Users can also generate custom animated characters when needed, which provides flexibility while maintaining the ethical use of AI technology.

Overall, the proposed system successfully reduces the manual effort required in traditional lip synchronization methods and enables quick generation of talking animated videos. The project demonstrates the practical application of AI in animation and multimedia generation while also addressing concerns related to responsible and ethical use of technology.

X. FUTURE SCOPE

The TALKSYNC system can be further improved by adding more advanced features and expanding its capabilities. Currently, the system supports three languages: English, Hindi, and Marathi for speech generation. In the future, support for additional languages can be added so that users from different regions can easily create lip-synchronized animated videos in their preferred language.

Another possible improvement is expanding the collection of animated characters and adding more customization options. Users could be given the ability to modify facial expressions, hairstyles, and other character features to create more personalized animated avatars.

The system can also be enhanced by improving the realism of facial movements and expressions through more advanced AI models. This would help generate smoother and more natural-looking animations.

In addition, future versions of the system may focus on faster processing and real-time video generation. Integration with social media platforms could also allow users to directly share their generated videos online. These improvements would make TALKSYNC more efficient and useful for various multimedia and communication purposes.

REFERENCES

[1] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, “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*, pp. 484–492, 2020.
[2] “Speech-Driven Talking Face Generation Using Deep Learning,” *ScienceDirect*. Available: https://www.sciencedirect.com/science/article/pii/S2949719124000323
[3] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: Learning Lip Sync from Audio,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.
[4] T. Chen, R. Yin, X. Zhou, and J. Wang, “SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016.
[6] Y. Zhou, X. Liu, and Z. Liu, “Talking Face Generation with Audio-Visual Synchronization,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1235–1245, 2021.
[7] H. Xie, Y. Wang, Z. Zhou, and W. Zeng, “MakeItTalk: Speaker-Aware Talking-Head Animation,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, 2020.
[8] J. Wang, X. Qian, and H. Li, “Talking Face Generation Guided by Speech and Facial Motion,” *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

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