International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Real-Time Adaptive Speech Mixing and Audio-Video Synchronization

[1]Dr. P. Harini, [2]Yamini Chunduri,[3]Usha Sri Gunti, [4]Annapurna Balla
[1]HOD & Professor, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.
[2,3,4]U. G. Student, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

Abstract -*The rapid growth of online communication, virtual meetings, and multimedia applications has increased the need for high-quality audio and video synchronization. In many real-time communication systems, delays between audio and video, background noise, and uneven speech levels reduce the overall quality of user interaction. These issues can affect communication, especially in applications such as video conferencing, online education, live streaming, and digital broadcasting. To address these challenges, this project presents a Real-Time Adaptive Speech Mixing and Audio-Video Synchronization system that enhances speech quality while maintaining accurate synchronization between audio and video streams.*

The proposed system captures audio and video simultaneously and processes them using adaptive speech enhancement techniques. It removes background noise, balances the audio levels of multiple speakers, and automatically synchronizes speech with the corresponding video frames. The synchronization module continuously monitors timing differences and adjusts the playback to minimize delays without requiring manual intervention. This adaptive approach improves speech clarity, reduces latency, and ensures smooth lip synchronization during communication. The developed system provides a reliable and efficient solution for modern multimedia applications by delivering synchronized, high-quality audio and video output. It can be effectively used in online meetings, virtual classrooms, multimedia editing, live broadcasting, and entertainment platforms where clear communication and accurate synchronization are essential for an improved user experience.

Keywords: *Audio-Video Synchronization, Speech Mixing, Adaptive Audio Processing, Noise Reduction, Real-Time Communication, Multimedia Systems, Signal Processing.*

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

INTRODUCTION

In today's digital era, multimedia communication has become an important part of our daily lives. People regularly use video conferencing, online classes, webinars, and live streaming platforms for communication, learning, and entertainment. For these applications to provide a better experience, both audio and video must remain synchronized. Even a slight delay between speech and lip movement can reduce communication quality and make conversations difficult to follow.

Another common problem is background noise, which affects speech clarity and makes it difficult for listeners to understand the speaker. Existing systems often struggle to handle these challenges, especially in real-time environments. This project aims to overcome these problems by developing a Real-Time Adaptive Speech Mixing and Audio-Video Synchronization system. The proposed system improves speech quality through noise reduction, automatically balances audio levels, and synchronizes audio with video in real time. By combining adaptive signal processing techniques with intelligent synchronization methods, the system provides a smoother and more natural communication experience for users across different multimedia platforms.

LITERATURE SURVEY

Audio-video synchronization has been an active area of research because of the increasing demand for high-quality multimedia communication. Earlier synchronization methods mainly depended on fixed timing mechanisms, which worked well only under stable network conditions. However, these approaches often failed when delays occurred due to network traffic or processing speed, resulting in noticeable mismatches between speech and video.

Recent studies have introduced adaptive synchronization techniques that automatically detect and correct timing differences in real time. Researchers have also focused on improving speech quality using noise reduction, echo cancellation, and automatic gain control. Machine learning and deep learning models have further enhanced speech detection and synchronization by learning patterns from audio and video data. Although these techniques have significantly improved multimedia communication, achieving high synchronization accuracy with minimal processing delay continues to be an important research challenge.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

RELATED WORK

Several research studies have explored different approaches to improve speech processing and multimedia synchronization. Traditional systems mainly relied on manual synchronization or predefined delay settings, which were not flexible enough to adapt to changing network conditions. As a result, users often experienced delays between speech and video during live communication.

Recent advancements have introduced intelligent synchronization methods that continuously monitor both audio and video streams. Adaptive algorithms automatically detect synchronization errors and adjust playback timing whenever necessary. Many researchers have also integrated speech enhancement techniques to reduce background noise and improve voice clarity. These improvements have made multimedia applications more reliable and efficient, especially in video conferencing, live broadcasting, and online education. However, there is still a need for systems that provide better synchronization while maintaining low processing time and high speech quality.

EXISTING SYSTEM

Most existing multimedia systems process audio and video separately before combining them for playback. Although this approach works under normal conditions, it often creates synchronization problems when network delays or processing variations occur. As a result, users may notice that the speaker's lip movements do not match the audio, making communication less effective.

Another limitation is the poor handling of background noise and multiple speakers. Existing systems usually do not adjust audio levels automatically, which may result in uneven sound quality during conversations. Manual synchronization is often required to correct timing issues, making the overall process time-consuming and less efficient. These challenges highlight the need for an intelligent system that can automatically enhance speech quality and maintain accurate synchronization in real time.

PROPOSED SYSTEM

To address the limitations of existing multimedia systems, this project introduces a Real-Time Adaptive Speech Mixing and Audio-Video Synchronization system that automatically processes both audio and video streams. The system captures audio and video simultaneously, enhances speech by removing background noise, balances audio levels, and synchronizes speech with the corresponding video frames.

The synchronization module continuously

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

compares audio and video timing and automatically corrects any delay that occurs during processing. Unlike traditional systems, the proposed solution adapts to changing conditions without requiring manual adjustments. This helps maintain smooth communication, accurate lip synchronization, and better speech clarity. The system is designed to support applications such as online meetings, live streaming, digital broadcasting, and multimedia editing, where synchronized audio and video are essential for providing a better user experience.

SYSTEM ARCHITECTURE

The proposed Real-Time Adaptive Speech Mixing and Audio-Video Synchronization system follows a layered architecture to ensure efficient processing and smooth communication between different modules. It consists of five main layers: User Layer, Presentation Layer, Application Layer, Data Layer, and External Libraries**.**

The User Layer allows users to register, log in, and upload videos for processing. The Presentation Layer provides a simple web interface with pages such as Login, Dashboard, Upload, and Result, making the system easy to use.

The Application Layer, developed using the Flask framework, is the core of the system. It performs user authentication, extracts audio from uploaded videos, analyzes speech, reduces background noise, enhances voice quality, synchronizes audio with video, and generates the final output.

The Data Layer uses an SQLite database to store user information, session details, and uploaded video records securely. The system also integrates external libraries such as MoviePy**,** Librosa**,** Noisereduce**,** SoundFile**,** and NumPy to perform video processing, speech enhancement, noise reduction, and numerical operations efficiently. This architecture ensures accurate synchronization and improved multimedia quality.

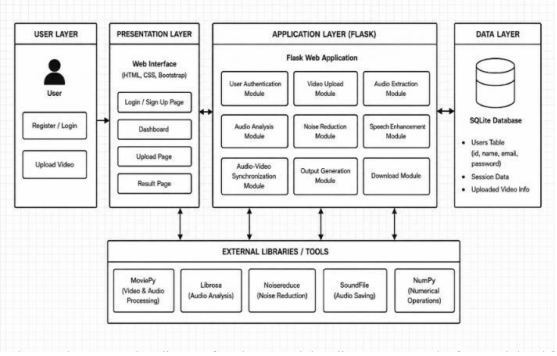

Fig 1: System Architecture

METHODLOGY DESCRIPTION

The system extracts audio from uploaded videos, removes background noise, enhances speech quality, synchronizes audio with video frames, and generates a clear, synchronized output for real-time multimedia communication.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

RESULTS AND DISCUSSION

The home page is the entry point of the application. It provides users with options to register or log in before accessing the multimedia processing features.

Fig 2: Home Page

The login page authenticates users using their registered email address and password. After successful authentication, the user is redirected to the dashboard.

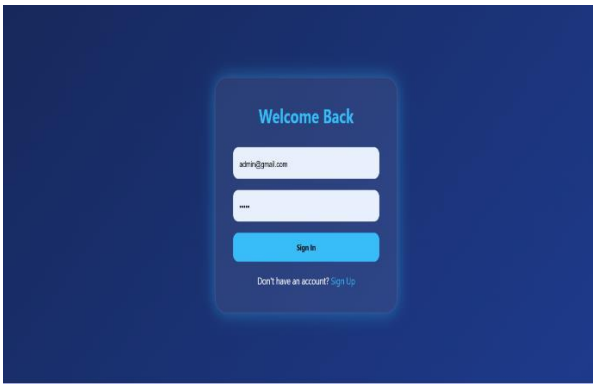

Fig 3: SignIn Page

The dashboard provides access to the main functionality of the application. From this page, users can upload video files for processing.

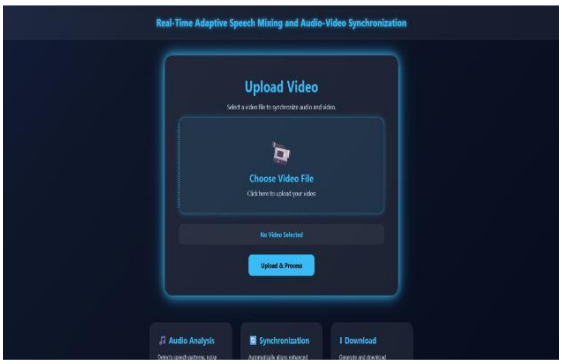

Fig 4: Dashboard Page

The extracted audio undergoes background noise reduction and speech enhancement. The processed audio provides improved clarity while preserving the original speech.

The enhanced audio is synchronized with the original video using MoviePy. This process ensures that the audio remains aligned with the video throughout playback.

After synchronization, the application generates the final output video with enhanced speech quality. The processed video can be previewed and downloaded by the user.

Fig 5: Results Page

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

CONCLUSION

The Real-Time Adaptive Speech Mixing and Audio-Video Synchronization system provides an efficient solution for improving multimedia communication by combining intelligent speech enhancement with accurate synchronization techniques. The system successfully reduces background noise, balances speech levels, and synchronizes audio with video in real time. This results in clear communication, reduced latency, and an improved user experience. The project demonstrates that adaptive signal processing can significantly enhance multimedia applications such as online meetings, digital broadcasting, live streaming, and virtual learning. Overall, the proposed system offers a reliable and practical solution for modern real-time communication environments.

FUTURE SCOPE

The proposed system can be further improved by integrating advanced Artificial Intelligence and Deep Learning techniques to increase synchronization accuracy and speech enhancement performance. Future versions may support multilingual speech processing, speaker identification, automatic subtitle generation, emotion recognition, and AI-based lip synchronization. Cloud-based deployment can enable remote multimedia processing and large-scale live streaming. Integration with 5G networks and edge computing technologies can further reduce communication delays and improve real-time performance. These enhancements will make the system more intelligent, scalable, and suitable for future multimedia communication applications.

REFERENCES

[1] D. P. Harini, “Electronic Health Record Stage Identification using Principal Component Analysis,” *Machine Intelligence Research,* vol. 18, no. 01, pp. 864- 873, 17 8 2024.

[2] D. P. Harini, “Two Level Intrusion Detection For Detecting Intruders in Multitier Web Applications,” *International Journal of Engineering & Science Research,* vol. 3, no. 9, pp. 472-478, 2013.

[3] D. P. Harini, “Analyze and Predict of Human Cyber Attackers using Artificial Neural Network,” *Journal of Basic Science and Engineering,* vol. 21, no. 1, pp. 1529 - 1536, 8 7 2024.

[4] Z. Yang et al., "Audio-Visual Speech Codecs: Rethinking Audio-Visual Speech Enhancement by Re-Synthesis," arXiv preprint arXiv:2203.17263, Mar. 2022. [Online]. https://arxiv.org/abs/2203.17263

[5] C. Chuang et al., "Improved Lite Audio-Visual Speech Enhancement," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2022. [Online]. https://ieeexplore.ieee.org/document/9734408

[6] P. Pan et al., "Selective Listening by Synchronizing Speech with Lips," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2022. [Online]. https://ieeexplore.ieee.org/document/9734404

[7] K. Cheng et al., "VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing in the Wild," arXiv preprint arXiv:2211.14758, Nov. 2022. [Online]. https://arxiv.org/abs/2211.14758

[8] R. Mira et al., "LA-VocE: Low-SNR Audio-Visual Speech Enhancement Using

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Neural Vocoders," arXiv preprint arXiv:2211.10999, Nov. 2022. [Online]. https://arxiv.org/abs/2211.10999

[9] A. Chern et al., "Audio-Visual Speech Enhancement and Separation by Utilizing Multi-Modal Self-Supervised Embeddings," arXiv preprint arXiv:2210.17456, Oct. 2022. [Online]. https://arxiv.org/abs/2210.17456

[10] J. Hong et al., "Visual Context-driven Audio Feature Enhancement for Robust End-to-End Audio-Visual Speech Recognition," arXiv preprint arXiv:2207.06020, Jul. 2022. [Online]. https://arxiv.org/abs/2207.06020

[11] X. Li et al., "Audio-Visual Multi-Channel Speech Separation, Dereverberation and Recognition," arXiv preprint arXiv:2204.01977, Apr. 2022. [Online]. https://arxiv.org/abs/2204.01977

[12] A. Golmakani et al., "Audio-Visual Speech Enhancement with a Deep Kalman Filter Generative Model," arXiv preprint arXiv:2211.00988, Nov. 2022. [Online]. https://arxiv.org/abs/2211.00988

[13] Y. Xu et al., "Improving Visual Speech Enhancement Network by Learning Audio-Visual Affinity with Multi-head Attention," arXiv preprint arXiv:2206.14964, Jun. 2022. [Online]. https://arxiv.org/abs/2206.14964

[14] A. Ephrat et al., "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," ACM Trans. Graph., vol. 37, no. 4, pp. 1–11, 2018. [Online]. https://arxiv.org/abs/1804.03619

[12] R. Gao and K. Grauman, "VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021. [Online]. https:/arxiv.org/abs/2101.03149

[13] P. Ma et al., "AV-HuBERT: Self-Supervised Learning of Audio-Visual Speech Representation," in Proc. Int. Conf. Learning Representations (ICLR), 2022. [Online]. https://arxiv.org/abs/2201.02184

[14] J. S. Chung and A. Zisserman, "Lip Reading Sentences in the Wild," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017. [Online]. https://arxiv.org/abs/1611.05358

[15] A. Ephrat et al., "Looking to Listen: Audio-Visual Speech Separation," ACM Trans. Graph., vol. 37, no. 4, pp. 1–11, 2018. [Online]. https://arxiv.org/abs/1804.03619