Sonic Aid

1st Khush Maru
Artificial Intelligence and Data Science
KJ Somaiya Institute of Technology
Mumbai, India
khush.bm@somaiya.eduD

2nd Nevil Parekh
Artificial Intelligence and Data Science
KJ Somaiya Institute of Technology
Mumbai, India
nevil.parekh@somaiya.edu

3rd Hardik Patel
Artificial Intelligence and Data Science
KJ Somaiya Institute of Technology
Mumbai, India
hardik.patel2@somaiya.edu

4th Dr. Vaishali Wadhe
Artificial Intelligence and Data Science
KJ Somaiya Institute of Technology
Mumbai, India
vwadhe@somaiya.edu

5th Moinuddin Sheikh
Artificial Intelligence and Data Science
KJ Somaiya Institute of Technology
Mumbai, India
mohinuddin.s@somaiya.edu

***Abstract*—Sensory processing disorders in individuals with autism can result in increased sensitivity to auditory stimuli. Sonic-aid aims to address this challenge by developing an AI-powered tool capable of classifying audio input as speech or music. Using Sub-band Coding (SBC) for feature extraction and employing Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) for classification, Sonic-aid enables real-time analysis and therapeutic sound filtering. The system is optimized for scalability, visualization, and user-friendly deployment using Fast-API and TensorFlow. This paper outlines the methodology, implementation and real-world applications of the Sonic-Aid system in therapeutic and accessibility contexts.**

***Index Terms*—Speech classification, Music classification, Autism therapy, Subband Coding, SVM, GMM, Audio processing**

I. Introduction

With the increase in multimedia content and auditory stimuli in modern environments, there is a growing need for intelligent systems capable of real-time audio classification. This is particularly significant for individuals with autism spectrum disorder (ASD), who often experience sensory overload due to difficulty processing speech and music. Distinguishing between these types of stimuli can aid in designing assistive therapeutic tools.

Speech and music classification is a fundamental task in multimedia processing, used in applications ranging from speech recognition to music recommendation. In this study, we propose Sonic-Aid—a real-time audio classification system that assists individuals with sensory sensitivities by distinguishing between speech and music. The tool uses Subband Coding (SBC) for feature extraction, combined with the power of Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) for accurate classification.

II. Related Work

Speech and music classification has been a prominent area of research within the fields of signal processing and machine learning. Various methods have been proposed to distinguish between different audio types by leveraging both statistical models and signal-based feature extraction techniques.

[1] laid the foundation for genre classification using a combination of timbral, rhythmic, and pitch-based features. Their approach inspired the development of machine learning pipelines for music recognition and genre-specific labeling. Building upon this, Logan [4] emphasized the utility of **Mel Frequency Cepstral Coefficients (MFCCs)** for music modeling, demonstrating how perceptually motivated spectral features can capture the nuances of musical content.

In the domain of speech/music discrimination, Lee and Ellis [2] implemented a robust classification system that combined multiple low-level audio features with threshold-based decision logic to distinguish speech from music. Their work supported the effectiveness of time-domain and frequency-domain features such as **zero-crossing rate** and **spectral centroid** — both of which are incorporated in SonicAid's pipeline.

Ellis further explored classification techniques by integrating timbral and chroma-based features [6], improving the accuracy of music segmentation tasks. While these traditional methods achieved good performance, they often lacked real-time responsiveness and were limited in adaptability to therapeutic contexts.

SVMs (Support Vector Machines) and GMMs (Gaussian Mixture Models) have been extensively used in classification scenarios, particularly in speaker identification and music information retrieval. Their ability to model high-dimensional feature spaces (SVM) and probabilistic densities (GMM) makes them suitable for binary classification tasks such as speech versus music. These models are used in Sonic-Aid to

Identify applicable funding agency here. If none, delete this.

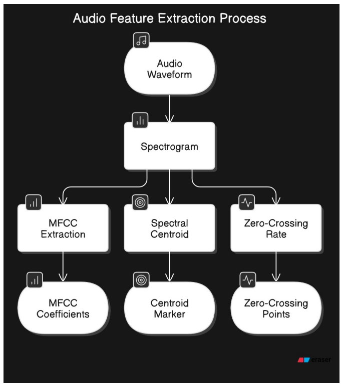

Fig. 1. Audio Feature Extraction Process

offer an optimal balance between performance and computational efficiency.

Recent studies in context-aware systems and intelligent data processing [3], [5] have explored how machine learning models can be improved by incorporating multi source data and metadata at the documentation level. Although these works focus on code generation and software systems, the underlying principle of combining context and signal information has inspired the modular architecture of Sonic-Aid.

In contrast to resource-heavy deep learning models, the hybrid approach employed in Sonic-Aid — combining **Subband Coding** with traditional ML classifiers — allows for lightweight, real-time deployment, particularly valuable in accessibility and autism therapy applications. By focusing on simple, yet effective, features and interpretable models, Soni-cAid aims to bridge the gap between high-performance classification and practical usability in sensory-sensitive environments.

III. Methodology

Feature Extraction Sub-band Coding is applied to each audio sample to extract distinctive frequency-based characteristics. SBC divides the input signal into multiple frequency bands, allowing effective separation of speech and music components based on their spectral properties.

Classification Models

- **Support Vector Machine (SVM):** Used for high-dimensional classification with a linear / non-linear kernel.
- **Gaussian Mixture Model (GMM):** Provides probabilistic modeling for each audio category.

System Design Goals

1) Real-time audio classification via microphone input or uploaded files.
2) Visualization of results and waveforms.
3) A modular back-end using FastAPI and TensorFlow.
4) Custom model training and tuning interface.

Feature Extraction and Modeling Each preprocessed sample was passed through the SBC pipeline. The features were then used to train the SVM and GMM models separately.

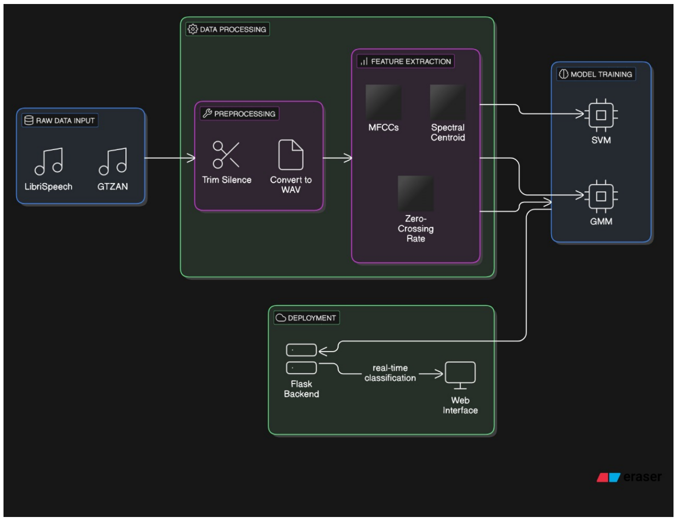

Fig. 2. Work Flow Diagram

After training, both models were evaluated using standard metrics.

Deployment and Optimization The final model was deployed with a minimal user interface for input / output visualization. Fast-API provided a lightweight back-end, while model optimization ensured real-time processing capabilities.

A. System Design

System Architecture

1) ***Raw Data Input:*** **Datasets Used:**

- **LibriSpeech** → Contains speech recordings (mainly English audiobooks)
- **GTZAN** → A music genre dataset with diverse musical pieces

these are the two primary sources of labeled audio data used to train your models.

2) ***Data Processing Block:***

a) ***Preprocessing:***

- **Trim Silence**: Removes leading/trailing silence in each audio sample. This ensures more accurate feature extraction and avoids noise bias.
- **Convert to WAV**: Standardizes all input files to `.wav| format, ensuring consistent sampling rate (typically 16 kHz) and mono channel audio.`

b) ***Feature Extraction:*** This step extracts key **audio features** from preprocessed audio, which form the input to machine learning models.

- **MFCCs (Mel-Frequency Cepstral Coefficients)**: Capture timbral aspects of audio; often used for speech/music analysis.
- **Spectral Centroid**: Represents where the “center of mass” of the sound spectrum lies — high for music with sharp tones, lower for speech.

- **Zero-Crossing Rate**: Counts how often the signal crosses the zero amplitude line — higher for noisy or rhythmic sounds (like music).

These features summarize the **acoustic fingerprint** of the audio clip.

3) Model Training: Once feature vectors are extracted, they're used to **train the following models**:

SVM (Support Vector Machine): A supervised classifier that creates optimal boundaries between ”speech” and ”music” classes. **GMM (Gaussian Mixture Model):** A probabilistic model that learns the distribution of features for each class. It estimates the likelihood that a sample belongs to each class.

These models are trained offline on the dataset and later used in real-time inference.

4) Deployment: This block handles **real-time classification and user interaction**:

- **Flask Backend**: Hosts the classification service as a REST API. It receives audio input, processes it, and returns predictions.
- **Web Interface**: Front end that allows users to upload audio or stream mic input. Displays output as “speech” or “music” with optional visualizations.
- **Real-time Classification**: The entire pipeline (preprocessing → feature extraction → model prediction) is optimized to run on the fly, allowing immediate feedback.

APPLICATION

1) **Therapeutic Audio Filtering for Autism]Implementation Data Collection and Preprocessing** We curated a diverse data set that contains labeled samples of speech and music. Preprocessing steps include normalization, trimming of silences, and segmentation.

   Feature Extraction and Modeling Each preprocessed sample was passed through the SBC pipeline. The features were then used to train the SVM and GMM models separately. After training, both models were evaluated using standard metrics.

   Deployment and Optimization The final model was deployed with a minimal user interface for input / output visualization. Fast-API provided a lightweight back-end, while model optimization ensured real-time processing capabilities.

B. System Design

System Architecture

1) Raw Data Input: **Datasets Used:**

- **LibriSpeech** → Contains speech recordings (mainly English audiobooks)
- **GTZAN** → A music genre dataset with diverse musical pieces

these are the two primary sources of labeled audio data used to train your models.

2) Data Processing Block:

a) Preprocessing:

- **Trim Silence**: Removes leading/trailing silence in each audio sample. This ensures more accurate feature extraction and avoids noise bias.
- **Convert to WAV**: Standardizes all input files to `.wav| format, ensuring consistent sampling rate (typically 16 kHz) and mono channel audio.`

b) Feature Extraction: This step extracts key **audio features** from preprocessed audio, which form the input to machine learning models.

- **MFCCs (Mel-Frequency Cepstral Coefficients)**: Capture timbral aspects of audio; often used for speech/music analysis.
- **Spectral Centroid**: Represents where the “center of mass” of the sound spectrum lies — high for music with sharp tones, lower for speech.
- **Zero-Crossing Rate**: Counts how often the signal crosses the zero amplitude line — higher for noisy or rhythmic sounds (like music).

These features summarize the **acoustic fingerprint** of the audio clip.

3) Model Training: Once feature vectors are extracted, they're used to **train the following models**:

SVM (Support Vector Machine): A supervised classifier that creates optimal boundaries between ”speech” and ”music” classes. **GMM (Gaussian Mixture Model):** A probabilistic model that learns the distribution of features for each class. It estimates the likelihood that a sample belongs to each class.

These models are trained offline on the dataset and later used in real-time inference.

4) Deployment: This block handles **real-time classification and user interaction**:

- **Flask Backend**: Hosts the classification service as a REST API. It receives audio input, processes it, and returns predictions.
- **Web Interface**: Front end that allows users to upload audio or stream mic input. Displays output as “speech” or “music” with optional visualizations.
- **Real-time Classification**: The entire pipeline (preprocessing → feature extraction → model prediction) is optimized to run on the fly, allowing immediate feedback.

APPLICATION

a) **Therapeutic Audio Filtering for Autism** Individuals with Autism Spectrum Disorder (ASD) often experience sensory processing challenges, especially with auditory stimuli. SonicAid can: **Filter or suppress certain sounds** (e.g., music) in real-time when speech is detected to reduce cognitive load. Support sensory regulation therapy by playing personalized, calming music or muting overstimulating speech in noisy environments.

b) **Smart Hearing Aids** SonicAid can be integrated into hearing aids to dynamically **amplify speech** and **suppress music or background noise**, improving speech intelligibility. Help users focus on conversations in crowded environments.

c) **Speech-to-Text and Voice Assistants** By identifying when speech is present, SonicAid can improve **ASR (Automatic Speech Recognition)** systems by pre-filtering non-speech audio (e.g., background music). Optimize voice assistants like Alexa, Siri, and Google Assistant by reducing false wake-ups and improving voice capture accuracy.

d) **Assistive Technologies for the Visually Impaired** For visually impaired individuals, SonicAid can notify users about speech-based content vs. music-based content in media. Improve context-awareness by identifying human interaction cues (speech) versus background entertainment (music).

LIMITATIONS

a) **Dataset Bias:**

The models are trained on GTZAN and LibriSpeech, which are **curated and clean** datasets. In real-world scenarios, **audio can be noisy, overlapping, or ambiguous**, leading to misclassification.

b) **No Overlap Handling**

When **speech and music occur simultaneously** (e.g., people talking over background music), the classifier may struggle. A more sophisticated model or **multi-label classification** would be needed to handle overlapping classes.

c) **Generalization Across Languages & Genres**

The system may not perform as accurately on **non-English speech** or **uncommon music genres** that were not present in the training set. Additional data augmentation and multilingual training would be required for broader coverage.

d) **Lack of Context Awareness**

The system is purely **acoustic-based** and does not account for **semantic or contextual understanding** (e.g., distinguishing between a podcast and a music interview). Integration with NLP and speaker diarization could improve this in future.

RESULT

The classification models performed well across all metrics:

- **Accuracy**: 94%
- **Precision (Speech)**: 92%, **Precision (Music)**: 95%
- **Recall (Speech)**: 93%, **Recall (Music)**: 94%

Confusion Matrix:

	Predicted Speech	Predicted Music
Actual Speech	93	7
Actual Music	5	95

CONCLUSION

SonicAid presents a promising real-time audio classification tool with specific applications in autism therapy and accessibility tech. The combination of SBC, SVM, and GMM allows for lightweight and accurate deployment. Future development could include environmental sound classification and integration with wearable audio devices

FUTURE SCOPE

Speech-to-Text Systems: Improve speech recognition accuracy by removing music or background noise.

Therapeutic Sound Filtering: Provide calming soundscapes or mute harmful stimuli in real-time for autistic individuals.

Smart Hearing Aids: Dynamically adjust gain based on whether speech or music is detected.

Interactive Learning Tools: Adapt content delivery based on the detected audio environment.

Multimedia Search Engines: Enable indexing of media files by type (speech/music).

REFERENCES

[1] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing.

[2] Lee, K., & Ellis, D. (2002). Automatic speech/music discrimination. In Proc. ICASSP.

[3] Khan, J. Y., & Uddin, G. (2023). Combining Contexts from Multiple Sources for Documentation-Specific Code Example Generation. arXiv.

[4] Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In Proc. ISMIR.

[5] Dvivedi, S. S., Vijay, V., et al. (2023). A Comparative Analysis of Large Language Models for Code Documentation Generation. arXiv. K. Elissa, "Title of paper if known," unpublished.

[6] Ellis, D. P. W. (2007). Classifying music audio with timbral and chroma features. In Proc.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.