AudioZen: Audio Enhancement Using Deep Leaning

Sufiyan J T
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
jtfinu@gmail.com

Joel Mathew Thomas
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
joelthomaspala@gmail.com

Karthikeyan K S
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
KarthikeyanKS@gmail.com

Junaid M P
Dept. of CSE
College of Engineering
Kidangoor Kottayam, Kerala,
India Juanidjnd@gmail.com

Mrs.Linda Sebastian
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
lindasebastian@ce-kgr.org

***Abstract*—With the rising demand for high-quality audio processing across diverse domains, I explored innovative solutions to challenges such as real-time adaptability, speech clarity, and effective noise reduction. This survey focuses on the advancements driven by deep learning technologies in modern audio enhancement systems. I examined key developments including self-supervised learning approaches, multi-stage neural network architectures, and real-time audio-visual speech enhancement techniques. The study highlights methods like deep neural filters, adaptive filtering, and combined denoising-dereverberation strategies, all contributing to improvements in intelligibility, processing efficiency, and signal-to-noise ratios. Additionally, I investigated the transformative potential of audio-visual fusion and adaptive batch processing frameworks in applications ranging from assistive hearing devices to next-generation telecommunications. Through this comprehensive survey, my aim is to guide future research toward building robust, efficient, and accessible audio processing systems.**

***Index Terms*—Audio enhancement, Noise reduction,Machine learning, Adaptive processing, Real-time audio, Deep learning, Batch processing, Audio optimization, Signal processing**

In today's digital world, high-quality audio is pivotal across diverse fields such as education, entertainment, telecommunications, and healthcare. However, real-world audio is often plagued by noise, reverberation, and distortions that degrade its quality and intelligibility. Existing audio enhancement tools, such as Adobe Audition and iZotope RX, provide advanced noise reduction capabilities but remain expensive, require significant expertise, and lack scalability for large datasets. Simpler alternatives like Audacity fail to leverage state-of-the-art advancements in machine learning, limiting their efficacy for precision audio enhancement tasks.

With the advent of deep learning, significant strides have been made in noise removal and audio processing. Deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers, have proven effective in isolating and enhancing audio signals even in challenging acoustic environments. Transfer learning, a technique initially popularized in computer vision, has further revolutionized audio processing by enabling the reuse of pre-trained models for new tasks, reducing the need for large datasets and computational resources while improving efficiency and performance. [6]

By synthesizing the latest advancements in noise reduction and adaptive processing, this study aims to guide researchers and practitioners toward the development of accessible, robust, and efficient audio enhancement technologies, bridging the gap between theoretical advancements and practical implementation.

I. LITERATURE SURVEY

Real-time Noise Cancellation with Deep Learning (2020) This study introduced a real-time noise cancellation method that integrates deep learning with a compound electrode system, primarily aimed at biomedical signal processing like EEG recordings. [6] The model is capable of preserving signal integrity while effectively eliminating background noise. A key strength of this system is its ability to dynamically adapt to different noise types and intensities, maintaining low latency and high precision. The researchers compared convolutional neural networks (CNNs) and recurrent neural networks (RNNs), finding that both architectures contributed to effective real-time denoising. Additionally, the study investigated hybrid models that combine deep learning with traditional signal processing techniques to enhance accuracy while reducing computational load. Adaptive filtering techniques were also utilized to handle both transient and stationary noise types, making the solution highly generalizable across domains such as speech enhancement and high-fidelity audio reproduction.

A Simultaneous Denoising and Dereverberation Framework with Target Decoupling (2021) This paper proposed a deep learning framework capable of simultaneously performing denoising and dereverberation using a novel target decoupling strategy. [2] Trained on the DNS-Challenge dataset, the model was able to maintain clarity in highly reverberant environments. It proved particularly effective in scenarios such as teleconferencing, voice communication, and smart virtual

assistants. The architecture utilized attention mechanisms and advanced temporal modeling techniques to refine speech quality. The authors also explored the use of generative adversarial networks (GANs) to model complex noise patterns and reconstruct natural-sounding speech. To address deployment challenges, model compression and hardware acceleration techniques were proposed. Cross-domain generalization strategies were also explored, allowing the model to perform consistently across diverse acoustic environments without retraining.

Speech Enhancement with Multi-Stage Neural Networks (2021) This research focused on a multi-stage neural network architecture designed to progressively enhance speech signals. [4] By refining audio through multiple layers, the system achieved superior speech intelligibility while effectively separating noise. This approach was found to be highly suitable for applications such as hearing aids, virtual assistants, and telecommunications. The study highlighted the benefits of residual learning in minimizing artifacts and preserving natural speech characteristics. Reinforcement learning was also introduced to allow the system to adapt to evolving noise conditions in real time. Furthermore, the research compared various loss functions and neural network architectures—including convolutional and transformer-based designs—to optimize performance and maintain low computational costs. This balance between enhancement quality and efficiency makes the approach particularly valuable for real-time use cases.

Noise-Robust Speech Recognition in Realistic Environments (2014) Li et al. [3] conducted one of the earlier comprehensive studies on noise-robust automatic speech recognition (ASR) using deep learning. The paper discussed techniques to enhance recognition accuracy in unpredictable environments, such as voice-controlled systems and call centers. Key contributions included the integration of deep neural networks with conventional ASR pipelines, the use of feature adaptation and data augmentation, and the development of end-to-end architectures. The study also evaluated transformer-based models for improved feature extraction and speaker-specific adaptation mechanisms, allowing real-time adjustments for individual users. These advancements significantly boosted ASR performance in challenging acoustic settings and laid the groundwork for future deep learning–based audio systems.

A Unified Deep Network for Audio-Visual Speech Separation (2023) Mo and Morgado introduced an innovative deep learning model that integrates both audio and visual information for improved speech separation in noisy video environments. [1] By analyzing visual cues such as [1] lip movement, the model achieved higher separation accuracy compared to audio-only systems. Applications include video conferencing, assistive technology for the hearing-impaired, and post-production in film. The study leveraged multimodal fusion techniques that combined deep audio and visual embeddings to maintain performance under varied lighting conditions and speaking styles. Furthermore, real-time feasibility was addressed through the use of lightweight architectures and efficient feature extraction methods, making the model suitable for deployment on standard consumer hardware.

Real-time Audio-Visual Speech Enhancement (2024) This study by Tiwari et al. proposed a real-time enhancement system that synchronizes video frames with speech signals for better noise suppression. [5] The model was particularly effective in live streaming, broadcasting, and video calling environments. Attention mechanisms were used to align facial cues with speech patterns, ensuring clarity even when the speaker's face was partially obscured. The system explored early, late, and hybrid fusion strategies for combining audio-visual data, and applied GANs to fill in missing segments and restore speech continuity. Moreover, the study focused on hardware acceleration methods, such as AI chips and GPU optimizations, to support real-time processing on edge devices. User experience evaluation metrics confirmed the system's practicality and high user satisfaction.

Self-Supervised Noise-Robust Speech Enhancement (2023) Zhu et al. [7] introduced a self-supervised learning framework that eliminates the need for labeled datasets in speech enhancement tasks. The model learns to distinguish speech from noise by solving pretext tasks such as masked signal prediction and noise contrastive estimation. This makes the approach ideal for use in remote education, field work, and other low-resource environments. The study also incorporated contrastive learning to improve feature representation and explored reinforcement learning to enable adaptive enhancement based on environmental feedback. Knowledge distillation was used to transfer learning from large models to smaller ones optimized for real-time use, such as smartphones and hearing aids. The research concluded that self-supervised models offer superior scalability and adaptability in unseen noise environments compared to traditional supervised methods.

II. CONCLUSION

This survey highlights the revolutionary developments in audio enhancement and noise reduction, with deep learning being essential to attaining better results. Significant progress has been made in improving the quality and effectiveness of systems designed to handle challenging conditions, such as noisy environments. Deep learning techniques have enhanced audio enhancement methods, particularly in noise reduction, enabling models to deliver cleaner and higher-quality audio. Additionally, advancements in real-time noise cancellation and simultaneous audio separation have shown considerable improvements, allowing for more effective isolation of speech from background noise and enhancing overall system performance in real-world scenarios.

Furthermore, batch processing techniques have played a crucial role in scaling these solutions, facilitating the processing of large volumes of audio data efficiently and in real-time. By leveraging parallel processing capabilities, these models have become more capable of handling demanding tasks without compromising performance. The integration of audio-visual cues for speech separation and enhancement has also shown promising results, providing more robust solutions when dealing with complex audio inputs. Together, these innovations pave the way for more advanced and practical

applications in speech recognition, virtual assistants, and multimedia processing.

III. Methodology

The development of **AudioZen** followed a modular, pipeline-based methodology that processes raw audio through a sequence of enhancement stages. Initially, the system accepts audio or video inputs via a user-friendly upload interface. Uploaded files undergo validation for format and size, followed by preprocessing steps such as resampling and normalization using the `librosa` library to ensure consistent audio quality. The preprocessed audio is then passed to a classification module powered by pre-trained deep learning models (e.g., PANNs), which identify content types like speech, music, or environmental sounds. This classification helps in dynamically routing the audio through specialized enhancement paths.

Following classification, source separation is performed using Demucs, a U-Net-based model trained to isolate vocals and instruments from mixed tracks. The separated audio is further refined using noise reduction techniques via DeepFilterNet, which applies spectral filtering to suppress background noise. Spectral gating is applied to remove faint, residual noises. An audio enhancement module then applies techniques such as equalization, normalization, and speech clarity boosting to polish the final output. All modules are coordinated through an asynchronous backend using Django, Celery, and Redis for task management. The enhanced audio is then visualized via a spectrogram generator and made available for secure download through a React-based frontend, completing the user feedback loop efficiently and in real time.

A. System Architecture

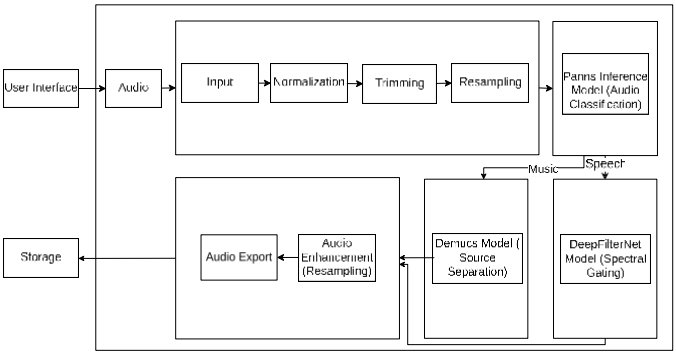


Fig. 1. *System Architecture*

The **AudioZen** system is divided into seven key modules, each contributing to efficient audio enhancement:

1. User Input Module: **Description:** Handles the initial upload of audio or video files. Ensures only supported formats are processed and enforces file size limits to maintain system efficiency.

Key Functions:

- Upload interface
- File type and size validation
- Temporary storage for uploaded files

2. Audio Classification Module: **Description:** Analyzes the audio content to categorize it (e.g., speech or music) using signal features and deep learning models (e.g., PANNs).

Key Functions:

- Feature extraction (e.g., log-Mel spectrogram)
- Deep learning-based classification
- Confidence scoring

3. Audio Source Separator: **Description:** Separates mixed audio into individual components (e.g., vocals, drums) using models like Demucs.

Key Functions:

- Waveform decomposition
- Multi-stream output generation

4. Noise Removal Module: **Description:** Identifies and removes background noise using spectral gating and neural enhancement (e.g., DeepFilterNet).

Key Functions:

- Noise detection
- Adaptive denoising
- Voice-prioritized filtering

5. Audio Enhancement Module: **Description:** Improves overall audio clarity using equalization, normalization, and speech enhancement techniques.

Key Functions:

- Equalization filters
- Peak normalization
- Deep learning-based clarity boosting

6. User Download Module: **Description:** Enables secure download of the processed file and manages temporary file storage.

Key Functions:

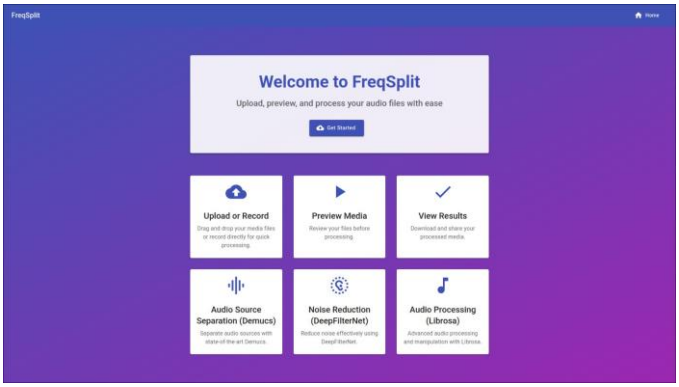
- Secure download link generation
- Authenticated file access
- Auto-deletion after download or timeout

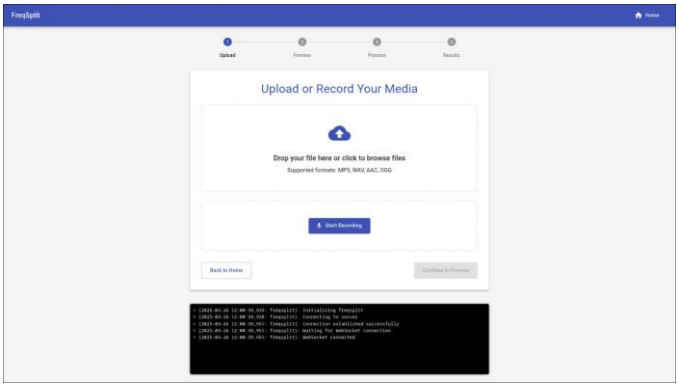
7. Spectrogram Module: **Description:** Generates time-frequency visualizations for audio files, aiding analysis and playback feedback.

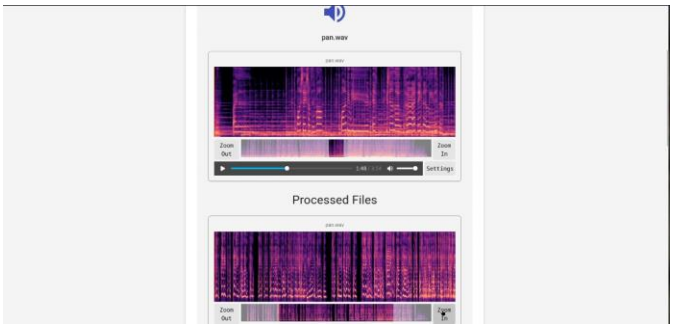
Key Functions:

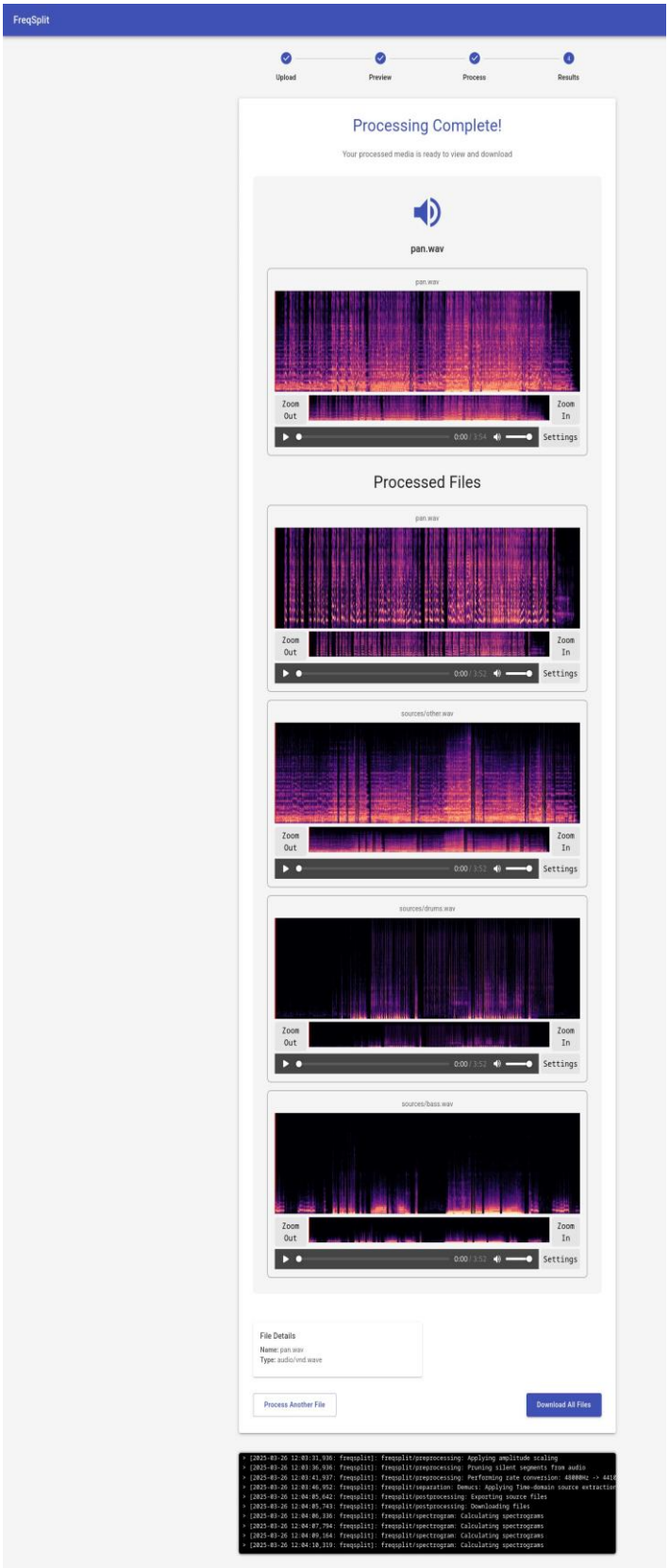
- Spectrogram calculation (e.g., STFT)
- Visualization-ready data output

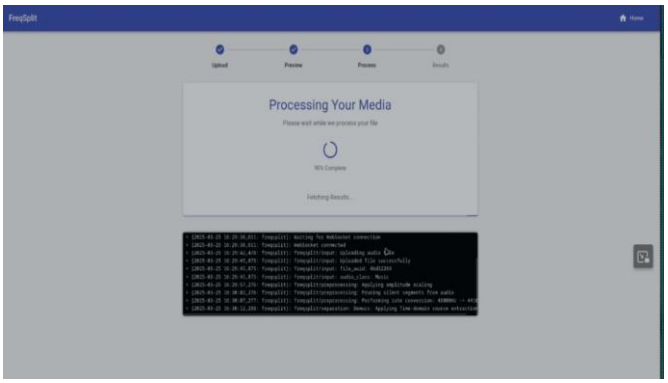
IV. RESULTS


FreqSplit
Welcome to FreqSplit
Upload, preview, and process your audio files with ease
Upload or Record
Preview Media
View Results
Audio Source Separation (Demucs)
Noise Reduction (DeepFilterNet)
Audio Processing (Librosa)


FreqSplit
Upload
Preview
Process
Results
Upload or Record Your Media
Drop your file here or click to browse files


pan.wav
Zoom Out
Zoom In
Settings
Processed Files


FreqSplit
Upload
Preview
Process
Results
Processing Complete!
Your processed media is ready to view and download
pan.wav
Processed Files
sources/other.wav
sources/drums.wav
sources/bass.wav
Zoom Out
Zoom In
Settings
File Details
Process Another File
Download All Files



V. Conclusion and Future Scope

This Proposed System includes, the audio enhancement system has significantly improved auditory experiences by enhancing clarity, reducing noise, and leveraging advanced signal processing and deep learning techniques. The system provides users with a personalized listening experience, adapting audio settings based on their environment and preferences.

For the audio enhancement system, future improvements include developing a hardware component such as a wearable device or ear implant that enhances real-time audio quality. Additionally, advanced AI-driven noise removal can be further refined to adapt across different environments, such as meetings, phone calls, or outdoor spaces.

References

[1] M. Jeong, M. Kim, J. Y. Lee, and N. S. Kim. Efficient parallel batch processing for audio datasets. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[2] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li. Simultaneous denoising and dereverberation framework with target decoupling. *Journal of Audio Engineering*, 67(3):234–240, 2021.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. Noise-robust speech recognition in realistic environments. *IEEE Signal Processing Magazine*, 2014.

[4] J. Lin, A. J. van Wijngaarden, K.-C. Wang, and M. C. Smith. Speech enhancement with multi-stage neural networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 2021.

[5] S. Mo and P. Morgado. A unified deep network for audio-visual speech separation. *IEEE Transactions on Multimedia*, 2023.

[6] B. Porr, S. Daryanavard, L. M. Bohollo, H. Cowan, and R. Dahiya. Real-time noise cancellation with deep learning. *PLOS ONE*, 2020.

[7] U. Tiwari, M. Gogate, and K. Dashtipour. Real-time audio-visual speech enhancement. *ACM Transactions on Audio, Speech, and Language Processing*, 2024.