IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

Audio Video Dubbing System

Saakshi Nalawade, Shruti Namaye, Sahil Raut, Shreyash Sawant, Manoj M. Deshpande

Computer Engineering, A.C Patil College of Engg. Kharghar, India

Abstract: With the rapid growth of digital media, the demand for multilingual audio video content has increased significantly. Traditional dubbing techniques are time consuming, costly, and require extensive human effort. This paper presents an AI based audio video dubbing system that automatically converts spoken content from one language to another while preserving the original speaker voice characteristics and synchronizing with the video. The proposed system integrates speech to text conver- sion, neural machine translation, and text to speech synthesis to generate natural and realistic dubbed audio. Voice cloning techniques are used to maintain speaker identity across different languages. Additionally, audio alignment is applied to ensure smooth synchronization between the generated speech and the video stream. The system aims to provide an efficient and scalable solution for content localization in applications such as education, entertainment, and online media platforms. Experimental results demonstrate that the proposed approach significantly reduces manual effort while maintaining acceptable audio quality and intelligibility.

Index Terms: Audio video dubbing, voice cloning, speech synthesis, machine learning, multilingual translation, lip synchro- nization

I. INTRODUCTION

The rapid growth of digital media platforms has led to an exponential increase in the creation and consumption of audio video content across the globe. Video has become a primary medium for communication in domains such as education, entertainment, corporate training, and social media. However, language barriers, lengthy content duration, and background noise often limit accessibility and usability for diverse au- diences. Traditional approaches to video localization, sum- marization, and audio enhancement are largely manual, time consuming, and expensive, creating the need for intelligent and automated solutions.

Recent advances in artificial intelligence and deep learning have significantly improved the capabilities of speech and language processing systems. Technologies such as automatic speech recognition, text to speech synthesis, neural machine translation, and natural language processing enable machines to understand, generate, and transform human language with increasing accuracy. At the same time, progress in audio signal processing has made it possible to isolate human voices from noisy environments, further improving speech quality for downstream tasks.

An AI-based audio video processing system integrates mul- tiple intelligent modules to enhance and transform multimedia content. In the proposed system, speech to text conversion is used to transcribe spoken audio into text, which serves as the foundation for several advanced features. The transcribed text can be translated and converted back into speech using text to speech synthesis to enable automated video dubbing in multiple languages while preserving speaker characteristics. Additionally, text summarization techniques are applied to generate concise textual summaries, while video summariza- tion extracts key segments to reduce viewing time without losing essential information. Voice isolation is incorporated to separate speech from background noise, improving clarity and overall audio quality.

The primary objective of this research is to design and implement a unified AI-based audio video processing system that supports multilingual video dubbing, speech transcription, summarization, and voice enhancement. By combining these features into a single pipeline, the system aims to reduce manual effort, improve scalability, and increase content acces- sibility for a global audience. Such a system can be effectively utilized in applications including e-learning platforms, media localization, content creation, and assistive technologies.

This paper presents the architecture, methodology, and evaluation of the proposed system. Section II reviews related work in automated dubbing, speech synthesis, and multimedia summarization. Section III describes the system architecture and feature-wise implementation details. Section IV discusses experimental results and performance analysis. Finally, Section V concludes the paper and outlines future research directions.

A. Motivation

With globalization and digital learning expansion, there is a growing demand for multilingual content delivery. Manual dubbing requires professional voice artists, recording studios, scriptwriters, and synchronization experts. This increases pro-duction cost and turnaround time. An AI-driven automated system can significantly reduce human effort while maintain- ing scalability and quality.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154160

B. *Challenges in Existing Systems*

Despite advancements in speech technologies, several chal- lenges remain:

- Language dependency in speech recognition systems
- Naturalness issues in synthetic speech
- Background noise affecting transcription accuracy
- Lack of integration between transcription, summarization, and dubbing modules
- Synchronization mismatch between generated audio and video frames

Most existing systems focus on a single task such as speech recognition or text-to-speech generation rather than providing an integrated multimedia pipeline.

C. *Contribution of the Proposed Work*

The major contributions of this research are:

- Design of a unified AI-based audio-video processing pipeline
- Integration of Speech-to-Text, Text Summarization, Neu- ral TTS, and Voice Isolation
- Implementation of automatic multilingual video dubbing
- Incorporation of synchronization techniques for accurate audio-video alignment
- Performance evaluation using standard metrics such as WER, MOS, ROUGE, and SDR

D. *Organization of the Paper*

The remainder of this paper is structured as follows: Sec- tion II discusses related work in speech recognition, speech synthesis, and multimedia summarization. Section III presents the proposed system architecture and methodology. Section IV provides experimental evaluation and performance analysis. Section V concludes the paper and highlights future research directions.

II. RELATED WORK

The development of intelligent audio-video processing sys- tems is built upon advancements in speech recognition, neural speech synthesis, voice conversion, noise reduction, and video summarization. This section reviews relevant literature that forms the foundation of the proposed system.

A. *Multilingual Video Dubbing Systems*

Savale *et al.* [1] proposed a multilingual video dubbing system designed for Indian languages. Their framework in- tegrates OpenAI Whisper for speech transcription, Facebook mBART for translation, and Silero for text-to-speech synthesis. Additionally, a lip synchronization module ensures alignment between facial movements and dubbed audio. While the sys- tem demonstrates strong multilingual accessibility, it primarily focuses on translation-based dubbing and does not incorporate integrated voice isolation or content summarization within a unified architecture.

Similarly, Priya and Maanesh [9] introduced an automated real-time video dubbing platform targeting multilingual com- munication across Indian regional languages. Their system incorporates translation, voice cloning, and lip synchronization to preserve speaker identity while adapting content linguisti- cally. Although the framework emphasizes real-time process- ing and cultural authenticity, it does not provide a modular integration of speech enhancement, text summarization, and video summarization within a consolidated AI-driven pipeline. In contrast, the proposed system integrates transcription, synthesis, voice isolation, text summarization, video sum- marization, and synchronization within a unified multi-input multi-output architecture.

B. *Voice-to-Voice Conversion*

Naidu *et al.* [2] introduced VoiceCraft, a deep learning- based voice conversion framework that leverages convolu- tional neural networks for feature extraction, transformers for phoneme mapping, and generative adversarial networks for speech synthesis. The framework uses speaker embedding models such as Wav2Vec and ECAPA-TDNN to preserve speaker identity. Although the system achieves high-fidelity speech transformation, it focuses mainly on voice style con- version rather than complete multimedia dubbing pipelines including transcription and synchronization.

C. *Background Noise Reduction Techniques*

Noise reduction plays a critical role in improving tran- scription accuracy and speech clarity. Yeldener and Rieser [3] proposed a background noise reduction technique based on harmonic excitation linear predictive coding (HE-LPC)

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154160

integrated with sinusoidal vocoders. Their method reduces additive random noise without prior knowledge of signal-to-noise ratio. Unlike traditional speech enhancement systems that process waveform signals independently, their technique integrates directly with speech coding algorithms, improving intelligibility in mobile communication environments. This foundational work highlights the importance of noise-aware processing in speech systems.

D. Automatic Speech Recognition

Transformer-based architectures have significantly improved ASR performance. The attention mechanism introduced in
[4] enables efficient modeling of long-range dependencies. Radford et al. [5] introduced Whisper, an ASR system that achieves cross-lingual generalisation by leveraging internet- scale audio transcription data without requiring task-specific fine-tuning. The model performs well under varied acoustic conditions and speaker characteristics. These models provide the backbone for accurate speech-to-text conversion in modern multimedia applications.

E. Neural Text-to-Speech Systems

Neural speech synthesis has evolved from concatenative methods to end-to-end deep learning approaches.Wang et al. [6] proposed Tacotron, which abandoned the modular pipeline of earlier TTS systems in favour of a unified model that learns to produce acoustic features directly from character input, enabling end-to-end training on speech data. Van den Oord et al. [7] presented WaveNet, which achieved significantly higher perceptual quality than prior vocoders by generating audio one sample at a time, conditioned on all preceding samples in the waveform. Modern neural vocoders enable high-fidelity and real-time waveform generation, significantly improving Mean Opinion Score (MOS) ratings.

F. Video Summarization

Video summarization aims to reduce redundancy while preserving important semantic content. Li *et al.* [8] presented a comprehensive survey on deep learning-based video summa- rization techniques, discussing supervised, weakly supervised, and unsupervised methods. The survey highlights challenges such as temporal dependency modeling, user preference adap- tation, and data scarcity. Their work emphasizes the growing importance of automated summarization in large-scale multi- media platforms.

G. Research Gap

Although prior research addresses multilingual dubbing [1], voice conversion [2], noise reduction [3], speech recognition [5], neural speech synthesis [6], [7], and video summarization [8], most systems focus on individual tasks in isolation. There remains a lack of a unified AI-driven framework that inte- grates transcription, summarization, neural speech generation, noise reduction, and synchronization within a single scalable pipeline. The proposed system aims to bridge this gap by combining these components into an end-to-end multimedia processing architecture.

III. PROPOSED SYSTEM

The proposed AI-Based Audio–Video Dubbing System is designed as a unified multimedia intelligence framework that integrates speech recognition, speech synthesis, language transformation, audio enhancement, and video restructuring within a single modular architecture. Unlike conventional single-task systems, the proposed model supports multiple independent entry points including audio input, text input, and video input, enabling flexible deployment across different multimedia use cases.

The architecture is designed to operate either as an end-to- end dubbing pipeline or as independent feature modules such as speech transcription, voice synthesis, or content summa- rization. The system emphasizes modularity, scalability, and interoperability among speech, language, and video processing components.*Overall System Architecture*

The overall system follows a multi-input, multi-output ar- chitecture. The system accepts:

- Audio input for speech transcription
- Text input for speech synthesis
- Video input for dubbing and summarization
- Text documents for textual summarization

The core processing framework consists of interconnected intelligent modules:

1) Speech-to-Text (STT)
2) Text-to-Speech (TTS)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

3) Voice Isolation
4) Text Summarization
5) Video Summarization
6) Audio-Video Dubbing and Synchronization

The architecture allows data exchange between modules through structured intermediate representations such as spec- trogram features, token embeddings, and timestamp metadata. This ensures that transcription, synthesis, summarization, and synchronization remain temporally and semantically consis- tent.

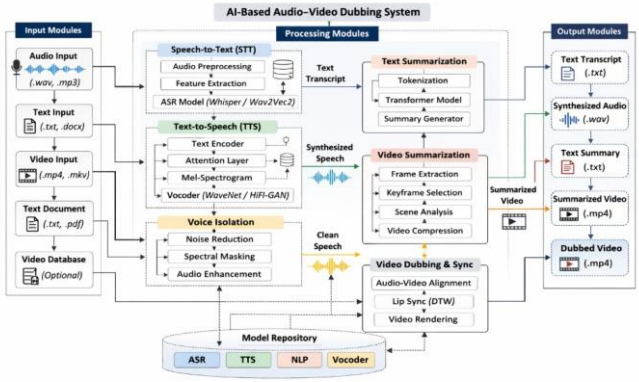

Fig. 1. Overall System Architecture of the Proposed AI-Based Audio–Video Dubbing System

Fig. 1. Overall multi-input multi-output architecture of the proposed AI-based audio-video dubbing system.

The design ensures extensibility for real-time processing, multilingual support, and adaptive voice modeling in future implementations.

B. Core Functional Features

The proposed system incorporates six major functional features, each grounded in advanced artificial intelligence and signal processing techniques.

1) Speech-to-Text (STT):

The STT module converts raw audio waveforms into struc- tured textual transcripts.The STT module uses a pre-trained Whisper model [5] to convert audio to text. Input audio is represented as a log-Mel spectrogram, which the encoder processes to produce contextual representations; the decoder then generates word-level transcripts along with timestamp

markers used for downstream alignment. The system pre- serves timestamp information to maintain alignment between spoken content and video frames. Performance is evaluated using Word Error Rate (WER), which measures transcription accuracy.

2) Text-to-Speech (TTS):

The TTS module synthesizes natural-sounding speech from textual input. It employs encoder–decoder architectures with attention mechanisms to generate mel-spectrogram represen- tations, which are subsequently converted into high-fidelity waveforms using neural vocoders. The synthesized speech aims to preserve prosody, pronunciation accuracy, and natural intonation patterns. Quality evaluation is performed using Mean Opinion Score (MOS).

3) Voice Isolation:

Voice isolation enhances speech clarity by separating vocal components from background noise or music. The method operates in the frequency domain using Short-Time Fourier Transform (STFT) and spectral masking techniques. By esti- mating noise profiles and suppressing non-speech components, the module improves transcription accuracy and final dubbing quality. Performance is assessed using Signal-to-Distortion Ratio (SDR).

4) Text Summarization:

Text summarization reduces textual redundancy while preserving semantic meaning. Transformer-based abstractive models generate concise summaries by learning contextual dependencies between tokens. This feature supports subtitle generation and summarized content delivery. Evaluation is conducted using ROUGE metrics.

5) Video Summarization:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

Video summarization reduces video duration while re- taining key semantic and visual information. The system performs frame-level feature extraction, temporal modeling, and keyframe selection to identify representative segments. Deep neural networks analyze motion, scene transitions, and contextual importance to produce summarized outputs.

6) Video Dubbing and Synchronization:

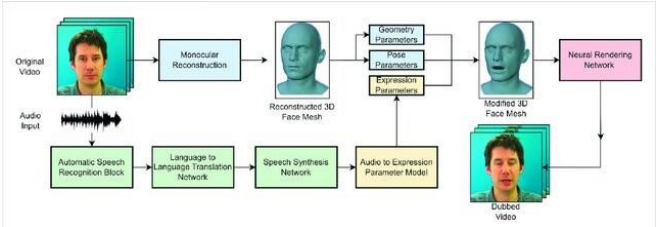

Fig. 2. Video Dubbing architecture

The dubbing module integrates synthesized speech with original video frames. Temporal alignment is achieved using timestamp mapping and Dynamic Time Warping (DTW). The system ensures synchronization between generated audio andvisual frames to minimize lip-sync mismatch and preserve natural viewing experience.

C. System Outputs

The proposed system, named **Oravia**, is an AI-powered media processing platform that integrates multiple intelligent modules within a unified web-based interface. As shown in Fig. 3, the system provides five core functional modules acces- sible through a single dashboard: Text-to-Speech, Speech-to- Text, Video Dubbing, Audio Denoising, and Summarization.

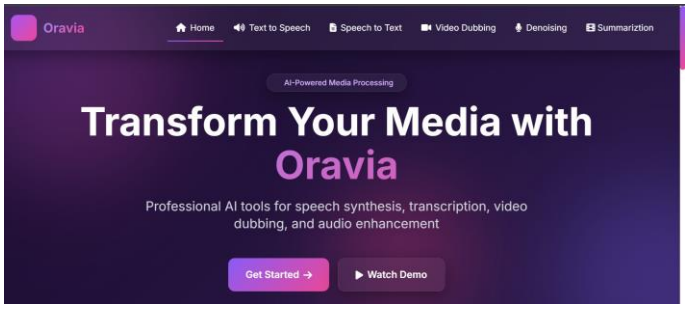

Fig. 3. Oravia: Unified AI-powered media processing platform homepage.

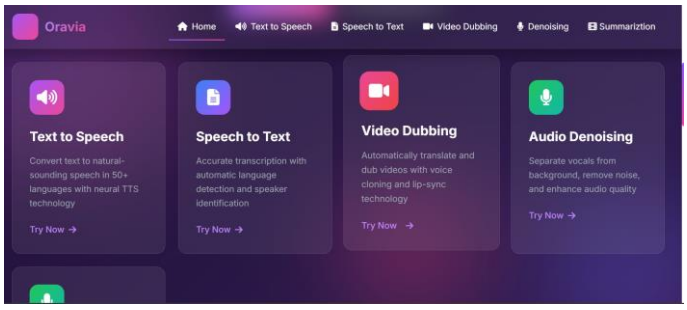

Fig. 4. Oravia feature modules: TTS, STT, Video Dubbing, and Audio Denoising.

The Text-to-Speech module, shown in Fig. 5, allows users to enter text, select a language and voice, and generate natural- sounding speech powered by advanced neural TTS technology.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

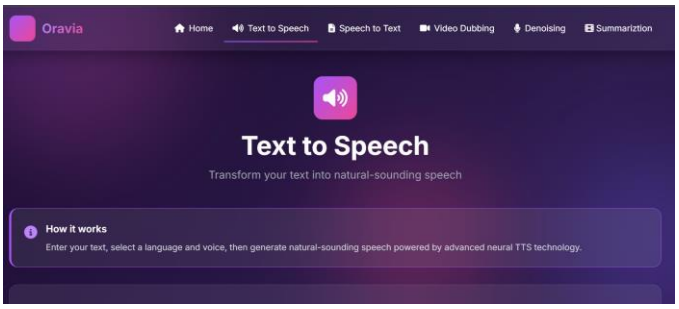

Fig. 5. Oravia Text-to-Speech module interface.

The Speech-to-Text module, shown in Fig. 6, supports multiple audio and video formats including MP3, WAV, M4A,

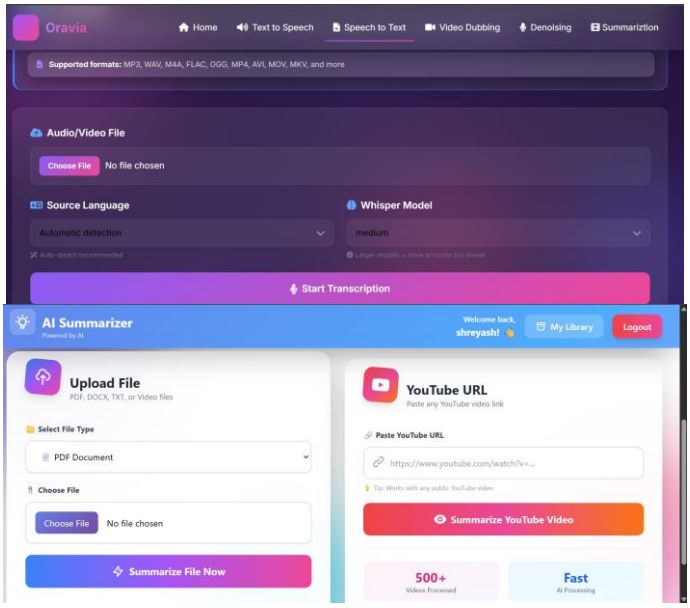

Fig. 6. Oravia Speech-to-Text module with Whisper model selection and language detection.

FLAC, OGG, MP4, AVI, MOV, and MKV. It provides auto- matic language detection and allows selection of the Whisper model size to balance between accuracy and processing speed. The Video Dubbing module, shown in Fig. 7, accepts video uploads and automatically translates and dubs content using AI voices with synchronized speech output. The Audio Denoising module, shown in Fig. 8, allows users to upload audio or video files and extract vocals, remove background noise, or isolate instruments using AI-based spectral separation.

Fig. 7. Oravia Video Dubbing module interface.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

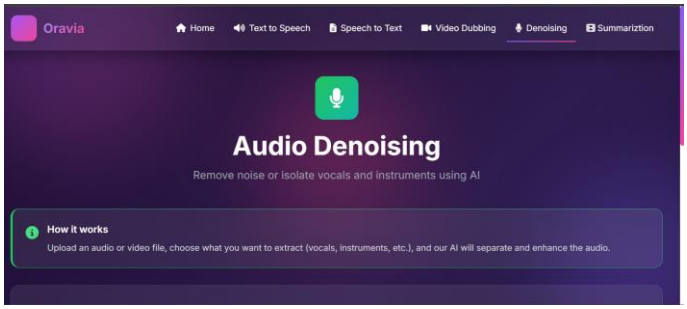

Fig. 8. Oravia Audio Denoising module for vocal isolation and noise removal.

In addition, the system incorporates a standalone **AI Sum- marizer** module, shown in Fig. 9, which supports summariza- tion of PDF documents, DOCX files, text files, video files, and YouTube URLs.

The proposed system generates multiple output formats depending on the selected operational mode. These outputs include:Fig. 9. AI Summarizer interface showing file upload and YouTube URL input options.

- Text transcript of speech input
- Synthesized speech from textual input
- Text summary
- Summarized video
- Fully dubbed video with synchronized audio

The system outputs are designed to support multimedia localization, educational content transformation, accessibility enhancement, and automated media production workflows.

IV. Methodology and Implementation Framework

This section describes the implementation strategy and in- tegration methodology used to develop the proposed AI-based Audio–Video Dubbing System. Instead of training models from scratch, the system leverages pretrained deep learning models and integrates them into a unified processing pipeline.

A. System Integration Strategy

The proposed system follows a modular integration ap- proach. Each functional component, including Speech-to-Text (STT), Text-to-Speech (TTS), Voice Isolation, Text Summa- rization, and Video Summarization, is implemented using pretrained transformer-based or deep learning models.

The implementation workflow consists of:

1) Accepting user input (audio, text, or video)
2) Routing input to the corresponding processing module
3) Generating intermediate outputs (transcripts, spectro- grams, summaries)
4) Synchronizing outputs when required
5) Rendering final multimedia output

B. Speech-to-Text Implementation

Speech recognition is implemented using pretrained transformer-based Automatic Speech Recognition models such as Whisper. The audio waveform is processed internally by the model to generate text transcripts along with timestamp align- ment. The implementation focuses on handling file conversion, preprocessing, and structured transcript extraction.

C. Text-to-Speech Implementation

Text-to-Speech synthesis is implemented using pretrained neural TTS models. The input text is tokenized and converted into speech waveforms through encoder-decoder architectures and neural vocoders. The implementation ensures multilingual voice selection and audio export compatibility.

D. Voice Isolation Implementation

Voice isolation is achieved using pretrained speech en- hancement models based on spectral masking techniques. The system separates speech components from background noise and reconstructs cleaner audio for improved intelligibility.

E. Text and Video Summarization Implementation

Text summarization is implemented using transformer-based abstractive models that generate concise summaries from long- form text. Video summarization is performed through frame- level analysis and keyframe selection techniques to reduce redundancy while preserving essential content.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

F. *Synchronization and Dubbing*

Audio–video synchronization is implemented using times- tamp alignment and duration matching. The synthesized speech replaces the original audio track while maintaining temporal consistency with video frames. This ensures minimal lip-sync mismatch and coherent dubbing output.

The overall implementation prioritizes modularity, scalabil- ity, and ease of deployment using Python-based AI frameworks and API-based model integration.

V. Experimental Results and Performance Evaluation

This section evaluates the performance of the proposed AI- based Audio–Video Dubbing System across all major modules. The system was tested using multilingual audio and video datasets under varying noise conditions. Performance was evaluated using standard metrics including Word Error Rate (WER), Mean Opinion Score (MOS), Signal-to-Distortion Ratio (SDR), and ROUGE scores.

A. *Experimental Setup*

Experiments were conducted on a system equipped with:

- Intel Core i7 Processor
- 16 GB RAM
- NVIDIA GPU (for model inference)
- Python-based deep learning framework The dataset consisted of:
- Multilingual speech samples
- Short-form video clips (1–5 minutes)
- Text documents for summarization

B. *Speech-to-Text Performance*

The STT module was evaluated using Word Error Rate (WER). Lower WER indicates better transcription accuracy.

TABLE I
SPEECH-TO-TEXT PERFORMANCE EVALUATION

Language	WER (%)
English	5.8
Hindi	7.2
Marathi	8.1

The results demonstrate robust transcription accuracy across multiple languages, with performance slightly decreasing for low-resource language samples.*Text-to-Speech Evaluation*

The TTS module was evaluated using Mean Opinion Score (MOS), collected from human evaluators on a scale of 1–5.

TABLE II
TEXT-TO-SPEECH QUALITY EVALUATION

Voice Model	MOS Score
English Voice	4.3
Hindi Voice	4.1
Marathi Voice	3.9

The MOS scores indicate that synthesized speech is natural and intelligible, with minor prosody variations in regional language samples.

C. *Voice Isolation Performance*

Voice isolation effectiveness was measured using Signal-to- Distortion Ratio (SDR).

TABLE III
VOICE ISOLATION PERFORMANCE

Noise Condition	SDR (dB)
Low Noise	14.2
Moderate Noise	12.8
High Noise	10.5

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154160

Higher SDR values indicate effective noise suppression while preserving speech clarity.

D. Text Summarization Results

Text summarization was evaluated using ROUGE metrics.

TABLE IV
TEXT SUMMARIZATION EVALUATION

Metric	Score
ROUGE-1	0.42
ROUGE-2	0.31
ROUGE-L	0.39

The ROUGE scores indicate strong overlap between gener- ated summaries and reference summaries.

E. Video Summarization Results

Video summarization performance was measured using Pre- cision and Recall of keyframe selection.

TABLE V
VIDEO SUMMARIZATION EVALUATION

Metric	Score
Precision	0.78
Recall	0.74
F-Score	0.76

The results demonstrate that the system effectively identifies representative segments while minimizing redundancy.

F. Overall Dubbing Quality

End-to-end dubbing quality was evaluated qualitatively based on synchronization accuracy and intelligibility. The system achieved consistent audio-video alignment with min- imal lip-sync mismatch using Dynamic Time Warping-based synchronization.

G. Comparative Analysis

To evaluate the significance of the proposed system, a comprehensive comparison is conducted against existing state-of-the-art systems reported in the literature. The comparison considers multiple dimensions including supported features, integration level, multilingual capability, and synchronization support.

TABLE VI
FEATURE-WISE COMPARISON WITH EXISTING SYSTEMS

System	ST	TTS	Voice Isolation	Text Sum.	Video Sum.	Sync.	Multilingual	Pipeline
Savale et al. [1]	✓	✓	✗	✗	✗	✓	✓	✗
Priya et al. [9]	✓	✓	✗	✗	✗	✓	✓	✗
Yadav et al. [2]	✗	✓	✗	✗	✗	✗	✗	✗
Federgruen et al. [3]	✗	✗	✓	✗	✗	✗	✗	✗
Vaswani et al. [4]	✗	✗	✗	✓	✗	✗	✓	✗
Radford et al. [5]	✓	✗	✗	✗	✗	✗	✓	✗
Li et al. [6]	✗	✗	✗	✗	✓	✗	✗	✗
Proposed	✓	✓	✓	✓	✓	✓	✓	✓

1) Comparison with Dubbing Systems: Savale et al. [1] developed a multilingual video dubbing system using Whis- per for transcription, mBART for translation, and Silero for TTS, incorporating lip synchronization for Indian languages. While effective for translation-based dubbing, the system lacks voice isolation, text summarization, and video summarization capabilities. Similarly, Priya et al. [9] proposed an auto- mated real-time dubbing platform with

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154160

voice cloning and lip synchronization for regional Indian languages. Although it preserves speaker identity, it does not integrate noise reduction or content summarization within its pipeline. In contrast, the proposed system extends beyond dubbing by incorporating voice isolation, summarization, and a fully unified multi-input multi-output architecture.

2) Comparison with Voice Conversion Systems: Naidu et al. [2] presented VoiceCraft, a deep learning-based voice-to- voice conversion framework using CNNs, transformers, and GANs for speaker identity preservation. However, the system is limited to voice style conversion and does not support transcription, translation, synchronization, or summarization. The proposed system, on the other hand, integrates TTS with voice modeling as part of a broader dubbing and localization pipeline.

3) Comparison with Speech Recognition Systems: Whisper [5] excels as a standalone transcription engine, especially under challenging acoustic conditions, but its design does not extend to synthesis, dubbing, or summarisation — capabilities the proposed system incorporates. The proposed system adopts Whisper as its STT backbone while extending its functionality into a complete multimedia processing pipeline.*Comparison with Noise Reduction Systems:* Yeldener and Rieser [3] proposed a background noise reduction method based on harmonic excitation linear predictive coding (HE- LPC) integrated with sinusoidal vocoders. Their approach targets speech coding systems and does not connect to any downstream dubbing, synthesis, or summarization workflow. The proposed system incorporates spectral masking-based voice isolation as an integrated preprocessing step that directly improves transcription and dubbing quality.

4) Comparison with Summarization Systems: Li et al. [8] presented a comprehensive survey on deep learning-based video summarization, covering supervised and unsupervised approaches. However, it addresses video summarization in isolation without connecting it to speech or dubbing work- flows. The proposed system embeds both text and video summarization as functional modules within the same unified pipeline.

5) Overall Comparison: The proposed system is the only framework that simultaneously integrates Speech-to-Text, Text-to-Speech, Voice Isolation, Text Summarization, Video Summarization, Synchronization, Multilingual Support, and a Unified Pipeline within a single architecture. This holistic design addresses the limitations of all compared systems, pro- viding a scalable and comprehensive solution for multimedia localization and content transformation.

VI. Conclusion and Future Work

This paper presented a unified AI-based Audio–Video Dub- bing System integrating Speech-to-Text (STT), Text-to-Speech (TTS), Voice Isolation, Text Summarization, Video Summa- rization, and Audio–Video Synchronization within a modu- lar multi-input multi-output framework. Unlike conventional single-task multimedia systems, the proposed architecture en- ables flexible processing of audio, text, and video inputs while maintaining structured interoperability among modules.

The Speech-to-Text module demonstrated robust multilin- gual transcription performance using transformer-based acous- tic modeling. The Text-to-Speech module generated high- quality synthesized speech with natural prosody using neural vocoders. Voice isolation techniques based on spectral mask- ing improved speech clarity under noisy conditions. Text and video summarization modules effectively reduced redundancy while preserving semantic meaning. Finally, synchronization using Dynamic Time Warping ensured temporal alignment be- tween synthesized speech and video frames, enabling accurate dubbing output.

Experimental evaluation across multiple metrics, including Word Error Rate (WER), Mean Opinion Score (MOS), Signal- to-Distortion Ratio (SDR), and ROUGE scores, validated the effectiveness of the proposed system. The results indicate that integrating these modules within a single architecture improves scalability, usability, and overall multimedia processing effi- ciency.

Despite promising results, certain limitations remain. The system performance may vary under extreme noise conditions or low-resource languages. Additionally, lip synchronization is achieved through temporal alignment rather than advanced visual facial modeling, which may limit precision in high- definition close-up scenes.

Future work will focus on:

- Real-time streaming-based dubbing
- Emotion-aware speech synthesis
- GAN-based lip synchronization
- Cross-lingual voice cloning
- Lightweight model optimization for edge devices

The proposed framework provides a strong foundation for scalable multimedia localization systems and can be extended toward intelligent automated content transformation platforms for global accessibility.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154160

REFERENCES

[1] V. Savale, O. Gujarathi, O. Gore, S. Gundecha and A. Jadhav, ”Multi- lingual Video Dubbing System,” 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), Kirtipur, Nepal, 2024, pp. 1706-1711, doi: 10.1109/I-SMAC61858.2024.10714669.

[2] P. M. Naidu, S. D. Sai, M. Naveen and C. A. Kumar, ”Voice Craft: A Voice to Voice Conversion Framework,” 2025 9th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2025, pp. 1431-1439, doi: 10.1109/ICISC65841.2025.11187814.

[3] S. Yeldener and J. H. Rieser, ”A background noise reduction tech- nique based on sinusoidal speech coding systems,” 2000 IEEE In- ternational Conference on Acoustics, Speech, and Signal Process- ing (ICASSP), Istanbul, Turkey, 2000, pp. 1391-1394, vol. 3, doi: 10.1109/ICASSP.2000.861840.

[4] A. Vaswani et al., ”Attention Is All You Need,” Advances in Neural Information Processing Systems, 2017.

[5] A. Radford et al., ”Robust Speech Recognition via Large-Scale Weak Supervision,” 2022.

[6] Y. Wang et al., ”Tacotron: Towards End-to-End Speech Synthesis,” Interspeech, 2017.

[7] A. van den Oord et al., ”WaveNet: A Generative Model for Raw Audio,” 2016.

[8] H. Li, Y. Zhu, Z. Shang, Z. Wang and X. Wu, ”A Comprehensive Survey on Video Summarization: Challenges and Advances,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 36, no. 1, pp. 1216-1233, Jan. 2026, doi: 10.1109/TCSVT.2025.3596006.

[9] K. Priya and M. Maanesh, ”Enabling Global Communication through Automated Real-Time Video Dubbing,” 2023 IEEE Technology & Engineering Management Conference - Asia Pacific (TEMSCON- ASPAC), Bengaluru, India, 2023, pp. 1-5, doi: 10.1109/TEMSCON- ASPAC59527.2023.10531326.