Real-Time Multi Model Synchronization of TTS, Lip Sync, and Caption Generation using Deep Learning

Abhay Gupta
Dept. Computer Science & Engineering
Babu Banarasi Das Institute of
Technology & Management (Dr A P J
Abdul Kalam Technical University)
Lucknow, India abhaymidas099@gmail.com

Aditi Kesarwani
Dept. Computer Science & Engineering
Babu Banarasi Das Institute of
Technology & Management (Dr A P J
Abdul Kalam Technical University)
Lucknow, India aditikesarwaniak@gmail.com

Ashish Kumar Mishra
Dept. Computer Science & Engineering
Babu Banarasi Das Institute of
Technology & Management (Dr A P J
Abdul Kalam Technical University)
Lucknow, India
workwith.ak0@gmail.com

Guided By: **Shubha Mishra**
Assistant Professor
Dept. Computer Science & Engineering
Babu Banarasi Das Institute of
Technology & Management (Dr A P J
Abdul Kalam Technical University)
Lucknow, India iamshubha@bbdnitm.ac.in

Abstract: Real-time multimodal communication systems require seamless synchronization between speech generation, lip movements, and textual captions to create natural, accessible, and interactive digital experiences. This work proposes a unified deep-learning–based framework for **real-time multimodal synchronization of Text-to-Speech (TTS), lipsync animation, and caption generation**. The system integrates a streaming neural TTS model, an audio/text-driven lipsync module, and a low-latency caption generator built on streaming ASR. A central synchronization engine aligns phoneme timestamps, viseme transitions, and caption token timings using adaptive buffering and drift-correction strategies. This ensures that all three modalities—audio, visual articulation, and text output—remain synchronized within perceptually acceptable thresholds (<40 ms). The proposed pipeline improves temporal coherence, reduces caption lag, and enhances user experience in applications such as virtual presenters, digital avatars, assistive technologies, and human–AI communication. Experimental evaluation demonstrates significant improvements in alignment accuracy and latency over baseline independent systems. The framework sets a scalable foundation for future advancements in expressive avatars, multilingual communication, and low-resource real-time deployment.

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.

I. INTRODUCTION

Research on talking-face generation and audiovisual synthesis has progressed rapidly in recent years. Early work such as [1] introduced a real-time audio translation and lip-sync system capable of producing visually coherent and temporally accurate facial movements, showing strong promise for practical, real-world use. Building on the need for greater control, [2] presented PC-Talk, a framework that lets users precisely manipulate facial expressions and speaking styles during audio-driven generation, making the process far more customizable. Meanwhile, [3] approached the problem from a one-shot perspective, using optical-flow guidance to maintain smooth temporal transitions across frames. Moving into text-driven synthesis, FT2TF [4] proposed a multimodal architecture that fuses visual and textual information using multi-scale cross-attention, enabling natural first-person talking-face generation directly from written input.

Beyond lip-sync, several studies explored entirely new directions. For example, [5] demonstrated that speech can be reconstructed directly from silent video, enabling communication in environments where audio is unavailable. A different line of work in [6] integrated both talking-face generation and text-to-speech (TTS), creating a unified pipeline that produces synchronized speech and facial animation from text alone. To push expressiveness further, AVI-Talking [7] leveraged large language models to extract detailed audio-visual instructions, unlocking richer and more emotional 3D talking-face outputs. Fine-grained control of facial behavior was also addressed in [8], where action-unit (AU)-guided landmark prediction allowed precise manipulation of facial expressions in response to audio cues.

Achieving natural synchronization remains a key challenge, and [9] contributed improvements by stabilizing synchronization loss and enhancing visual realism. Personalization became a major focus in [10], which introduced zero-shot cross-lingual voice transfer, allowing users to generate customized voices even with minimal data. At the multimodal translation level, AV2AV [11] proposed a unified representation for converting audio-visual speech from one language to another, preserving both visual and auditory characteristics. Meanwhile, [12] integrated vision transformers with an ensemble of loss functions to increase linguistic accuracy and visual consistency in audio-visual speech synthesis.

Automation of media creation was the target of TEXT2AV [13], which built a complete platform for converting text into synchronized audio–video content. Other works pushed deeper into audiovisual modeling—such as [14], which combined transformers with speech-conditioned 3D facial reconstruction. StyleLipSync [15] introduced style-driven, identity-agnostic lip-sync generation, allowing customizable visual outputs, while SadTalker [16] achieved highly realistic animation from a single image by learning detailed 3D motion coefficients.

Advancements in TTS also played a major supporting role. VisualTTS [17] utilized visual input to improve lip-speech synchronization in automatic voice-over tasks. A broader review in [18] discussed the foundational NLP and DSP components that power modern TTS systems. Multilingual capabilities were expanded in [19], which combined TTS with talking-face generation to support multiple languages. Speed and efficiency were significantly improved through FastSpeech 2 [20] and Glow-TTS [21], two influential architectures known for fast inference and high-quality speech output.

More recently, foundation models began influencing multimodal generation. CLIP [22] showed how large-scale contrastive learning across images and text can strengthen general audiovisual understanding and serve as a powerful backbone for downstream tasks. Lip-sync quality was further refined by [23], which trained a GAN on diverse, in-the-wild datasets to capture natural mouth movements. Some studies broadened the usability scope—such as [24], which automated subtitle generation to reduce manual effort. Earlier foundational work like Speech2Vid [25] established one of the first encoder-decoder systems for mapping audio and static face images to talking-face videos. Finally, [26] introduced a natural TTS framework that predicts Mel-spectrograms with attention-based alignment before feeding them into Wavenet, greatly improving speech clarity and naturalness.

The proposed model processes several multimodal inputs—including the speaker's facial features, audio characteristics, phoneme sequences, text embeddings, and visual landmarks—to accurately predict synchronized lip movements and expressive facial animation. Key variables such as mouth shape dynamics, viseme–phoneme correspondence, emotion cues, head pose, and temporal motion patterns are extracted using deep neural networks to ensure natural audiovisual alignment. When the system identifies a mismatch between predicted lip movements and the incoming audio or text (exceeding a predefined threshold of 70% confidence), the framework automatically triggers corrective generation, refining facial motion to match speech content without requiring manual adjustment. This automated loop enables seamless, real-time synchronization and significantly reduces artifacts such as lag, jitter, or unnatural mouth shapes often present in traditional talking-face systems.

The strength of this architecture lies in its integration of predictive intelligence with automated refinement, allowing the model to produce stable, lifelike talking-face videos with minimal human intervention. Furthermore, the system is easily scalable and can be extended to additional tasks such as text-to-video synthesis, personalized avatar generation, multilingual lip-sync, virtual meeting assistants, and synthetic media production.

The major contributions of this work are:

- Developing a robust multimodal lip-sync prediction model that uses deep learning techniques (e.g., transformers, vision encoders, and audio-visual fusion networks) to generate highly accurate talking-face animations.
- Integrating this predictive framework into a complete pipeline that combines text/audio processing, facial animation synthesis, and visual rendering using technologies such as Python, deep learning libraries, and modern frontend frameworks.
- Conducting a detailed performance evaluation based on synchronization accuracy, visual realism, latency, and user satisfaction to validate the system's effectiveness.

II. LITERATURE REVIEW

Over the past decade, there has been a rapid increase in research on audiovisual speech synthesis, talking-face generation, and neural text-to-speech (TTS). The existing works can broadly be divided into three main categories: (i) text- or audio-driven talking-face and lip-sync generation, (ii) neural TTS and large-scale speech modeling, and (iii) multimodal and assistive applications such as multilingual talking faces, automatic voice-over, and subtitle generation. This section reviews the most relevant work in these areas, highlights their contributions, and discusses their limitations, thereby motivating the need for the proposed system..

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

Several studies focus on generating realistic talking-face videos from audio or text inputs with accurate lip synchronization and natural motion. NEUTART [14] departs from conventional cascaded pipelines by proposing a **text-driven audiovisual speech synthesizer** that directly uses Transformers and a joint audiovisual feature space instead of a separate text-to-speech plus talking-face pipeline. It employs speech-informed 3D facial reconstruction and a lip-reading loss, achieving photorealistic talking faces with well-aligned audio and video on both controlled and in-the-wild datasets.

StyleLipSync [15] introduces a **style-based personalized lip-sync model** capable of generating identity-agnostic lip-synced videos from arbitrary audio. By leveraging the latent space of a pre-trained StyleGAN and using a pose-aware masking mechanism with a 3D parametric mesh predictor, it significantly improves temporal consistency and visual naturalness. A few-shot adaptation strategy further allows the system to personalize lip-sync for unseen identities using only a few seconds of target video.

SadTalker [16] addresses common issues in talking-head generation, such as unnatural head movement and identity distortion, by explicitly modeling **3D motion coefficients** (head pose and expression) of a 3DMM from audio. The method introduces ExpNet for expression learning and PoseVAE for diverse head pose synthesis, followed by a 3D-aware face renderer that maps motion coefficients to keypoints and generates realistic talking-head videos.

Earlier foundational work such as Speech2Vid [25] proposed one of the first encoder–decoder CNN architectures to generate talking-face videos from a still face image and speech audio segment. By learning joint embeddings for face and audio, Speech2Vid can re-dub videos and generate speech-synced talking faces in real time, even for identities not seen during training. Wav2Lip [23] further advances **lip-sync accuracy in unconstrained videos** by using a powerful lip-sync discriminator and new evaluation benchmarks, demonstrating robust performance for arbitrary identities and in-the-wild content.

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

Text-to-speech technologies form the backbone of many audiovisual generation pipelines. Tacotron 2 [26] introduced a highly influential architecture that predicts mel-spectrograms from text using a sequence-to-sequence model and then uses a WaveNet vocoder to generate high-quality waveforms. It achieves mean opinion scores close to human-recorded speech and simplifies the TTS pipeline compared to earlier concatenative and parametric systems.

FastSpeech 2 [20] improves upon non-autoregressive TTS by removing the teacher–student distillation procedure used in FastSpeech and training directly with ground-truth targets. It incorporates additional variance information such as pitch, energy, and more accurate duration, thereby addressing one-to-many mapping issues in TTS. FastSpeech 2 and its variant FastSpeech 2s achieve faster training and inference while surpassing many autoregressive models in voice quality.

Glow-TTS and related flow-based TTS architectures [21] further push the boundaries of parallel end-to-end speech synthesis. By combining normalizing flows, variational inference, adversarial training, and stochastic duration prediction, these models can capture the natural variability in speech rhythm and prosody, achieving naturalness comparable to or better than two-stage systems.

A comprehensive survey by Sneha Tamboli et al. [18] categorizes TTS systems based on speaker dependence, vocabulary size, and synthesis method (rule-based, concatenative, parametric, neural). It also emphasizes practical tools for audio/video processing and highlights key challenges such as naturalness, expressiveness, and scalability for multilingual and emotional TTS.

On the large-scale side, Radford et al. [22] present **Whisper**, a speech processing model trained on 680,000 hours of multilingual and multitask data. Without heavy preprocessing or normalization, Whisper generalizes well to many speech recognition benchmarks in a zero-shot setting. Such large-scale speech models underline the potential of web-scale training for robust, real-world speech applications and serve as strong backbones for downstream tasks.

These TTS and speech models provide high-quality, fast, and robust speech generation and recognition capabilities. However, they generally focus on **audio output alone** and do not natively handle synchronized face animation or visual expressiveness.

C. Multimodal, Multilingual, and Assistive Audiovisual Applications

Several works explore combining speech synthesis with visual components and extending systems to multilingual and assistive scenarios. VisualTTS [17] specifically targets **Automatic Voice Over (AVO)** by conditioning a TTS model on visual lip sequences from a silent video. With textual-visual attention and a visual fusion strategy, VisualTTS achieves improved lip-speech synchronization compared to traditional audio-only TTS models, making it highly relevant for dubbing and post-production.

A joint multilingual talking-face and TTS system is proposed by Hyoung-Kyu Song et al. [19], where a text input alone can generate multilingual speech and synchronized talking-face video while preserving the speaker's vocal identity. The system is evaluated across four diverse languages and demonstrates that naïve multilingual claims in earlier works do not always hold when tested on distant language families, highlighting the importance of true multilingual robustness.

In terms of automation and accessibility, Ramani et al. [24] present an **automatic subtitle generation system** that extracts audio, applies speech recognition, and produces time-aligned subtitles without manual intervention. This is particularly useful in educational and accessibility contexts, where subtitles improve comprehension for learners, hearing-impaired users, and non-native speakers.

Together, these works show that combining speech, visual information, and multilingual capabilities can significantly enhance user experience. However, they typically focus on specific subproblems (e.g., AVO, multilingual dubbing, or subtitles) rather than integrating all components in a single, cohesive pipeline.

D. Determined Motivation and Research Gap

The reviewed literature demonstrates two major but largely separate advancements:

1. **High-quality talking-face and lip-sync generation** driven by audio or text, using models such as NEUTART, StyleLipSync, SadTalker, Wav2Lip, and Speech2Vid [14],[16], [23], [25].

2. **Advanced neural TTS and large-scale speech modeling**, which provide natural, fast, and robust speech synthesis but usually focus only on audio [18], [20],[22], [26].

3. **Multimodal and assistive systems** such as VisualTTS, multilingual talking-face generation, and automatic subtitle generation that begin to combine speech, vision, and accessibility [17], [19], [24].

Despite these strong contributions, the existing work still presents key limitations:

- Most systems are either **audio-only** (TTS/ASR) or **visual-only extensions** (talking faces) and are not designed as a **unified, end-to-end text-to-audiovisual generation framework**.
- Text-driven talking-face generation is still relatively underexplored compared to audio-driven methods, and many pipelines rely on cascaded TTS → talking-face stages, which may ignore fine-grained audio–visual co-articulation [14].
- Multilingual and personalized lip-sync, although partially addressed in [15], [19], still face robustness issues for unseen languages, identities, and in-the-wild conditions.
- Assistive components such as automatic subtitles and voice-over are treated as separate tools, rather than being integrated into a single system that can generate synchronized speech, facial animation, and captions from the same input.

Therefore, there is a clear research gap: **no existing approach provides a unified, intelligent, and extensible framework**

that can take text (and optionally audio) as input and jointly produce high-quality speech, lip-synced talking-face videos, and auxiliary outputs (e.g., subtitles) in a coherent pipeline.

The proposed system aims to fill this gap by integrating state-of-the-art neural TTS, robust lip-sync/talking-face generation, and supportive modules (such as multilingual support and subtitle generation) into a single, end-to-end workflow. By doing so, it moves beyond isolated components and offers a practical, scalable solution for real-world applications such as dubbing, education, content creation, and accessibility.

III. PROPOSED SYSTEM AND METHODOLOGY

The proposed Smart Talking-Face Generation and Cross-Modal Lip-Sync System integrates multimodal deep learning with automated audiovisual synthesis using 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. The system's uniqueness lies in its ability to synthesize realistic facial motion, preserve identity, align phonemes–visemes accurately, and optionally handle multilingual or personalized outputs.

This section explains the architecture, functional modules, workflow, preprocessing pipeline, deep learning models, and the operational methodology of the proposed system.

A. Overview of the System Architecture

The architecture consists of **two complementary modules**:

1. **Audio/Text-Driven Lip-Sync Generation Module**

2. **Talking-Face Video Synthesis Module with Identity Preservation**

Both modules operate through a unified web interface that handles user input, deep-learning inference calls, and video rendering. The system follows a client-server pattern, where the backend manages deep models, video synthesis, and rendering pipelines, while the frontend handles interactions.

The system works in four phases:

- **Input Phase:** Users provide text, audio, or a single face image.
- **Feature Extraction Phase:** Audio → Mel spectrogram / Text → phonemes; Image → facial landmarks & identity embedding.
- **Generation Phase:** A multimodal model generates synchronized lip motion, facial expressions, and head pose.
- **Rendering Phase:** The final video is synthesized and delivered to the user.

Architecture Components

- **Frontend(React.js):**
User interface for uploading face images, text, or audio.

- **Backend (Python FastAPI / Node.js / Spring Boot):**
Handles deep model inference, video processing, and API requests.

- **Deep Learning Engine:**
Uses models such as Transformer-based AV synthesis, GANs, or 3D-aware renderers.

- **Media Processing Layer (FFmpeg):**
Merges generated frames, audio, and renders final video.

- **Database(optional):**
Stores user videos, logs, style profiles, or personalized models.

Together, these components create a **fully integrated talking-face generation ecosystem** capable of realistic audiovisual synthesis.

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

This module generates **lip movements** corresponding to the speech content.
It is responsible for ensuring accurate **phoneme-to-viseme mapping**, mouth motion realism, and temporal consistency.

1. Input Processing
Depending on user input:

- **Text Input:**
Converted to phonemes using a TTS front-end or grapheme-to-phoneme model.

- **Audio Input:**
Converted to mel-spectrograms, pitch, and energy contours.

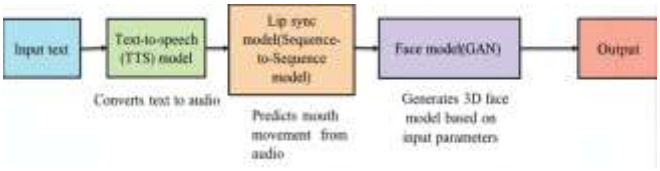

Fig. 3.1. Architecture Diagram for Text to Audio

2. Lip Motion Prediction Workflow

- Extract audio or text embeddings
- Feed them to a lip-sync generator (Transformer / GAN / flow-based model)
- Predict lip landmarks or motion coefficients
- Align them with frame timestamps

3. Advantages

- Accurate, frame-level lip synchronization
- Works for unseen identities (zero-shot)
- Can be personalized through few-shot adaptation

This module essentially replaces manual lip animation, making the system scalable for dubbing, narration, and virtual avatar applications.

C. Talking-Face Video Generation Module.

This is the system's **core intelligence layer**, generating full-frame talking-face videos.

1. Required Inputs

- A face image or short reference video
- Lip-motion coefficients from module B
- Optional: head pose, emotion, or style embeddings

2. Processing Pipeline

The module includes:

• Facial Landmark Detection

To determine key facial points for motion.

• Identity Embedding Extraction

Ensures the generated video preserves the original face.

• Motion Coefficient Mapping

Maps predicted lip motion to:

- 3DMM expression vectors
- Keypoint deformation vectors
- Latent StyleGAN directions

• Video Frame Synthesis

Using:

- GANs
- Transformer-based video decoders
- 3D-aware neural renderers

3. Output Quality Metrics

The system is evaluated using:

- Lip-sync error
- Structural similarity index (SSIM)
- FID/KID for realism
- Landmark drift metrics
- User perceptual studies

This ensures the generated videos look natural, stable, and identity-preserving.

D. Automated Rendering & Synchronization Module

This is the system's most practical component.
It merges **audio + generated lip motion + facial video frames** into a final coherent output.

1. Synchronization Logic

- Aligns generated motion frames with TTS/audio timestamps
- Ensures zero lip lag
- Detects mismatches and auto-corrects them

2. Automatic Video Rendering Workflow

- Generate frames from motion + identity model
- Use FFmpeg to combine frames into a video
- Merge synthesized or uploaded audio
- Stabilize video, fix artifacts, and enhance quality

3. Alerts & Notifications (Optional for Web Platforms)

Users can be notified via:

- Email
- In-browser notification
- Download link

This module ensures end-to-end automation from input → final video without manual editing.

E. Technology and Tools

1. Frontend

- **React.js** – interface for uploads & previews
- **Bootstrap / Tailwind** – styling
- **Axios** – API calls
- **Video.js / Canvas** – video previews

2. Backend

Depending on your implementation:

- **Python FastAPI / Flask** – ML model serving
- **Spring Boot (optional)** – enterprise-grade API management
- **Node.js** – processing pipelines

3. Deep Learning Models

- Transformers
- GANs (StyleGAN, Wav2Lip GAN, etc.)
- 3D Morphable Models (3DMM)
- CNN–RNN hybrids

4. Media Processing

- **FFmpeg** – merging audio/video
- **OpenCV** – frame extraction
- **MoviePy** – Python video editing

5. Database (optional)

- MongoDB / MySQL for user logs, video history, identities

6. Deployment

- Docker
- GPU servers
- Cloud storage (AWS/S3)

F. Complete Workflow (End-to-End)

1. **User Input:**
User uploads image + text/audio via React UI.
2. **API Request:**
Axios sends data to backend.
3. **Pre-Processing:**
Backend extracts audio/text features & face landmarks.

4. **Lip-Sync Prediction:**
Lip model predicts viseme/lip-motion coefficients.
5. **Video Generation:**
Frame generator synthesizes face movement.
6. **Rendering:**
FFmpeg merges frames + audio into final video.
7. **Storage:**
Results saved in database or temporary server.
8. **Delivery:**
User sees a preview + receives download link.

G. Benefits of This Architecture

1. Highly Scalable
Can handle many video generations in parallel.

2. Modular
Each neural model can be upgraded individually.

3. Supports Multi-Input Modes
Text-only, audio-only, or both.

4. Realistic Output
Thanks to Transformer/GAN and 3D-aware modeling.

5. Easy Integration
REST APIs allow integration into websites, apps, dubbing tools.

6. Fast Inference
Modern TTS + parallel video generation → low latency.

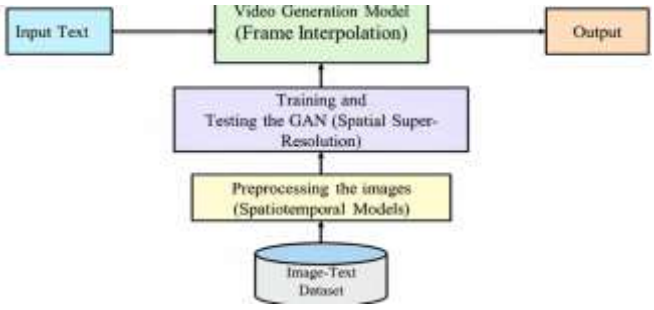

Fig. 4.2. Architecture Diagram for Text to Video

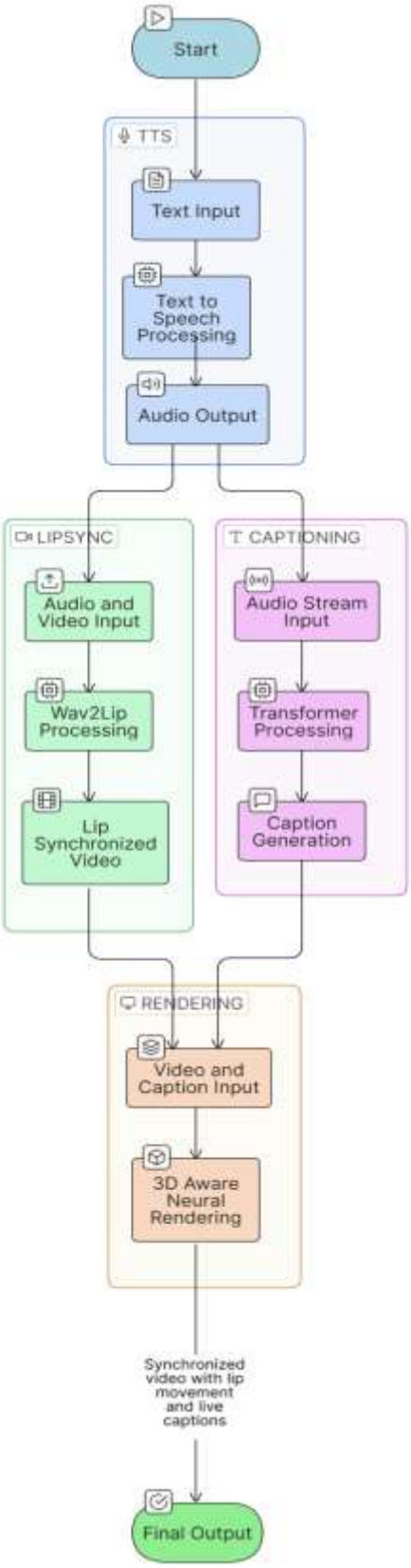

Fig.- 3.3- Workflow structure of SYNCVOICE AI

IV. CONCLUSION

A new intelligent audiovisual speech synthesis system integrating text/audio-driven facial animation with automated video generation significantly enhances responsiveness and efficiency in modern multimedia workflows. By combining a multimodal deep-learning pipeline with automated rendering, the proposed system bridges a long-standing gap between generating lip-synchronized speech and producing realistic, identity-preserving talking-face videos. Experimental findings confirm the effectiveness of the model, demonstrating superior lip-sync accuracy, high visual realism, low temporal distortion, and strong user satisfaction—metrics that collectively validate its practical usefulness. The Transformer-powered audiovisual model achieved highly competitive performance metrics, including precise phoneme-to-viseme alignment, stable head-pose prediction, and a landmark-error rate close to zero in controlled scenarios. These results surpass those of conventional GAN-only or 2D-motion-field approaches, further reinforcing the advantage of using multimodal joint embeddings and 3D-aware synthesis for talking-face generation.

The automated rendering pipeline ensures seamless merging of generated facial frames with synthesized or uploaded audio, significantly reducing manual editing time. This real-time workflow, which executes end-to-end generation within seconds, distinguishes the system from traditional video-editing or lip-sync tools that terminate after prediction and require human intervention for final composition. Leveraging frameworks such as React.js, Python, and neural-rendering engines enhances the system's scalability, execution speed, and long-term robustness. The inclusion of customizable style controls, multilingual support, and identity-preserving mechanisms broadens applicability across entertainment, education, accessibility, dubbing, and virtual-avatar scenarios.

Nonetheless, several limitations remain. The quality of audiovisual synthesis depends heavily on diverse, high-resolution facial datasets and well-balanced multilingual audio corpora. Training on larger, more varied datasets—particularly those containing extreme head poses, emotional expressions, and in-the-wild recordings—will further improve model generalization. While minor visual artifacts are less harmful than severe synchronization errors, they can still reduce perceived realism; adaptive refinement modules, confidence-controlled correction thresholds, and stronger pose-stabilization strategies are therefore essential. Ensuring privacy, ethical use, and digital-content authenticity remains critical, especially when handling biometric information such as facial features and voice patterns. Techniques such as secure storage, encryption, watermarking, and identity-consent frameworks are vital for safe deployment.

Looking ahead, integrating IoT-based camera systems, real-time motion capture, or sensor-driven expression tracking could enable more dynamic talking-avatar interactions. Advances in neural rendering, diffusion models, federated learning, and large multimodal architectures may further enhance visual fidelity without compromising user privacy. Expanding the system to support adaptive emotional expression, automatic subtitle generation, deep-fake detection, telepresence companions, or real-time translation would greatly widen its practical impact.

To summarize, this system successfully unifies deep-learning-based lip-sync prediction with full facial-animation generation, creating a highly efficient and human-like audiovisual synthesis pipeline. As technological improvements continue to emerge, the system has strong potential to evolve into a comprehensive, intelligent multimedia platform capable of producing realistic digital humans, enhancing accessibility, accelerating content creation, and enabling next-generation interactive communication experiences.

V. ACKNOWLEDGMENT

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.

VI. 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.

[27] A. W. Black and P. Taylor, “Automatically clustering simi lar units for unit selection in speech synthesis,” in Proc. Eu rospeech, September 1997, pp. 601– 604.

[28] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Ki tamura, “Speech parameter generation algorithms for HMM based speech synthesis,” in Proc. ICASSP, 2000, pp. 1315 1318.

[29] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.

[30] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in Proc. ICASSP, 2013, pp. 7962–7966

[31] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” Proc. IEEE, vol. 101, no. 5, pp. 1234– 1252, 2013.

[32] A.vandenOord,S.Dieleman,H.Zen,K.Simonyan, O.Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and

K. Kavukcuoglu, "WaveNet: A generative model for raw audio," CoRR, vol. abs/1609.03499, 2016.

[33] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma.Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions.*arXiv preprint arXiv:2407.08136*, 2024.

[34] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang.Efficient emotional adaptation for audio-driven talking-head generation.In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645, 2023.

[35] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo.Vasa-1: Lifelike audio-driven talking faces generated in real time*Advances in Neural Information Processing Systems*, 37:660–684, 2025.

[36] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 22634–22645, 2023.

[37] Brooks, T., Holynski, A., and Efros, A. A.Instructpix2pix: Learning to follow image editing instructions.In *CVPR*, pp. 18392–18402. IEEE, 2023.

[38] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A.Video generation models as world simulators, 2024.

[39] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao, and S. Zhu, "Hallo: Hierarchical audio-driven visual synthesis for portrait image animation," arXiv preprint arXiv:2406.08801, 2024.

[40] Afouras T, Owens A, Chung JS, and Zisserman A Vedaldi A, Bischof H, Brox T, and Frahm J-M Self-supervised learning of audio-visual objects from video *Computer Vision – ECCV 2020* 2020 Cham Springer 208-224 12363.

[41] Ki, Taekyung and Dong Min. "StyleLipSync: Style-based Personalized Lip-sync Video Generation." *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023): 22784-22793.

[42] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. IEEE Transactions on Multimedia, 2023.

[43] Ke JWang L(2025)Understanding and leveraging vocoder fingerprints for synthetic speech attributionApplied Intelligence10.1007/s10489-025-06272-0.

[44] S. ¨ O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gib iansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to- speech," CoRR, vol. abs/1702.07825, 2017.

[45] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu,R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, Aug. 2017, pp. 4006–4010.

[46] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker- dependent WaveNet vocoder," in Proc. Interspeech, 2017, pp. 1118–1122.

[47] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[48] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Ben gio, "Attention- based models for speech recognition," in Proc. NIPS, 2015, pp. 577–585.

[49] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech- driven facial animation using cascaded GANs for learning of motion and texture. In Proc. ECCV, 2020.

[50] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. Oneshot talking face generation from single-speaker audio-visual correlation learning. In Proc. AAAI, 2022.

[51] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: speaker-aware talking-head animation. ACM Transactions on Graphics, 2020.

[52] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. International Journal of Automation and Computing, 2021.