Automated Lip Syncing for Chatbots and Virtual Assistants: Enhancing Audio-Visual Interaction for Seamless Communication

Rishi Shah
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT),
Changa, Anand, Gujarat,
22dce113@charusat.edu.in

Nisarg Chaudhari
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT),
Changa, Anand, Gujarat,
22dce011@charusat.edu.in

Tanvi Sheikh
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT),
Changa, Anand, Gujarat,
22dce116@charusat.edu.in

Dweepna garg
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT),
Changa, Anand, Gujarat,
dweepnagarg.ce@charusat.ac.in

Parth Goel
Department of Computer Science & Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT),
Changa, Anand, Gujarat, India
parthgoel.ce@charusat.ac.in

***Abstract*— This paper explores the integration of Wav2Lip lip-sync technology into virtual assistants and chatbots, enhancing their interactivity by turning static images into video with synchronized lip movements. By combining text-to-speech (TTS) and video synthesis, the system aims to create more engaging and human-like digital assistants. Current chatbots and virtual assistants, such as those found on websites and devices like Google Assistant and Alexa, lack personalized visual elements, limiting user connection and engagement. Our proposed solution utilizes the Bark AI library to generate personalized and vivid speech from text, converts images to video using MoviePy, and ensures accurate lip-sync with Wav2Lip technology. The final output is a video where the assistant's lips move in sync with the spoken audio, providing a more realistic and personalized user experience.**

Keywords— Chatbots, Virtual Assistant, Communication

I. INTRODUCTION

Virtual assistants and chatbots have become important tools for handling tasks and query solutions. However, they lack a personalized human-like presence, relying instead on text and voice interactions. By introducing more interactive components, these systems can transform from static, emotionless tools into more engaging experiences. Wav2Lip lip-sync is an exciting method that overcomes this issue by improving the visual representation of the virtual assistant. This method transforms a simple image into a video in which the assistant's lips move in time with the spoken words, making interactions feel more authentic and human-like. Researchers have explored lip-syncing technologies for applications in e-learning and customer service, noting that incorporating visual representation through video and synchronized speech can drastically enhance engagement and improve information retention [1]. Moreover, lip-sync technologies have been successfully applied in diverse environments, including the creation of multilingual avatars for broader reach and communication [2].

Current chatbots and virtual assistants, which are commonly available on websites and smart devices such as Google Assistant and Alexa, sometimes lack a visual component, reducing emotional connection and interaction with customers. Many websites now use chatbots for handling customer inquiries and support, but they are usually text-based, limiting interaction to words on a screen. The lack of a speaking avatar in these systems limits the potential for a more personalized and engaging experience. By including lifelike, speaking avatars, these technologies have the ability to make discussions more natural and engaging. This additional visual aspect may help clients feel more connected to the assistant, resulting in higher overall satisfaction and user engagement. Prior work by Iannizzotto et al. demonstrated that virtual assistants equipped with visual and speech synchronization received overwhelmingly positive feedback from users, who reported a preference for engaging with virtual assistants that had a more human-like visual and auditory presence [3]. Similarly, other studies have shown that integrating both verbal and non-verbal communication methods, like lip-sync video, into virtual assistants can significantly improve customer service and satisfaction [4].

In order to achieve this enhanced experience, we propose using Bark AI's text-to-speech (TTS) technology, resulting in a more personalized and realistic speaker experience than typical solutions such as pyttsx3. The generated voice is then synchronized with an image, which is converted to video using the MoviePy library. The Wav2Lip technology allows perfect lip-sync between the generated video and voice, resulting in a more lifelike and engaging interaction between the user and virtual assistant. Using Bark AI, the virtual assistant can give more realistic and expressive discussions, improving the overall user experience. This approach aligns with other text-to-speech advancements, such as the work of Wang & Székely, who highlight the importance of realistic speech generation in enhancing the user's

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

interaction with digital agents [5]. Mekni et al. found that when virtual assistants simulate human-like behaviors—such as natural voice modulation and lip-syncing— users tend to experience higher satisfaction and are more likely to develop an emotional connection with the technology [6]

The main contributions of this paper are summarized below:

- Bark AI is utilized to generate multilingual text-to-speech, allowing the virtual assistant to adapt seamlessly to various languages and voices.
- MoviePy transforms static images into videos, providing a visual presence for the assistant without requiring complex video editing.
- Wav2Lip technology synchronizes lip movements to the generated speech, creating a realistic, speaking appearance for the assistant.
- This accessible system can be applied in fields like customer service and education, enhancing the engaging qualities of virtual assistants through both speech and visuals.

Our structure of the paper is as follows:

In section II, an overview of related works is given. Section III explains the proposed methodology. Results and discussion of the paper is highlighted in section 4. Finally, the conclusion and future direction summarized in section 5.

II. Related Works

Xu, A. et al. created a customer service chatbot in the year 2017. This chatbot was trained using LSTM networks and evaluated on the basis of 4 things of responses by conducting a survey. The chatbot response was as same as the request was. If it was emotional then the reply is emotional and if it was informational(seeking) the reply was with same informational tone. These tells us that usage of chatbots can have a different impact on users which can significantly solve their problems [7].

Dibitonto, M. et al. conducted survey and implement a Virtual assistant named LiSA (Link Student Assistant) to gather information of University , college events etc. They found good statistical data about the usage of students to retrieve University data by different means. They tested the Chatbot like this and found a very good response in it[[8].

Tulshan A. S. & Dhage S. N. conducted a comprehensive survey on Google Assistant, Siri, Cortana, and Alexa, analyzing their voice- based functionalities[9].

Prajwal et al. concerning your work on converting text to audio, image to video, and creating audio-lipsynced video using Wav2Lip is the method they developed for realistic lip synchronization in unconstrained environments[10].

Wu X. et al. addressed challenges in speech-driven video synthesis, such as achieving precise lip-sync and high video quality, but noted that recent tools like Wav2Lip offer highly accurate lip-syncing, making them valuable in developing realistic virtual assistants [11].

Gupta, A. et al. highlighted that while tools like Wav2Lip achieve reliable lip-sync across different identities, they currently face limitations in generating ultra-high-resolution videos, which are essential for realistic virtual assistant applications [12].

Schumacher, D., & LaBounty Jr. discussed recent advancements in models like Bark, which show promise in generating realistic, multilingual speech. However, they emphasized the need for further improvements in accuracy and control for optimized performance [13].

Tathe, A. et al. explored end-to-end speech synthesis models like Bark, XLSR Wav2Vec2, and mBART, noting their effectiveness in seamless text-to-audio conversion and cross-lingual applications [14].

Table 1 Previous works and their findings

Author(s)	Year	Focus Area	Key Contribution
Xu, A. et al [7].	2017	Customer Service Chatbot	Developed a chatbot for customer service, using LSTM networks to tailor responses based on emotional and informational tones, significantly improving user satisfaction.
Dibitonto, M. et al [8].	2018	Virtual Assistant for University Information	Designed LiSA, a virtual assistant to help students access university data, events, and information efficiently, showing positive user engagement.
Tulshan, A. S. & Dhage, S. N [9].	2019	Comparison of Popular Virtual Assistants	Conducted a comprehensive survey on Google Assistant, Siri, Cortana, and Alexa, analyzing their voice-based functionalities and user interaction effectiveness.
Prajwal et al [10].	2020	Realistic Lip-Syncing in Videos	Developed a method for realistic lip synchronization

			in unconstrained environments using Wav2Lip, enhancing lip-sync precision across different settings.
Wu, X. et al [11].	2021	Speech-Driven Video Synthesis	Identified challenges in lip-sync precision and video quality, highlighting Wav2Lip's advancements in accurate lip-sync for realistic virtual assistants.
Gupta, A. et al [12].	2022	High-Resolution Lip-Synced Videos	Analyzed Wav2Lip's performance across identities, noting its limitation in generating ultra-high-resolution videos essential for lifelike virtual assistants.
Schumacher, D. & LaBounty Jr [13].	2023	Multilingual Speech Synthesis	Examined Bark's capabilities in realistic multilingual speech generation, noting areas for improved accuracy and control to optimize virtual assistant performance.
Tathe, A. et al [14].	2023	End-to-End Speech Synthesis	Explored the use of Bark, XLSR Wav2Vec2, and mBART models in seamless text-to-audio conversion, particularly for cross-lingual applications in virtual assistants.

Table I describes the key findings from the similar research work done on the topic and the observation from the paper. It also highlights advancements in enhancing virtual assistants, particularly in speech synthesis, lip-syncing, and multilingual support, while noting ongoing challenges in achieving high realism and user engagement.

III. Proposed Methodology

This section will cover the whole workflow of the Video of Chatbot/Virtual assistant. It is mainly divided in three main sections done by using different technologies ensuring smooth streamlining of all three to make the workflow perfect.

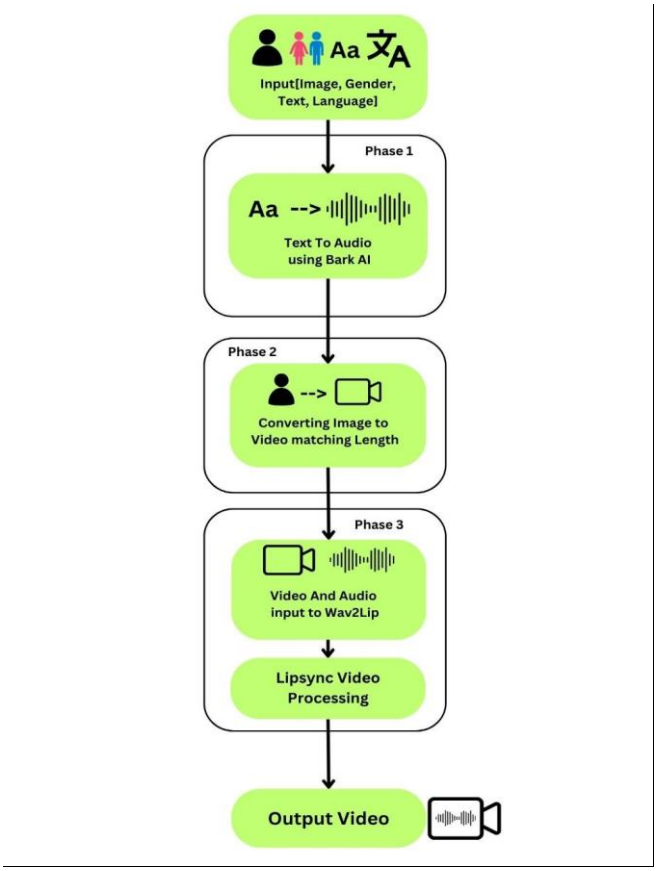

Figure 1 Workflow of the Video Chatbot/ Virtual Assistant

Figure 1 illustrates the workflow for generating a video with synchronized audio and lip movements using Bark AI and Wav2Lip. In this process, users provide inputs such as an image, text, language, and gender. In Phase 1, Bark AI converts the text to audio. Phase 2 then converts the static image into a video that matches the audio length. In Phase 3, the video and audio are input into Wav2Lip, which processes the lip synchronization to create a realistic speaking appearance. The final output is a video with both synchronized audio and visual elements.

A. Phase 1

The first phase involves converting the input text into audio, forming the foundation of the project. For this, we use Suno-Bark AI, a Text-To-Speech (TTS) module known for its wide-ranging features. Bark is ideal for generating speech due to its support for multiple languages, including English, Hindi, French, and German, as well as its ability to choose between male and female voices. One of Bark's standout features is its capacity to add natural elements like background noise, humming, or laughter, making the audio more engaging and dynamic.

Bark stores its speaker data in numpy arrays (.npy files) via their Hugging Face repository, with information on coarse, semantic, and fine speech processing. With over 130 voices available, Bark ensures the audio fits seamlessly with the multilingual virtual assistant's goals, creating a realistic and lifelike experience.

B. Phase 2

Once the audio is generated, the next task is to create a video that matches the duration of the audio. This is crucial for accurate lip synchronization in later stages. In this phase, we use MoviePy, a Python library for video creation and manipulation. The process begins by taking a still image as input and creating a video that lasts for the entire duration of the audio.

MoviePy allows precise timing adjustments to ensure that the video aligns perfectly with the audio file. This step is essential for setting the stage for lip synchronization, as any discrepancies between the audio and video would lead to poor results. MoviePy's flexibility also allows for additional edits such as transitions and text, giving us the freedom to fine-tune the video to meet the project's requirements.

C. Phase 3

The final phase focuses on syncing the virtual assistant's lip movements with the generated speech using Wav2Lip. This technology analyzes the audio and generates lip movements for the video, whether based on a still image or a moving video. By doing so, it adds a level of realism that is crucial for making the virtual assistant appear natural and engaging.

Wav2Lip uses a Generative Adversarial Network (GAN) consisting of a generator and a discriminator. The generator creates the lip movements, while the discriminator checks their accuracy until the results look natural. Wav2Lip's architecture includes a face encoder, an audio encoder, and a face decoder. The face encoder captures facial features, while the audio encoder analyzes speech features like tone and rhythm. These components come together in the face decoder to generate synchronized lip movements.

By combining these three phases—text-to-audio conversion, image-to-video creation, and lip synchronization—we can create a lifelike virtual assistant capable of delivering personalized and engaging experiences for users. Each phase ensures that the final product runs smoothly, with speech and visual elements perfectly synchronized.

IV. RESULTS & DISCUSSIONS

The system developed in this project successfully achieves the desired output of generating lip-synced videos for virtual assistants and chatbots. Using the combination of Bark AI for text-to-speech (TTS) and Wav2Lip for lip synchronization, along with MoviePy for video creation, the model can take a static image, convert it into a video, and sync the lips to the speech. The resulting videos maintain a high level of accuracy in lip movement, aligning perfectly with the spoken audio. This makes the virtual assistant appear more natural and engaging, significantly enhancing user interaction.

Multiple languages and speakers were tested in the project, including English, Hindi, French, and German, with both male and female voices. Bark AI's flexibility allowed for smooth text-to-speech conversion across these languages, with the audio produced being clear and lifelike. The Wav2Lip model ensured that the lips moved in sync with the generated audio, regardless of the language or speaker selected. The resulting videos are dynamic and visually appealing, with no noticeable delays or mismatches between the audio and lip movements.

The use of a single still image to create the video through MoviePy proved to be effective. The video length was adjusted to match the duration of the audio, ensuring that Wav2Lip could perform accurate lip-syncing. This process, combined with the ability to customize the virtual assistant's appearance and voice, makes the model highly versatile for a range of applications, including chatbots, customer service tools, and digital assistants.

The system excels in creating realistic virtual assistants that both speak and visually engage with users. Its accurate lip-sync, multi-language support, and smooth integration of text-to-speech and video generation make this project a significant advancement toward more human-like digital interactions.

The potential for chatbots and virtual assistants to become integral in daily life is rapidly increasing, especially as businesses and tech companies adopt them for customer support and user interaction. This project illustrates how adding a visual, lip-synced face to these assistants, similar to services like Google Assistant or Alexa, can significantly elevate the user experience, making interactions more engaging and relatable. Initially, we encountered challenges converting text to audio with pyttsx3 but successfully transitioned to Bark AI via Hugging Face, which provided better multilingual support. After testing various lip-syncing options, Wav2Lip emerged as the top choice for accuracy, while Python's MoviePy efficiently converted images into videos, streamlining the workflow. As companies integrate such technologies, the capacity for personalized, visually engaging virtual assistants will set new standards in customer service, education, and more, making interactions feel more personal and human-like.

Output video link: -
https://youtube.com/shorts/oZ0zjm7s704?feature=share

Figure 2 Sample Image as Input

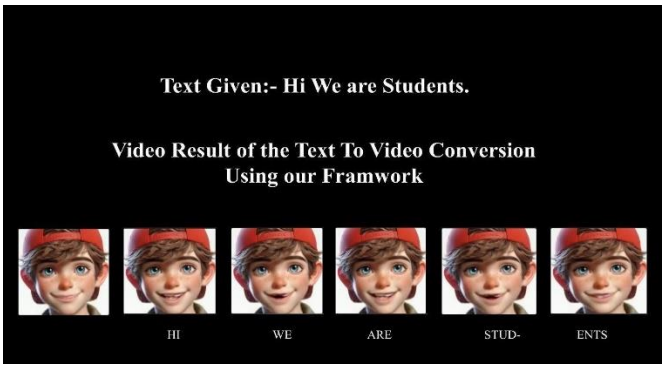

Figure 3 Images of Lipsync video

Figure 2 displays a sample input image of the animated character used for the text-to-video conversion. This image serves as the reference for the character's appearance, including details like facial features and expressions.

Figure 3 showcases frames from the generated lip-synced video sequence. Here, the text "Hi, we are students" has been converted into individual frames where the animated character articulates each word with precise lip movements. The frames demonstrate the effectiveness of the framework in creating realistic, synchronized lip-sync, where the character's mouth movements and expressions align seamlessly with the audio, providing an engaging and lifelike output video.

V. Conclusion and Future Works

The developed system presents a robust framework for generating lifelike, lip-synced videos from text, demonstrating strong potential to elevate virtual assistant technology. By integrating Bark AI for multilingual text-to-speech, MoviePy for converting static images into videos, and Wav2Lip for precise lip synchronization, the system produces virtual assistants that communicate in natural voices and visually synchronize lip movements with spoken words. This combination enhances user engagement, offering more relatable interactions across customer service, education, and entertainment. Key contributions include Bark AI's generation of adaptable, natural-sounding speech in multiple languages; MoviePy's efficient, accessible approach to dynamic video creation; and Wav2Lip's delivery of realistic, human-like visual experiences. Together, these elements create a scalable, versatile framework for globally adaptable virtual assistants that engage users audibly and visually.

Future work will focus on integrating Large Language Models (LLMs) to enable more sophisticated conversations, incorporating facial expressions and gestures for greater realism, and expanding language support to improve the system's global usability and interactivity.

References

[1] S. Abraham, V. Ewards, and S. Terence, "Interactive video virtual assistant framework with retrieval augmented generation for e-learning," in 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2024, pp. 1192–1199.

[2] G. Nanditha, K. V. Datla, G. Kevin, R. Nikitha, L. Pallavi, and C. M. Babu, "Multilingualsync: A novel method for generating lip-synced videos in multiple languages," in 2023 3rd Asian Conference on Innova- tion in Technology (ASIANCON). IEEE, 2023, pp. 1–5.

[3] G. Iannizzotto, L. L. Bello, A. Nucita, and G. M. Grasso, "A vision and speech enabled, customizable, virtual assistant for smart environments," in 2018 11th International Conference on Human System Interaction (HSI).IEEE, 2018, pp. 50–56.

[4] V. Obradovi´c, I. Rajak, M. Se˘cujski, and V. Deli˘c, "Text driven virtual speakers," in 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022, pp. 1170–1173

[5] S. Wang and ´E. Sz´ekely, "Evaluating text-to-speech synthesis from a large discrete token-based speech language model," arXiv preprint arXiv:2405.09768, 2024.

[6] M. Mekni, Z. Baani, and D. Sulieman, "A smart virtual assistant for students," in Proceedings of the 3rd International Conference on Applications of Intelligent Systems, 2020, pp. 1–6.

[7] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in Proceedings of the 2017 CHI conference on human factors in computing systems, 2017, pp. 3506–3510

[8] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a campus environment: design of lisa, a virtual assistant to help students in their university life," in Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20. Springer, 2018, pp. 103–116

[9] A. S. Tulshan and S. N. Dhage, "Survey on virtual assistant: Google assistant, siri, cortana, alexa," in Advances in Signal Processing and Intelligent Recognition Systems: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers 4. Springer, 2019, pp. 190–201.

[10] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 484–492.

[11] X. Wu, P. Hu, Y. Wu, X. Lyu, Y.-P. Cao, Y. Shan, W. Yang, Z. Sun, and X. Qi, "Speech2lip: High-fidelity speech to lip generation by learning from a short video," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22 168–22 177.

[12] A. Gupta, R. Mukhopadhyay, S. Balachandra, F. F. Khan, V. P. Namboodiri, and C. Jawahar, "Towards generating ultra-high resolution talking-face videos with lip synchronization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5209–5218.

[13] D. Schumacher and F. LaBounty Jr, "Enhancing bark text-to-speech model: Addressing limitations through meta's encodec and pretrained hubert."

[14] A. Tathe, A. Kamble, S. Kumbharkar, A. Bhandare, and A. C. Mitra, "End to end hindi to english speech conversion using bark, mbart and a finetuned xlsr wav2vec2," arXiv preprint arXiv:2401.06183, 2024.