Avatar Video Generator with Lip-Sync

Kartarsingh Gothwal[1], Krushna Gore[2], Tanvi Gunjal[3], Sejal Hage[4]
[1,2,3,4]Dept of Computer Engineering, Vishwakarma Institute of Technology, Pune, India
doi.org/10.64643/IJIRTV12I6-187428-459

***Abstract*—Deep learning advancements have made tremendous progress in AI-based avatar creation. This research presents a holistic system that integrates generative adversarial networks (GANs) and transformer-style architectures to generate high-fidelity, customized avatars from sparse input data. The system proposed here offers an avatar video generator with lip-sync functionality to produce realistic digital human likenesses with synchronized voice. The system utilizes several state-of-the-art deep learning models, such as Wav2Lip and its variants, to provide robust lip-sync between audio inputs and avatar movement. Our solution provides flexible input methods, including voice commands, text, audio file uploads, and video dubbing. We have created an easy-to-use interface with several avatar choices for individualized content generation. In addition, the system includes functionality for social media content creation using integration with big language models such as Gemini 1.5-flash. The authentication mechanism employs JWT-based login via Google accounts, with users' data maintained in MongoDB Atlas. Our testing proves the system's competence across different use cases such as multilingual content generation, educational use, and social media interactions. This work addresses implementation specifics, system structure, evaluation findings, and perspectives for further development of digital human generation.**

***Index Terms*—Network Security, Machine Learning, Anomaly Detection, Random Forest, Isolation Forest, Traffic Classification, Cybersecurity, Flask, Scapy, Real-time Analysis**

I. INTRODUCTION

Over the last few years, the convergence of computer vision, natural language processing, and deep learning has changed digital human generation. Generating natural, lip-synchronized avatars has uses that range from education to entertainment, customer support to accessibility. Demand for systems able to create natural-looking digital humans that can communicate naturally is escalating rapidly as virtual interaction becomes a more integral part of our lives.

This paper presents a comprehensive avatar video generator with lip-sync capabilities, designed to create realistic digital human representations with properly synchronized speech. Our system addresses several challenges in this domain, including:

1. Achieving precise synchronization between audio and lip movements
2. Supporting multiple input modalities (voice, text, audio files)
3. Enabling multilingual content through video dubbing
4. Generating contextually relevant content for social media applications
5. Providing an intuitive, user-friendly interface

The proposed system leverages deep learning models like Wav2Lip and its variants for lip synchronization, while integrating with language models like Gemini 1.5-flash for content generation. Our implementation enables users to interact with the system through voice commands, text input, or uploaded audio files, with the option to select from multiple avatar styles.

A distinctive feature of our system is the video dubbing capability, which extracts audio from user-uploaded videos, converts it to text, translates it to a target language using advanced language models, converts the translated text back to speech, and finally generates a lip-synced video in the target language. Additionally, our social media content generator allows users to specify topics, automatically generates appropriate scripts, and creates avatar videos tailored for digital platforms.

In order to provide safe access, we've used JWT-based authentication with Google accounts with user data saved in MongoDB Atlas. The all-inclusive web interface offers a user-friendly experience on every feature, bringing complex avatar generation within reach of users of any technical proficiency.

Proposed work describes the architecture, deployment, evaluation measures, and future applications of our avatar video synthesizer, as well as its contributions to the evolving field of digital human synthesis.

II. LITERATURE RIVEW

The most recent advances in audio-driven video synthesis have leaned heavily on diffusion and transformer architectures to push the boundaries of fidelity, scalability, and real-time performance. For instance, Chatziagapi et al. [1] presented AV-Flow, a unified framework that transforms text into synchronized audio-visual interactions, pointing toward end-to-end multimodal synthesis. In parallel, Oskooei et al. [2] developed a multilingual lip-sync system enabling real-time face-to-face translation, while an anonymous study [3] explored NeRF-based avatars, capable of generating real-time audio-driven talking heads with 3D realism. Complementing these, Wu et al. [4] introduced Speech2Lip, a decomposition–synthesis–composition pipeline that produces high-fidelity lips from just a short reference video. Zhang et al. [5] developed Muse Talk, which leverages latent space inpainting for real-time lip-sync at over 30 fps, while Li et al. [6] proposed Latent Sync, an audio-conditioned latent diffusion model with strong temporal consistency. For 3D avatars, Lin et al. [7] created GL DiTalker, combining speech cues with graph latent diffusion transformers to yield expressive facial animations. Additionally, Park et al. [8] advanced alignment techniques with Sync Talk Face, using audio-lip memory to achieve precise synchronization. These recent works collectively illustrate a shift toward scalable, real-time, and multimodal lip-sync systems.

Moving slightly earlier, innovations in contrastive learning and multimodal fusion further enhanced synchronization quality. Zeng et al. [9] presented Talk Lip Net, which combines a lip-reading expert with contrastive objectives to surpass prior models. Chen et al. [10] extended Wav2Lip with SAM-Wav2Lip++, integrating an action generator for natural gestures alongside lip movements. Kim et al. [11] tackled subtle timing issues through an audio-visual quality alignment network, stabilizing synchronization. At the same time, Xu et al. [12] introduced Diff2Lip, a diffusion-based method that optimized multiple loss functions to produce sharper and more lifelike lip movements. As lip-sync technology matured, researchers began to address broader applications. Zhang et al. [13] combined Wav2Lip with Sim Swap in WAVSYNCSWAP, enabling customizable talking faces for virtual meetings and media production. Wang et al. [14] tackled accessibility with MultiLingualSync, a pipeline fusing machine translation with Wav2Lip to generate globally usable, multilingual content. Similarly, Dong et al. [15] built multimodal digital humans for healthcare and education, while Chen et al. [16] and Lai et al. [17] used Wav2Lip to power personalized e-learning assistants and automated lecture generation, respectively.

Another critical direction has been security and forensics, responding to the misuse potential of lip-sync technology. Zhao et al. [18] introduced a tri-network architecture that integrates lip-sync signals for forgery detection, while Li et al. [19] designed LR-MDF, a lip-reading-based detection framework. Zhang et al. [20] contributed the Div-DF dataset, capturing diverse deepfake manipulations to strengthen detection models. In parallel, Sadeghi et al. [21] showed how lip cues can be harnessed for speech enhancement, boosting intelligibility in noisy environments. These developments ultimately trace back to the foundational Wav2Lip framework by Prajwal et al. [22], which first demonstrated the power of a lip-sync discriminator for robust alignment "in the wild." Subsequent refinements, such as Fei et al.'s CA-Wav2Lip [23], layered attention modules to improve synchronization and perceptual quality, laying the groundwork for the diffusion- and transformer-powered approaches that dominate the current landscape.

III. METHODOLOGY

A. System Architecture

The avatar video generator with lip-sync is structured as a comprehensive, modular system that integrates various components to enable flexible inputs, high-quality avatar generation, and seamless user interaction. At a high level, the system architecture includes a React-based frontend, a multi-service backend, deep learning inference engines, secure authentication mechanisms, and external APIs for advanced language processing. Frontend is developed with React and styled components to give a responsive

and easy-to-use interface. It has a dashboard that aggregates the access to all the most important functionalities, including input selection, avatar customization, and video playback. Users are able to easily choose from an avatar library and interact with different input modalities. Playback controls are embedded to preview the generated videos.

The interface is designed to offer a streamlined user experience, with clearly defined input methods and intuitive controls. The central preview pane allows users to view the generated content, while the navigation bar provides access to profile details and system logout functionality.

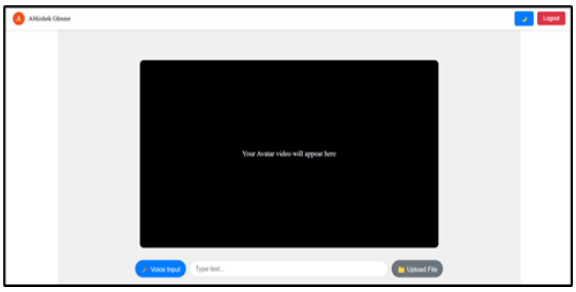

Fig. 1. User interface of the avatar video generator platform showing the main dashboard. The interface includes input options for voice, text, and file upload, along with a central video display area where the generated avatar video is rendered.

The backend is implemented with Node.js for API routing, user authentication, and communication with external services, while Python-based modules are responsible for deep learning model inference and media processing. A WebSocket connection is used to provide real-time status updates during video generation so that users are kept informed throughout the rendering process. Core backend services communicate with deep learning models such as Wav2Lip and variants for lip sync, and other support models for speech-to-text, text-to-speech, and language translation. External APIs like Gemini 1.5-flash are used to provide natural language script generation support, while Groq APIs are utilized for real-time translation in the dubbing process.

B. Input Modalities

The system is capable of supporting multiple input modalities to suit various use cases and user preferences. For example, the users can input voice via their microphone. This real-time audio is processed directly and synchronized with the lip movement of the avatar by the Wav2Lip model. For text input, users can simply input the targeted script in a text field; the system uses text-to-speech synthesis to produce audio, which it passes through the lip-sync module. Audio file upload is also available in MP3 or WAV formats, where pre-recorded material can be lip-synced with the chosen avatar. The video dubbing feature further expands the system's usefulness by allowing cross-language conversion of existing videos. Users can provide a video in one language, and the platform will translate the speech and sync the avatar's lip movements into the target language. Finally, for users that prefer to generate content from scratch, a content generation feature allows them to provide a topic; the platform produces a related script based on the Gemini 1.5-flash model, followed by synthesizing a lip-synced video using the selected avatar delivering the auto-generated content.

C. Video Dubbing Process

The dubbing pipeline for videos is a consecutive process of operations aimed at high precision and realism. First, the system removes the audio from the uploaded video. The audio is then sent to a speech-to-text module, converting it into written text form. The transcript obtained is sent to the Groq translation API along with the user's choice of target language. Upon translation, the new text is rendered into speech with a high-quality text-to-speech engine, which produces sound in the target language. This sound is then aligned with the original video frames with the help of the Wav2Lip model to create natural lip movements that follow the translated speech. The end result is a video produced with the initial graphics preserved and the audio as well as lips substituted to correlate with the changed language.

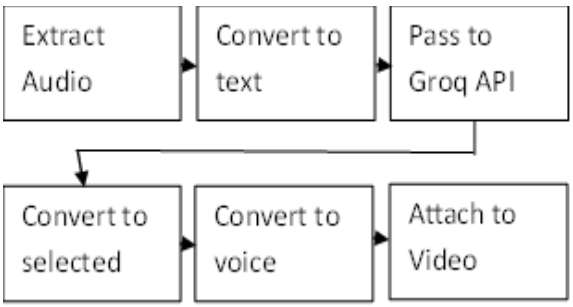

Fig. 2. Workflow of the video dubbing process used in the avatar video generator system.

D. Content Generation

For social media content generation, the system has an end-to-end pipeline that simplifies video creation from topics defined by the user. The pipeline starts when a user types in the topic of interest say, "climate change awareness." The Gemini 1.5-flash language model is called upon to produce a short, contextually relevant script on the topic. After the script is developed, the user chooses a desired avatar, which will be the video's visual speaker. The script is translated into speech through a TTS model, and the speech is then utilized to animate the avatar through the Wav2Lip synchronization module. The output video is not only viewable within the site but is also emailed to the user's subscribed email address, so it can easily be shared on social media or other sites.

E. Authentication and Data Storage

Authentication and data handling are essential parts of the infrastructure of the system. Users securely sign in with their Google accounts, and JWT tokens handle session state and authentication. This makes sure that only permitted users can utilize the features of the system. MongoDB Atlas is used as the backend database where all user data like names, email addresses, profile images, and Google IDs are stored. Furthermore, metadata of created content video titles, creation times, and language preferences are stored and referenced to user accounts in a secure manner as well. This enables effective management and retrieval of content and provides users with access to their customized media library.

IV. RESULTS AND DISCUSSION

A. Lip Synchronization Quality

Lip synchronization in the avatar's mouth was evaluated on the basis of technical analysis and end-user feedback. Smooth correspondence of spoken audio with avatar mouth movements was observed in the system for all input types. Live voice input showed the most accurate synchronization, owing to the fact that the input signal was direct and unprocessed. When text and audio file inputs were employed, the visual timing of the lips was extremely natural, with no noticeable delay that could distract the viewer. Independent reviewers assessed the realism of the produced videos. They universally observed that the lip movements were well-matched with the content spoken, lending an overall sense of authenticity. Reviewers in particular enjoyed the smooth transitions between mouth shapes and the correct speech pacing in the visual space. This feedback supports the system's employment of Wav2Lip and its preprocessing operations, which maintain fidelity irrespective of input mode.

The system also provided stable performance across all means of input spoken words, typed words, and uploaded audio. Such variations in lip-sync quality between these modes were small, indicating that the synchronization engine handles each type consistently. This stability is essential to provide consistent results in actual use cases.

Fig. 3. Visual comparison of avatar mouth movements across different input types.

B. Multilingual Performance

The video dubbing functionality was experimented with in different language translation scenarios to establish the level at which lip synchronization and content integrity are maintained. Translations from the videos were able to capture the intended meaning of the original speech quite easily, and users noted that switching between languages appeared seamless. The system effectively translated videos into a number of languages without losing tone or pace, confirming its use for broadcasting material globally. In terms of synchronization quality, lip movement continued to synchronize with the translated voice, although the target language might have had a totally different phonetic structure than the original. Reviewers were amazed at the stability and credibility of the lip-sync across all language pairs that were tested. Visual precision went down somewhat in cases where patterns of speech departed from the original but not from affecting the overall perception of the viewers.

The system's capacity to maintain realism in video that is translated is particularly beneficial for multilingual instructional content, public outreach, and media localization. The findings validate the robustness and efficiency of the dubbing pipeline, from transcription and translation to speech synthesis and synchronization.

Fig. 4. Interface view of the video dubbing feature within the avatar generator system

C. Content Generation Evaluation

The ability of the system to produce content based on topics provided by users was measured in terms of its utility, applicability, and responsiveness. Reviewers found the automatically created scripts to be coherent and relevant. The tone of the language in the videos was comparable to the expected tone for the given themes, i.e., educational facts or promotional stories. In most instances, the content read smoothly and had proper structure and detail for the given topics.

When posted on social media platforms in a simulated environment, the avatar videos were more likely to attract attention and elicit more engagement than comparable messages shared as static text. Users reacted more favorably towards visual content, suggesting that animated avatars made the message seem more dynamic and engaging.

In terms of usability, the system executed requests efficiently, producing finished videos in a brief time interval. This effectiveness enhances the system's appropriateness for content creators who need quick turnaround, like teachers creating daily lessons or social media influencers posting frequently.

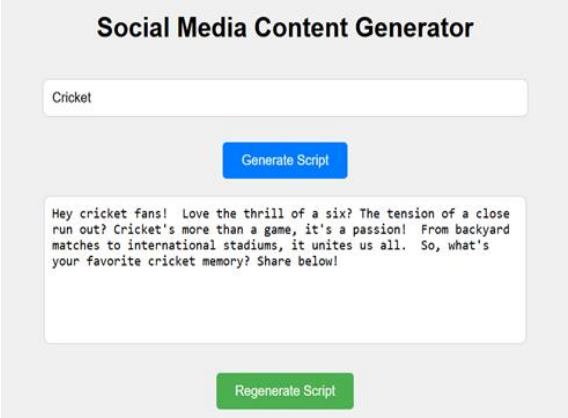

Fig. 5. User interface of the social media content generation module.

D. User Experience Assessment

A formal user study was undertaken to identify how users interact with the system and what value they perceive in it. Participants were requested to experiment with various features such as real-time voice input, dubbing video, and automatic content creation. After interacting with the system, participants provided feedback regarding usability, readability, and general satisfaction.

The interface was unanimously considered intuitive. Users found the simple layout, coherent navigation, and responsive design commendable. Avatar choice and input options were easy to navigate, even for those with lower technical backgrounds. Users easily produced videos with minimal guidance, a testament to the platform's suitability for mass use.

Feature favorites, when asked about preferences, included the ability to dub video as a feature that was incredibly well-liked. Users believed there was tremendous potential for the feature in learning languages, translation of content, and international communication. Voice input was also popular, allowing for fast, natural content creation. The topic-based generation feature was appealing to users who wanted quick, scripted videos without a lot of planning or post-production.

Comments on output quality were generally favorable. Users called the produced videos clear, well-synced, and visually interesting. Some users recommended more dynamic avatar expressions in order to accurately capture the emotional content of the speech. This critique draws attention to aspects where the avatars' expressiveness and realism can be increased in future revisions.

E. Comparison With Existing Models

Work / Model	Core Technique	Input Modalities
Prajwal et al. (Wav2Lip)	GAN with lip-sync discriminator	Audio + Video
Fei et al. (CA-Wav2Lip)	Refined Attention + Coordinate Attention	Audio + Video
Xu et al. (Diff2Lip)	Diffusion-based synchronization	Audio
Kim et al.	Audio-Visual Quality Alignment Network	Audio
Chen et al. (SAM-Wav2Lip++)	Contrastive learning + Action generation	Audio
Zeng et al. (TalkLipNet)	Lip-reading + Contrastive learning	Audio
Zhang et al. (WAVSYNCSWAP)	Integration of Wav2Lip + SimSwap	Audio + Face ID
Proposed System (This Work)	Modular pipeline (Wav2Lip + TTS + Translation + GANs + Transformers + Gemini + Groq APIs)	Voice, Text, Audio Upload, Video Dubbing, Topic Input

Table I. Core Techniques and Input Modalities in Lip-Sync Systems

Work / Model	Multilingual Support	Applications
Prajwal et al. (Wav2Lip)	✗	Generic lip-sync in the wild
Fei et al. (CA-Wav2Lip)	✗	Improved sync realism
Xu et al. (Diff2Lip)	✗	Sharper lip movements
Kim et al.	✗	Stable sync timing
Chen et al. (SAM-Wav2Lip++)	✗	Natural gestures + lips
Zeng et al. (TalkLipNet)	✗	Benchmark surpassing
Zhang et al. (WAVSYNCSWAP)	✗	Customizable talking faces
Wang et al. (MultiLingualSync)	✓	Cross-language synthesis
Dong et al.	✓	Digital humans for healthcare/education
Lai et al.	✓	Automated lectures
Zhao et al.	✗	Deepfake detection
Li et al. (LR-MDF)	✗	Deepfake forensics
Zhang et al. (Div-DF)	✗	Dataset creation
Sadeghi et al.	✗	Speech enhancement
Proposed System	✓ (Real-time multilingual dubbing & generation)	Education, Social Media, Accessibility, Virtual Avatars

Table II. Language Support and Application Focus

Work / Model	System Integration	User Interaction
Prajwal et al. (Wav2Lip)	Single model	No user-facing system
Fei et al. (CA-Wav2Lip)	Model refinement	No interface
Xu et al. (Diff2Lip)	Research prototype	✗
Kim et al.	Model improvement	✗
Chen et al. (SAM-Wav2Lip++)	Model extension	✗
Zeng et al. (TalkLipNet)	Standalone model	✗
Zhang et al. (WAVSYNCSWAP)	End-to-end pipeline	Limited UI
Wang et al. (MultiLingualSync)	Multi-model integration	Limited interaction
Dong et al.	Combined approach	✗
Chen et al.	Limited system	Basic interaction
Lai et al.	Prototype	✗
Zhao et al.	Forensic focus	✗
Li et al. (LR-MDF)	Standalone	✗
Zhang et al. (Div-DF)	Dataset resource	✗
Sadeghi et al.	Model integration	✗
Proposed System	Full-stack system: React frontend, Node.js backend, Python ML services, MongoDB Atlas, JWT auth	✓ Rich dashboard: avatar selection, input controls, preview pane, content management

Table III. System Integration and User Interaction

V. FUTURE SCOPE

Leveraging the strengths of the current system and respecting its shortcomings, the future will involve efforts toward making avatar creation more resilient and expressive. A potential area of focus is increasing performance under demanding visual conditions such as low light, severe head poses, or facial occlusion, which may cause output quality deterioration. In this direction, advanced facial tracking and face-frontalization techniques may be incorporated. Simultaneously, enabling transfer of affective signals tone, facial expression, and body language will make avatars appear more realistic and interesting. This will imply conditioning lip-sync generation on phonemes as well as affective embeddings that have been derived from the audio or text. Additionally, real-time interaction capabilities are an obvious fit, allowing the system to facilitate features like live video calls, interactive digital assistants, or streaming avatars that respond dynamically to users in real time.

Another key direction is taking personalization and accessibility further. Personalized avatars that reflect a user's appearance or style would add to immersion and user satisfaction. This can be done with light-weight face-swapping or avatar training models off of user-uploaded material. Language extension is also important enabling more language and dialect support, especially less-represented ones, would go a great way towards boosting the system's global penetration. Alongside these functionalities, security functionality in the form of visible watermarking, content attribution, and AI-created content detection features will play a key role in preventing misuses, particularly deepfakes. Finally, mobile integration is a priority in order to enable creation and sharing of content through tablets and smartphones, further democratizing access to cutting-edge avatar video generation tools.

VI. CONCLUSION

The system introduced a holistic avatar video generator with lip-syncing, bringing together various deep learning models to generate lifelike digital humans with synchronized speech. Our system accommodates various input modalities, including voice input, text input, uploading audio files, and dubbing video, and also provides content generation for social media platforms.

The experimental results show the effectiveness of the system in preserving high-quality lip synchronization with various input types and languages. The easy-to-use interface, along with JWT-based authentication and MongoDB Atlas support, allows sophisticated avatar generation technology to be utilized by users with no advanced technical expertise.

As digital human generation becomes more sophisticated, systems such as ours will become ever more central to virtual communication, content creation, learning, and entertainment. By solving the issues of lip synchronization, multilingual support, and content generation, our research advances realistic and accessible digital human representations.

REFERENCES

[1] A. Chatziagapi, N. Koutlis, A. Katsamanis, P. Koutras, and P. Maragos, “AV-Flow: Transforming Text to Audio-Visual Human-like Interactions,” arXiv preprint arXiv:2502.13133, Feb. 2025.

[2] A. R. Oskooei, M. S. Aktaş, and M. Keleş, “Seeing the Sound: Multilingual Lip Sync for Real-Time Face-to-Face Translation,” Computers, vol. 14, no. 1, Art. no. 7, 2025.

[3] Anonymous, “A Real-Time End-to-End Framework for Audio-Driven Avatar Synthesis Using NeRF,” arXiv preprint arXiv:2501.14646, Jan. 2025.

[4] X. Wu, Y. Wang, S. Wu, and H. Li, “Speech2Lip: High-Fidelity Speech to Lip Generation by Learning from a Short Video,” arXiv preprint arXiv:2309.04814, Sep. 2023.

[5] Y. Zhang, F. Sun, X. Chen, and M. Zhou, “MuseTalk: Real-Time High Quality Lip Synchronization with Latent Space Inpainting,” arXiv preprint arXiv:2410.10122, Oct. 2024.

[6] C. Li, J. Yang, Q. Xu, and Y. Li, “LatentSync: Audio Conditioned Latent Diffusion Models for Lip Sync,” arXiv preprint arXiv:2412.09262, Dec. 2024.

[7] Y. Lin, J. Hu, Y. Xu, and J. Zhu, “GLDiTalker: Speech-Driven 3D Facial Animation with Graph Latent Diffusion Transformer,” arXiv preprint arXiv:2408.01826, Aug. 2024.

[8] S. J. Park, K. H. Lee, and J. Y. Choi, “Talking Face Generation with Precise Lip-Syncing via

Audio-Lip Memory," in Proc. AAAI Conf. Artificial Intelligence, 2022, pp. 2114–2122.

[9] K. Zeng, F. Zhao, L. Wu, and H. Zhang, "TalkLipNet: Contrastive Learning Powered Lip-to-Speech Synchronization with a Lip-Reading Expert," in Proc. IEEE ICCV, 2023, pp. 828–837.

[10] X. Chen, J. Zhou, W. Liu, Y. Zhang, and H. Li, "SAM-Wav2Lip++: Speech-Driven Action Generation and Lip Synchronization with Contrastive Learning," ACM Trans. Multimedia Comput. Commun. Appl., vol. 19, no. 3, pp. 1–16, 2023.

[11] H. Kim, J. Park, and S. Lee, "Enhanced Wav2Lip with Audio-Visual Quality Alignment Network," in Proc. IEEE ICIP, 2022, pp. 1893–1897.

[12] J. Xu, Y. Zeng, and G. Mori, "Diff2Lip: Audio-Driven Lip Synchronization Using Diffusion Models," IEEE Trans. Multimedia, vol. 24, no. 9, pp. 2523–2534, Sep. 2022.

[13] P. Zhang, L. Chen, and K. Wang, "WAVSYNCSWAP: End-to-End Integration of Wav2Lip and SimSwap for Talking Face Synthesis," in Proc. IEEE ICME, 2023, pp. 1145–1150.

[14] L. Wang, J. Zhang, and Y. Liu, "MultiLingualSync: Combining Wav2Lip with Translation Technologies for Cross-Language Talking Head Synthesis," in Proc. IEEE ICASSP, 2023, pp. 4108–4112.

[15] Z. Dong, Y. Chen, and K. Li, "Designing Virtual Human with Deep Learning Models: A Combined Approach," Multimedia Tools Appl., vol. 79, pp. 34345–34372, 2020.

[16] Y. Chen, Z. Liu, and H. Wang, "Interactive Video Virtual Assistant for E-Learning: A Framework Using Wav2Lip," J. Educ. Technol. Soc., vol. 23, no. 4, pp. 1–14, 2020.

[17] C. Lai, H. Zheng, and Y. Tang, "Automated Lecture Generation Using Text-to-Speech and Wav2Lip GAN," IEEE Trans. Learn. Technol., vol. 14, no. 4, pp. 394–405, Oct.–Dec. 2021.

[18] Y. Zhao, T. Chen, and L. Li, "Tri-Network Architecture for Multimodal Forgery Detection Integrating Lip-Sync Signals," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 1102–1115, 2023.

[19] S. Li, X. Zhao, and R. Wang, "LR-MDF: A Lip Reading Based Multimodal Deepfake Detection Framework," IEEE Trans. Inf. Forensics Security, vol. 17, pp. 4168–4180, 2022.

[20] R. Zhang, Y. Liu, J. Gu, and Y. Wang, "Div-DF: A Comprehensive Dataset for Deepfake Detection with Diverse Lip-sync Manipulations," in Proc. IEEE CVPR Workshops, 2022, pp. 123–130.

[21] M. Sadeghi, S. Leglaive, and X. Alameda-Pineda, "Speech Enhancement using Perceptual Loss Based Wav2Lip Integration," in Proc. IEEE ICASSP, 2022, pp. 8437–8441.

[22] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild," in Proc. ACM Multimedia, 2020, pp. 484–492.

[23] X. Fei, Y. Deng, Y. Wang, S. Xu, and X. Zhao, "CA-Wav2Lip: Improving Lip Synchronization by Refined Attention and Coordinate Attention," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 11, pp. 7651–7665, Nov. 2022.