STUDY AND ANALYSIS OF DIFFERENT AUDIO SIGNALS USING ARTIFICIAL INTELLIGENCE

Dr. A.A. Tayade
Asst. Prof. & Head, Department of Computer Science, G. S. College, Khamgaon, arvindtayade40@gmail.com

Ms. Vaibhavi Chaporkar
Asst.Prof. , Department of Computer Application, Vidyabharati Mahavidylaya, Amravati vaibhavi.chaporkar@gmail.com

Dr. R. K. Nawasalkar
Asst.Prof. & Head, Department of Computer Science, G.S. Tompe College, Chandur Bazar ram.nav1978@gmail.com

Abstract
Audio signals, ranging from speech and music to environmental and biomedical sounds, carry complex patterns that require sophisticated analysis methods. Earlier approaches in signal processing relied on manually engineered features, which often struggled with noisy or diverse input conditions. The integration of Artificial Intelligence (AI), especially Machine Learning (ML) and Deep Learning (DL), has enabled automated feature extraction, improved classification accuracy, and more efficient real-time processing. This work examines AI-driven methodologies for analyzing various audio signal types, including spoken language, musical compositions, ambient sounds, and health-related acoustic data. Advanced models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures are discussed in the context of tasks like recognition, enhancement, and synthesis. Comparative insights show that AI-based techniques generally surpass traditional algorithms in precision, adaptability, and scalability, while also introducing challenges related to interpretability, computational demands, and low-resource data handling. The study offers a unified perspective on applying AI to audio signal analysis, with potential applications in voice assistants, music information retrieval, smart city sound monitoring, and AI-supported medical diagnostics.

Keywords: *Audio signal analysis, Artificial Intelligence, Deep learning, Speech processing, Music information retrieval, Environmental sound classification.*

1. Introduction

The interpretation and analysis of audio signals are foundational in many technological and scientific fields such as speech processing, music retrieval, environmental monitoring, and healthcare diagnostics. Audio signals time-varying waveforms representing sound ncode detailed information about their source and surrounding context. Traditionally, analyzing these signals depended on intricate signal processing methods that relied heavily on manual feature engineering and expert domain knowledge. However, the rise of Artificial Intelligence (AI), especially machine learning and deep learning, has revolutionized this domain.

AI allows systems to learn directly from data, recognize complex patterns, and make predictions without explicit programming for every possible case. In audio signal processing, AI automates feature extraction, classification, enhancement, and synthesis with unprecedented precision. Modern AI architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformers, and Generative Adversarial Networks (GANs) are capable of analyzing vast datasets, discovering subtle features beyond traditional techniques.

Applications of AI in audio are widespread. Speech technologies use AI for automatic speech recognition (ASR), speaker verification, and emotion recognition. In music, AI supports tasks like genre classification, music recommendation, source separation, and even creative composition. Environmental sound recognition assists in urban surveillance, wildlife monitoring, and smart cities. In medicine, AI analyzes biomedical audio signals such as heartbeats and respiratory sounds to provide early, non-invasive diagnostics.

Despite these advancements, challenges persist including limited labeled data, variable recording conditions, background noise, and constraints of real-time processing on limited hardware. Researchers are actively developing robust feature representations, domain adaptation methods, and efficient network designs to overcome these issues.

This paper presents an in-depth study of various audio signal types and their analysis through AI techniques. It discusses prominent feature extraction methods, state-of-the-art AI models, and diverse practical applications, aiming to offer a comprehensive view of AI's transformative impact on audio signal processing and future opportunities.

2. Audio Signal Characteristics & Feature Extraction

Effective AI-driven audio analysis begins with understanding the inherent properties of audio signals. Audio is rich, multidimensional data that

requires careful feature extraction to capture the most relevant information in a compact form suitable for machine learning.

2.1 Types of Audio Signals

- **Speech Signals:** Human vocalizations characterized by linguistic content, pitch, rhythm, and spectral formats.
- **Music Signals:** Structured sounds including melody, harmony, and rhythm, analyzed for tempo, pitch, and timbre.
- **Environmental Sounds:** Unstructured ambient noises such as traffic, machinery, or nature sounds, typically analyzed via temporal and spectral patterns.
- **Biomedical Audio:** Includes physiological sounds like heartbeats, breathing, and coughs, often low amplitude but diagnostically significant.

2.2 Key Signal Characteristics

- **Time-Domain:** Variations of the audio waveform over time.
- **Frequency-Domain:** Distribution of energy across frequencies, obtained via transforms like FFT.
- **Time-Frequency:** Combined representations showing frequency content changes over time (e.g., spectrograms).
- **Perceptual Attributes:** Features related to human auditory perception such as loudness and pitch.

2.3 Feature Extraction Techniques

Time-Domain Features:

- Zero Crossing Rate (ZCR): Counts waveform zero crossings, useful for voiced/unvoiced speech distinction.
- Energy/RMS: Measures signal power, useful for voice activity detection.
- Short-Time Energy (STE): Energy over short windows, aids in endpoint detection.

Frequency-Domain Features:

- Spectral Centroid: Indicates "brightness" of sound by locating spectrum center of mass.
- Spectral Bandwidth: Width of frequency bands holding significant energy.
- Spectral Roll-Off: Frequency below which a majority of spectral energy lies.
- Spectral Flux: Measures spectral change rate over time, helpful for detecting sound onsets.

Cepstral Features:

- Mel-Frequency Cepstral Coefficients (MFCCs): Capture perceptually relevant spectral properties, widely used in speech and audio tasks.
- Delta/Deltas: Capture dynamic changes by computing derivatives of MFCCs.
- Linear Predictive Coding (LPC): Models vocal tract, useful for speech synthesis and analysis.

Time-Frequency Representations:

- Spectrogram: Visualizes frequency content over time, often input to CNNs.
- Log-Mel Spectrogram: Applies Mel-scale filtering and log scaling for perceptual accuracy.
- Chroma Features: Represent pitch classes for music analysis.
- Constant-Q Transform (CQT): Offers frequency resolution aligned with musical scale.

Wavelet-Based Features:

- Discrete Wavelet Transform (DWT): Provides multiresolution analysis in time and frequency, effective for biomedical signals.
- Wavelet Packet Transform (WPT): Extension providing flexible frequency decomposition.

Higher-Level Features:

- OpenSMILE Toolkit: Extracts thousands of acoustic features used for emotion and paralinguistic analysis.
- Learned Embeddings: Deep learning models learn task-specific, discriminative features (e.g., Wav2Vec2, VGGish).

2.4 Feature Selection & Dimensionality Reduction

To avoid over fitting and reduce computational cost, techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), t-SNE/UMAP for visualization, and neural auto encoders are used.

2.5 Tools & Libraries

Popular tools include Librosa, PyAudio Analysis, OpenSMILE, and Essentia, which provide robust feature extraction functionalities.

2.6 Feature Selection Based on Application

Application Domain	Recommended Features
Speech Recognition	MFCCs, Spectrograms, Wav2Vec Embeddings
Emotion Detection	MFCCs, Prosodic Features, OpenSMILE
Environmental Sound Analysis	Log-Mel Spectrogram, Spectral Features
Music Information Retrieval	Chroma, CQT, MFCCs, Tempo, Key
Biomedical Audio	MFCC + DWT, Energy, Temporal Features

3. AI Architectures for Audio Processing

- **Deep Neural Networks:**
  CNNs excel on spectrograms/raw audio for classification; RNNs, LSTMs, GRUs handle temporal dependencies; CRNNs combine both spatial and temporal modeling.
- **Transformer Models:**
  Self-attention architectures (e.g., Wav2Vec2, HuBERT) lead in speech recognition and audio classification.
- **Generative & Autoregressive Models:**
  WaveNet synthesizes realistic waveforms; VAEs and GANs enhance data augmentation and source separation.
- **Neuromorphic & Reinforcement Learning:**
  Spiking neural networks offer efficient classification; reinforcement learning enables adaptive audio control.

4. Applications of AI in Audio

- **Speech Recognition & Speaker/Emotion Identification:**
  From DNNs to transformer-based models, achieving high accuracy in speech and emotion recognition.
- **Environmental Sound Classification:**
  CNNs and transformers classify ambient sounds for smart cities and wildlife monitoring.
- **Source Separation & Remixing:**
  AI isolates audio sources for remixing and restoration.
- **Biomedical Audio Analysis:**
  AI aids diagnosis by analyzing heartbeats, coughs, snoring sounds.
- **Brain-Computer Interfaces & Audio Reconstruction:**
  Emerging research reconstructs audio from brain signals.

5. Challenges & Strategies to Enhance Robustness

- Variability in recording quality tackled by data augmentation, noise-robust features, and domain-invariant embeddings.
- Limited/imbalanced datasets addressed through transfer learning, synthetic data generation, and few-shot learning.
- Temporal/contextual variability mitigated by RNNs, transformers, and context-aware features.
- Cross-domain generalization improved by domain adaptation and diverse training data.
- Real-time constraints handled by model compression, lightweight architectures, and hardware acceleration.
- Model interpretability enhanced via explainability tools and attention mechanisms.
- Label noise managed with robust loss functions and semi-supervised learning.
- Overlapping audio sources addressed through source separation and multimodal learning.

6. Comparative Numerical Analysis of Different Audio Signals

S.No	Audio Signal Type	Application Domain	Feature Complexity (1-5)	AI Model Used	Efficiency (%)	Data Availability (1-5)
1	Speech Signals	Speech Recognition	3	Transformer (e.g., Wav2Vec2)	95–98	5
2	Speaker Identification	Security, Voice Biometrics	3	CNN, Siamese Networks	90–96	4
3	Emotion Detection (Speech)	Affective Computing	4	LSTM, GRU, OpenSMILE	85–95	3
4	Music Classification	Genre/Instrument Recognition	4	CNN, CRNN	88–95	4
5	Environmental Sounds	Surveillance, Smart Cities	3	CNN, PANNs	85–94	4
6	Animal Vocalization	Wildlife Monitoring	3	CNN, Spectrogram-based	80–92	2
7	Machine Sounds	Industrial Fault Detection	2	Auto encoders, CNN	~90	3
8	Cough Sound	Health (e.g., COVID Detection)	3	CNN, Random Forest	85–95	3
9	Snore/Breathing Sounds	Sleep & Respiratory Monitoring	3	CNN, Wavelet + ML Models	85–92	2

Conclusion

AI-driven audio signal processing has matured rapidly, with deep learning architectures and rich feature representations enabling high performance across speech, music, environmental, and biomedical domains. Challenges remain in data scarcity, noise robustness, and domain generalization. Emerging trends like neuromorphic computing, end-to-end waveform modeling, and cross-modal learning are poised to further transform the field.

References

1. Costantini G., Casali D., Cesarini V. (2024). *New Advances in Audio Signal Processing*. Appl. Sci., 14(6), 2321
2. Hajihashemi V., Alavi Gharahbagh A., Hajaboutalebi N. et al. (2024). *Feature-Reduction Scheme for Acoustic Event Detection Systems*. Electronics 2024, 13(11), 2064
3. Thapa N., Lee J. (2024). *Dual-Path Beat Tracking: TCN + Transformer for MIR*. Appl. Sci., 14(24), 11777
4. Venkatesh S., Benilov A., Coleman P. et al. (2024). *Real-time Low-latency Music Source Separation using Hybrid Spectrogram-TasNet*. arXiv, Feb 2024
5. Barusco M., Borsatti F., Dalle Pezze D. et al. (2025). *From Vision to Sound: Advancing Audio Anomaly Detection with Vision-Based Algorithms*. arXiv, Feb 2025.
6. Khaleghpour H., McKinney B. (2025). *Unified AI for Accurate Audio Anomaly Detection*. arXiv, May 2025
7. Basu A., Chaudhari P., Di Caterina G. (2025). *Fundamental Survey on Neuromorphic Based Audio Classification*. arXiv, Feb 2025.
8. Salehi P. et al. (2024). *Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis*. arXiv, Nov 2024
9. Chan R.K., Wang B.X. (2024). *Complementary Acoustic-Phonetic and MFCC Features for Speaker Comparison*. Inf. Fusion 109, 102422
10. Sim J.H. et al. (2024). *Deep Learning Model for Cosmetic Gel Classification using Spectrogram*. ACS Appl. Mater. Interfaces 2024
11. Huang S. et al. (2025). *Survey of DNN Pruning Techniques*. IEEE TPAMI, 46, 10558–…
12. Kotti M., Moschou V., Kotropoulos C. (2024). *Speaker segmentation and clustering methods*. (Referenced through speaker diarization advancements)
13. Wikipedia (2025). *Music Source Separation overview and applications*. (Updated article summarizing recent source separation research)
14. Wikipedia (2023). *Mel-frequency cepstrum (MFCC)*—commonly used in speech/music tasks
15. Wikipedia (2025). *Computational Audiology*—AI applied to hearing diagnostics and monitoring
16. arXiv (2025, Feb). *SIToBI – Speech Prosody Annotation Tool for Indian Languages*
17. arXiv (2025, Feb). *InterGridNet: Audio Source Location Classification via ENF + CNN*. arXiv:2502.10011
18. arXiv (2025, Feb). *MTLM: New Language Model Training Paradigm for ASR*. arXiv:2502.10058
19. Reddit discussion (2025). *Speaker diarization, noise-robust STT, emotion detection remain unsolved areas*
20. Wikipedia (2023–24). *Retrieval-based Voice Conversion (RVC)*: speech-to-speech voice transformation technology.