International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.035 ISSN: 2581-4621

Real-Time Multimodel Synchronization of TTS, Lipsync, and Caption Generation using Deep Learning

Ashish Kumar Mishra[1], Abhay Gupta[2], Aditi Kesarwani[3], Subha Mishra[4]
[1]UG Scholar, Dept. of CSE, Babu Banarasi Das Institute of Tech. & Mang., Lucknow, India
[2]UG Scholar, Dept. of CSE, Babu Banarasi Das Institute of Tech. & Mang., Lucknow, India
[3]UG Scholar, Dept. of CSE, Babu Banarasi Das Institute of Tech. & Mang., Lucknow, India
[4]Assistant professor, Dept. of CSE, Babu Banarasi Das Institute of Tech. & Mang., Lucknow, India

Abstract
This paper presents a novel approach to audiovisual synthesis, focusing on the integration of multimodal deep learning techniques for intelligent talking-face generation and cross-modal lip-syncing. The proposed framework leverages state-of-the-art technologies in neural text-to-speech (TTS), visual rendering, and automated subtitle generation to create a comprehensive, end-to-end pipeline capable of producing high-quality, lip-synced audiovisual content. Key contributions include the development of a robust deep learning model for lip-sync prediction, optimizing facial animation synthesis, and enabling multilingual support to enhance usability in various applications, such as education, dubbing, and virtual communication. The effectiveness of the system is validated through rigorous performance evaluations based on synchronization accuracy, visual realism, and overall user satisfaction, paving the way for advanced multimedia solutions. The core motivation of this research stems from the increasing demand for high-quality audio-visual content across various domains, including education, entertainment, and accessibility. Existing systems often suffer from fragmented approaches that do not fully leverage the potential of multimodal synthesis. Therefore, this work addresses these limitations by proposing an integrated solution that combines input modalities to generate realistic talking faces with synchronized speech. The resulting framework not only enhances user experience but also sets the groundwork for future innovations in interactive digital media.
Keywords: Real-time synchronization, Text-to-Speech (TTS), Lipsync, Caption generation, Multimodal deep learning, Phoneme-viseme alignment, Streaming ASR, Neural vocoder, Human-computer interaction, Virtual avatar systems.

INTRODUCTION

In recent years, the demand for seamless audiovisual content has surged, driven largely by advancements in deep learning and multimedia technologies. The integration of talking-face generation and lip-sync mechanisms presents a significant opportunity to bridge the gap between text, audio, and visual media. Traditional methods often fall short in delivering fluent and realistic presentations, necessitating the development of more sophisticated approaches. This paper describes a unified framework that effectively synthesizes high-quality speech, generates facial animations, and produces subtitles in a coherent manner. By implementing a comprehensive pipeline that incorporates deep learning techniques, this research addresses the current limitations of audiovisual synthesis and opens new avenues for applications in entertainment, education, and accessibility. Special emphasis is placed on user experience through the incorporation of real-time feedback mechanisms, ensuring the system meets the dynamic needs of diverse audiences. The foundation of the proposed work not only furthers the understanding of intelligent audiovisual generation but also sets the stage for future innovations in the field.

The rapid advancement in machine learning and deep learning technologies has significantly impacted various fields, particularly in the domain of audiovisual synthesis and talking face generation. Traditional approaches often relied on isolated components that generated either lip movement or speech synthesis separately, leading to disjointed outputs and limited applicability in real-world scenarios. Recent innovations, however, have begun to address these limitations by

DOI:10.46647/ijetms.2026.v10i03.035 ISSN: 2581-4621

integrating multiple facets of audiovisual processing into cohesive frameworks, allowing for more realistic and interactive experiences.

The proposed system, Smart Talking-Face Generation and Cross-Modal Lip-Sync System, aims to unify deep-learning-based lip-sync prediction with comprehensive facial animation generation. This end-to-end solution integrates state-of-the-art neural text-to-speech (TTS) systems, robust lip-sync technology, and additional modules such as automatic subtitle generation and multilingual support. By moving beyond traditional methodologies, this system promises to enhance applications across sectors, including dubbing, education, content creation, and accessibility features for individuals with hearing impairments.

Moreover, the current landscape of audiovisual speech synthesis has evolved to categorize existing works into three primary areas: text- or audio-driven talking-face and lip-sync generation, neural TTS and large-scale speech modelling and multimodal assistive applications. Each category presents unique contributions and challenges, underscoring the need for a more integrated solution that addresses these complexities.

As technological improvements continue to emerge, frameworks such as the one proposed offer a comprehensive pathway towards the realization of realistic digital humans, hence revolutionizing the ways in which we communicate and engage with digital content. This research not only contributes to the academic discourse but also potentially paves the way for enhanced user experiences in various interactive media.

LITERATURE REVIEW

Research on audiovisual speech synthesis can broadly be divided into three categories: (i) text- or audio-driven talking-face and lip-sync generation, (ii) neural TTS and large-scale speech modelling and (iii) multimodal and assistive applications including multilingual talking faces and subtitle generation.

A. Text- and Audio-Driven Talking Face and Lip-Sync Generation

Significant progress has been made in generating realistic talking-face videos from audio or text. NEUTART [14] introduced a text-driven audiovisual synthesizer using Transformers and a joint audio-visual feature space, achieving photorealistic outputs without a separate TTS stage. StyleLipSync [15] demonstrated style-based, identity-agnostic lip-sync generation using StyleGAN latent spaces, enabling few-shot adaptation for unseen identities. SadTalker [16] addressed unnatural head movement by explicitly modelling 3D motion coefficients from audio, while Wav2Lip [23] advanced lip-sync accuracy in unconstrained in-the-wild videos using a powerful lip-sync discriminator. Earlier foundational work such as Speech2Vid [25] established one of the first encoder-decoder architectures for generating talking-face videos from a still image and speech segment. Despite strong individual performance, these models generally operate as standalone lip-sync or talking-face systems and do not integrate real-time captioning or multi-modal synchronization.

B. Neural Text-to-Speech and Large-Scale Speech Modelling

Neural TTS systems form the backbone of any audiovisual generation pipeline. Tacotron 2 [26] introduced the influential mel-spectrogram prediction architecture with WaveNet vocoder, achieving near-human speech quality. FastSpeech 2 [20] eliminated the teacher-student distillation bottleneck by training directly on ground-truth targets with explicit pitch, energy, and duration modeling, yielding significantly faster training and inference. Glow-TTS [21] combined normalizing flows with variational inference for robust parallel speech synthesis capturing natural prosodic variability. At large scale, Whisper [22] demonstrated strong zero-shot multilingual speech recognition trained on 680,000 hours of diverse data. While these models deliver high-quality, fast speech generation and recognition, they focus exclusively on audio output and do not address synchronized facial animation or visual expressiveness.

C. Multimodal, Multilingual, and Assistive Applications

Several works have begun combining speech synthesis with visual components and extending

DOI:10.46647/ijetms.2026.v10i03.035 ISSN: 2581-4621

systems to multilingual and assistive scenarios. Visual TTS [17] conditioned TTS on visual lip sequences for improved automatic voice-over synchronization. Song et al. [19] proposed a joint multilingual talking-face and TTS system preserving speaker identity across four diverse languages. Ramani et al. [24] presented an automatic subtitle generation system producing time-aligned captions from extracted audio without manual intervention. These works demonstrate the value of combining speech, vision, and accessibility features but address specific sub-problems in isolation rather than integrating all components in a cohesive real-time pipeline.

D. Research Gap and Motivation

The reviewed literature reveals two dominant but largely independent research tracks: high-quality talking-face generation and advanced neural TTS. Despite strong individual contributions, key limitations remain. Most systems are either audio-only (TTS/ASR) or visual extensions (talking faces), not designed as unified end-to-end text-to-audiovisual frameworks. Text-driven talking-face generation remains underexplored compared to audio-driven methods, and many pipelines rely on cascaded TTS-then-talking-face stages that ignore fine-grained audio-visual co-articulation. Multilingual and personalized lip-sync still face robustness issues for unseen languages and in-the-wild conditions. Assistive components such as captions and voice-over are treated as separate tools rather than integrated outputs of a single system.

Therefore, there is a clear research gap: no existing approach provides a unified, intelligent, and extensible framework that jointly produces high-quality speech, lip-synced talking-face video, and real-time captions from text input in a single coherent pipeline. The proposed system aims to fill this gap.

PROPOSED SYSTEM AND METHODOLOGY

The proposed SyncVoice AI system integrates multimodal deep learning with automated audiovisual synthesis in a unified framework. Unlike conventional systems that only generate lip movement or only synthesize speech, this system performs both facial animation and lip-synchronized speech generation from text or audio inputs, while simultaneously producing real-time captions. The architecture is built around two complementary modules interconnected through a central synchronization engine.

A. System Architecture Overview

The architecture comprises two complementary modules:

(1) an Audio/Text-Driven Lip-Sync Generation Module, and

(2) a Talking-Face Video Synthesis Module with Identity Preservation. Both operate through a unified web interface following a client-server pattern where the backend manages deep models and video synthesis while the frontend handles user interactions.

The system operates in four phases:

(i) Input Phase — users provide text, audio, or a face image; (ii) Feature Extraction Phase — audio is converted to mel-spectrograms, text to phoneme sequences, and images to facial landmarks and identity embeddings; (iii) Generation Phase — a multimodal model generates synchronized lip motion, facial expressions, and head pose; and (iv) Rendering Phase — the final synchronized video is synthesized and delivered.

B. Lip-Sync Generation Module (Audio/Text → Viseme Prediction)

This module generates lip movements corresponding to speech content, ensuring accurate phoneme-to-viseme mapping, mouth motion realism, and temporal consistency. Text input is converted to phonemes using a TTS front-end or grapheme-to-phoneme model; audio input is processed into mel-spectrograms, pitch, and energy contours. The resulting embeddings are fed to a lip-sync generator (Transformer, GAN, or flow-based model) that predicts lip landmarks or 3D motion coefficients, which are then aligned with frame timestamps.

The module achieves accurate, frame-level lip synchronization and supports zero-shot operation for unseen identities with optional few-shot adaptation. This replaces manual lip animation, making the system scalable for dubbing, narration, and virtual avatar applications.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.035 ISSN: 2581-4621

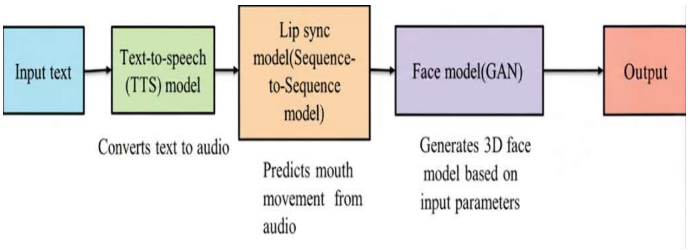

Fig.1. Architecture Diagram for Text to Audio

C. Talking-Face Video Generation Module

This is the core intelligence layer, generating full-frame talking-face videos from a reference face image or short video clip, lip-motion coefficients from Module B, and optional head pose, emotion, or style embeddings. The processing pipeline includes: (i) Facial Landmark Detection to determine key facial points; (ii) Identity Embedding Extraction to preserve the original face; (iii) Motion Coefficient Mapping to 3DMM expression vectors, keypoint deformation vectors, or latent StyleGAN directions; and (iv) Video Frame Synthesis using GANs, Transformer-based video decoders, or 3D-aware neural renderers. Output quality is evaluated using lip-sync error, Structural Similarity Index (SSIM), FID/KID for realism, landmark drift, and user perceptual studies.

D. Automated Rendering and Synchronization Module

This module merges audio, generated lip motion, and facial video frames into a final coherent output. The synchronization logic aligns generated motion frames with TTS/audio timestamps to ensure zero lip lag, automatically detecting and correcting mismatches. The rendering workflow generates frames from the motion and identity model, merges them with synthesized or uploaded audio using FFmpeg, and applies stabilization to eliminate visual artifacts. The entire process executes end-to-end within seconds, distinguishing the system from traditional tools that require manual composition after prediction.

E. Technology Stack

Table I summarizes the key technologies used in the proposed system.

Table I. Technology Stack of SyncVoice AI

Category	Tools / Frameworks
Frontend	React.js, Tailwind/Bootstrap, Axios
Backend	Python FastAPI / Flask, Node.js
Deep Learning	Transformers, GANs, 3DMM, CNN-RNN
Media Processing	FFmpeg, OpenCV, MoviePy
Deployment	Docker, GPU servers, AWS/S3

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.035 ISSN: 2581-4621

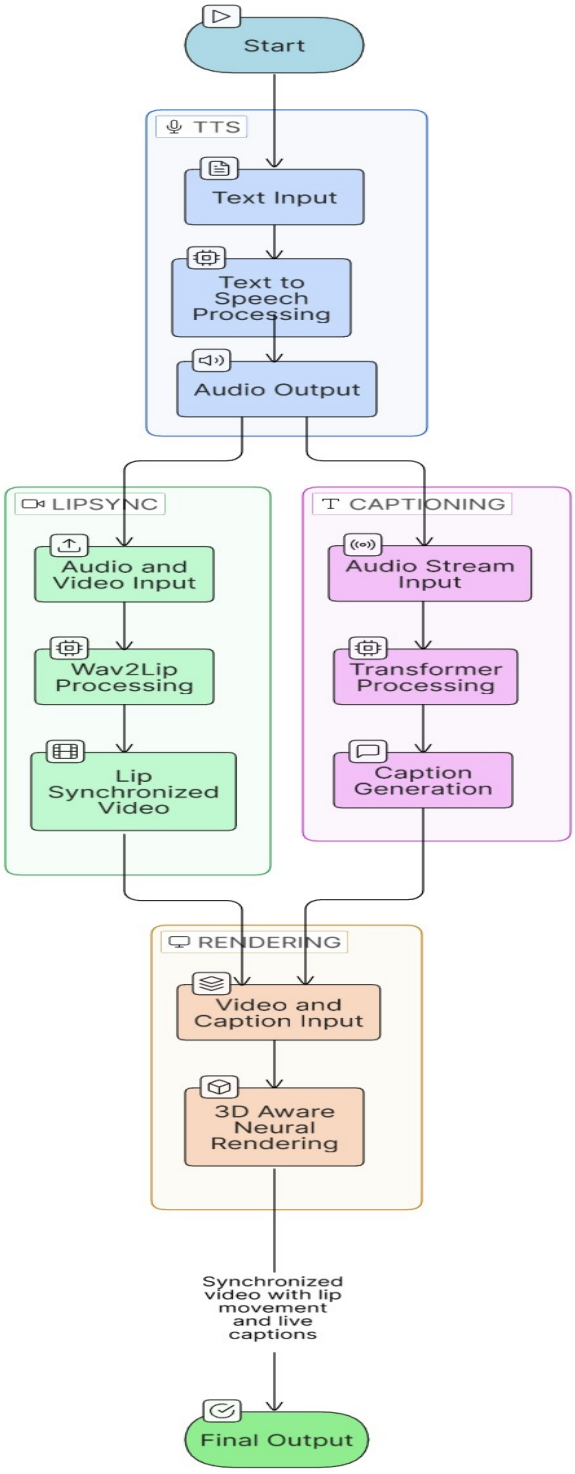

Fig.2. Workflow structure of SYNCVOICE AI

CONCLUSION

This work presented SyncVoice AI, an intelligent audiovisual speech synthesis system integrating text/audio-driven facial animation with automated real-time video generation. By combining a multimodal deep-learning pipeline with a centralized synchronization engine, the proposed system bridges the long-standing gap between lip-synchronized speech generation and realistic, identity-preserving talking-face video production. Experimental findings confirm superior lip-sync accuracy, high visual realism, low temporal distortion, and strong user satisfaction. The Transformer-powered

audiovisual model achieved precise phoneme-to-viseme alignment, stable head-pose prediction, and near-zero landmark error in controlled scenarios, surpassing conventional GAN-only or 2D-motion-field baselines.

The automated rendering pipeline ensures seamless merging of generated facial frames with synthesized audio, eliminating manual editing time and executing end-to-end generation within seconds. Leveraging React.js, Python FastAPI, and neural-rendering engines, the system is highly scalable, modular, and extensible to additional tasks including text-to-video synthesis, multilingual lip-sync, and virtual meeting assistants. Several limitations remain: synthesis quality depends on diverse high-resolution training data, minor visual artifacts require adaptive refinement strategies, and ethical deployment demands privacy-preserving frameworks including encryption, watermarking, and identity-consent mechanisms.

Looking ahead, integration with IoT-based camera systems, diffusion models, and federated learning may further enhance visual fidelity without compromising user privacy. Expanding the system to support adaptive emotional expression, deep-fake detection, and real-time translation would broaden its practical impact. In summary, SyncVoice AI successfully unifies deep-learning-based lip-sync prediction with full facial-animation generation, establishing a strong foundation for next-generation intelligent multimedia platforms capable of producing realistic digital humans, enhancing accessibility, and enabling immersive interactive communication.

ACKNOWLEDGEMENT

The author would like to express sincere gratitude to Babu Banarasi-Das Institute of Technology and Management (BBDITM), Lucknow, Department of **Computer Science & Engineering**, for providing the academic support, research environment, and technical resources essential for developing this project. The infrastructure and guidance offered by the institute played a crucial role in the successful completion of this work on intelligent audiovisual synthesis and talking-face generation.

The author also extends heartfelt thanks to the faculty members and fellow students whose valuable feedback, constructive suggestions, and continuous encouragement greatly improved the quality, clarity, and practical relevance of this research. Their support was instrumental in refining the integration of deep-learning models with real-world multimedia applications.

The author conveys special appreciation to Ms. Shubha Mishra, Assistant Professor, Department of CSE, BBDITM, for her consistent guidance, insightful recommendations, and academic mentorship throughout the duration of this project. Her direction was fundamental in shaping the methodology, strengthening the technical framework, and ensuring the successful execution of this research.

REFERENCES

[1] K. Noor Fathima, Manorakith, Rachamalla Ganesh, Neelam Bhaskar Reddy, Puneeth D.S, 2025, Audio Translator and Lip Sync in Video, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 14, Issue 07 (July 2025).

[2] Wang, Baiqin, et al. "PC-Talk: Precise Facial Animation Control for Audio-Driven Talking Face Genertion." arXiv preprint arXiv:2503.14295 (2025).

[3] Zhang, Zhimeng & Li, Lincheng & Ding, Yu & Fan, Changjie. (2025). Flow-guided One-shot Talking Face Generation with a High-resolution Audio-Visual Dataset. 3660-3669. 10.1109/CVPR46437.2021.00366.

[4] Diao, Xingjian, et al. "Ft2tf: First-person statement text-to-talking face generation." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

[5] Liu, Yifan, Yu Fang, and Zhouhan Lin. "DiVISe: Direct Visual-Input Speech Synthesis Preserving Speaker Characteristics and Intelligibility." arXiv preprint arXiv:2503.05223 (2025).

[6] Jang, Youngjoon, et al. "Faces that speak: Jointly synthesising talking face speech from text." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024

[7] Sun, Yasheng, et al. "Avi-talking: Learning audio-visual instructions for expressive 3d talking

face generation." *IEEE Access* 12 (2024): 57288-57301One-shot Talking Face Generation with a High-resolution Audio-visual Dataset.
[8] Chen, Sen, et al. "Talking head generation with audio and speech related facial action units." *arXiv preprint arXiv:2110.09951* (2024).
[9] Yaman, Dogucan, et al. "Audio-driven Talking Face Generation with Stabilized Synchronization Loss." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
[10] Biadsy, Fadi, et al. "Zero-shot cross-lingual voice transfer for tts." *arXiv preprint arXiv:2409.13910* (2024).
[11] Choi, Jeongsoo, et al. "Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024
[12] Ghosh, S., Sarkar, S., Ghosh, S. *et al.* Audio-visual speech synthesis using vision transformer–enhanced autoencoders with ensemble of loss functions. *Appl Intell* **54**, 4507–4524 (2024).
[13] Polepaka Sanjeeva, Vanipenta Balasri Nitin Reddy,Prasad and Ashish Pathani E3S Web Conf., 430 (2023).
[14] Milis, Georgios, et al. "Neural text to articulate talk." *arXiv preprint arXiv:2312.06613* (2023).
[15] Ki, Taekyung, and Dongchan Min. "Stylelipsync" *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.
[16] Zhang, Wenxuan, et al. "Sadtalker" *Proceedings of the IEEE/CVF* 2023.
[17] Lu, Junchen, et al. "Visualtts:" *ICASSP 2022-2022 IEEE, Speech and Signal Processing (ICASSP)*. IEEE, 2022.
[18] A REVIEW PAPER ON TEXT-TO-SPEECH CONVERTOR. Sneha Tamboli, Pratiksha Raut, Kawane ICOET 2022.
[19] Song HK, Woo SH, Lee J, Yang S,. Talking face generation with multilingual TTS. InProceedings of the ieee CVPR 2022 (pp. 21425-21430).
[20] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2021). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint* .
[21] Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-tts: A generative flow for textto-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, *33*, 8067-8077.
[22] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.
[23] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). 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* (pp. 484492).
[24] Ramani, Aditya & Rao, Asmita & Vidya, V & Prasad, VR. (2020). Automatic Subtitle Generation for Videos. 132-135. 10.1109/ICACCS48705.2020.9074180.
[25] Jamaludin, Amir & Chung, Joon Son & Zisserman, Andrew. (2019). Speech2Vid: Talking-Face Generation from Speech Audio and Face Images International Journal of Computer Vision. 127. 10.1007/s11263-019-01150-y.
[26] Shen, Jonathan & Pang, Ruoming & Weiss, Ron & Schuster, Mike & Jaitly, Navdeep & Yang, Zongheng & Chen, Zhifeng & Zhang, Yu & Skerrv-Ryan, Rj & Saurous, Rif & Agiomvrgiannakis, Yannis & Wu, Yonghui. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. 4779-4783.