e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

UNLEASHING TRANSFORMERS FOR AUDIO CLASSIFICATION: A STEP TOWARD INTELLIGENT VOICE SYSTEMS

P. Kamakshi Thai[*1], Premsai Paruchuri[*2], Gudaboina Lalith[*3], Saganti Akhila[*4]

[*1]Assistant Professor, Department Of CSE(AIML), ACE Engineering College, India.

[*2,3,4]Students Of Department Of CSE(AIML), ACE Engineering College, India.

DOI: https://www.doi.org/10.56726/IRJMETS71588

ABSTRACT

Voice recognition technology has transformed our interactions with digital devices, enabling features like virtual assistants, transcription, and security systems. Traditionally, voice recognition relied on hand-engineered features and classical machine learning, which struggled with variations in audio quality and speaker characteristics. Transformer-based models, such as Wav2Vec2, now offer advanced solutions by capturing long-range dependencies in sequential data and learning directly from raw audio. This system utilizes Wav2Vec2 for audio feature extraction and a custom transformer-based classifier to identify singers in audio recordings. Data augmentation techniques enhance the model's robustness across diverse audio conditions. The system achieved high accuracy, highlighting its potential applications in music streaming, copyright management, and audio content organization. This work demonstrates the transformative capabilities of transformers in audio processing and points toward future advancements in voice-based technologies.

Keywords: Voice recognition, Transformers, Wav2Vec2, Audio feature extraction, Audio classification, Data augmentation, Singers Identification, Machine Learning.

I. INTRODUCTION

Voice recognition technology has transformed human-computer interaction, enabling applications like virtual assistants, automated transcription, and security authentication. Initially, classical machine learning methods faced challenges in handling variations in speaker accents, background noise, and recording conditions. Deep learning revolutionized the field by automating feature extraction, with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) improving temporal and spatial audio processing. However, these models struggled with long-range dependencies in sequential data, leading to the adoption of transformer-based architectures. Originally developed for Natural Language Processing (NLP), transformers such as Wav2Vec2 excel in audio recognition by utilizing self-attention mechanisms and self-supervised learning to extract features directly from raw waveforms. These advancements enhance voice recognition across various environments and extend to applications in music streaming, copyright management, and audio content organization, driving innovation in intelligent voice systems.

II. LITERATURE SURVEY

[1] Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations" by Alexei Baevski et al. This paper introduces the Wav2Vec2 model, a self-supervised learning framework for extracting audio features directly from raw waveforms. The model's ability to capture both local and global dependencies in audio data forms the backbone of the proposed project.

[2] "Attention Is All You Need" by Vaswani et al. The seminal paper on transformer architecture that laid the foundation for models like Wav2Vec2. It introduces the self-attention mechanism, enabling transformers to effectively process sequential data and capture long-range dependencies

[3] "Deep Speech: Scaling up End-to-End Speech Recognition" by Hannun et al. This paper demonstrates the potential of deep learning in replacing traditional voice recognition pipelines with end-to-end systems. It highlights the limitations of RNNs, which transformers aim to overcome in this project.

[4] "Audiomentations : A Python Library for Audio Data Augmentation" by Kevin Francksen et. This paper emphasizes the importance of data augmentation techniques in improving model generalization across diverse audio environments. The strategies discussed here are integral to enhancing the robustness of the proposed system.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

[5] "Transfer Learning with Pretrained Audio Models" by Kong et al. This paper explores how pre-trained models like Wav2Vec2 can be fine-tuned for specific audio classification tasks. It underscores the effectiveness of transfer learning in achieving high accuracy with limited labeled data.

2.1 Existing Model

Current voice recognition systems employ various methodologies, primarily leveraging feature engineering techniques such as Mel-frequency cepstral coefficients (MFCCs) and linear predictive coding (LPC). Commercial systems generally offer higher accuracy but come with costs and potential proprietary limitations. Open-source alternatives provide flexibility but might compromise on accuracy and robustness. Both types of systems face difficulties in handling complex audio scenarios and extracting meaningful information from raw audio data.

2.2 Challenges in Existing System:

These methods, while effective to a degree, often require manual tuning and exhibit significant limitations in handling dialectal variations and diverse vocal characteristics. Traditional systems, like those utilizing HMMs and GMMs, face challenges in complex audio environments with background noise and overlapping speech, leading to lower accuracy. Both types of systems face difficulties in handling complex audio scenarios and extracting meaningful information from raw audio data.

III. PROPOSED METHODOLOGY

Our voice recognition system leverages transformer technology's self-attention mechanism to enhance feature extraction directly from raw audio signals. Unlike traditional methods that rely on manual feature engineering, our approach automates this process, improving efficiency and accuracy. The system is built on transformer-based architecture specifically designed for audio processing, incorporating self-attention layers to capture intricate vocal characteristics and context dependencies. By focusing on accurate classification and categorization rather than just denoising, our model ensures a more precise interpretation of voice inputs across various conditions.

3.1 Advantages of Proposed System

The proposed system offers significant advantages over traditional voice recognition methods, particularly in handling diverse and noisy environments. By eliminating manual feature engineering, it enhances scalability and adaptability, allowing for more robust audio processing. Evaluations on standard benchmarks demonstrate superior accuracy and computational efficiency compared to existing systems, highlighting the effectiveness of transformer-based models in voice recognition. This innovation not only improves feature representation but also addresses key limitations of earlier approaches, making it a more reliable and scalable solution for modern audio applications.

IV. ALGORITHM

Step-1: Load Pre-trained Model: A pre-trained model like facebook/wav2vec2-base is loaded using Hugging Face Transformers to leverage existing audio feature extraction capabilities.

Step-2: Fine-tune the Model: The model is trained on labelled singer data, updating its weights while keeping earlier layers frozen to retain general audio features.

Step-3: Prepare Labelled Data: Audio clips are collected, labeled, and pre-processed using waveform conversion, normalization, and length adjustments for consistency.

Step-4: Apply Data Augmentation: Techniques like Gaussian noise, pitch shifting, and time stretching are applied to enhance model robustness and generalization.

Step-5: Train the Model: The dataset is split into training, validation, and test sets. The model learns singer classification using Cross-Entropy Loss and an optimizer like AdamW.

Step-6: Evaluate the Model: Performance is measured using accuracy and F1 score, ensuring balanced evaluation and identifying areas for improvement.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

V. DATA FLOW DIAGRAM

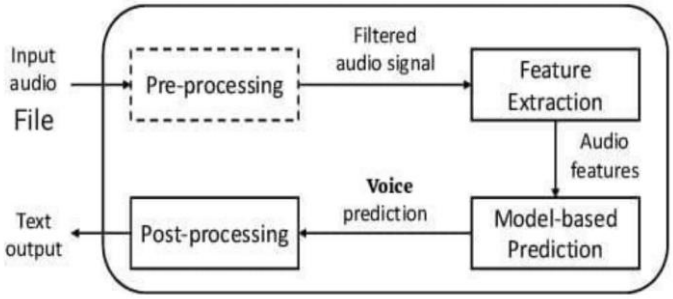

Figure: Data Flow Diagram

Audio Input: The system starts by ingesting raw audio files from a specified source.

Data Preprocessing: The raw audio undergoes resampling, normalization, and trimming/padding to create uniform input data.

Feature Extraction: The preprocessed audio is fed into the Wav2Vec2 model to extract meaningful features.

Classification: The extracted features are input into the transformer-based classification model, which outputs the predicted singer classes.

Post-Processing: The classification results are processed and converted into user-friendly formats, such as textual labels indicating the identified singer.

VI. CONCLUSION

This work demonstrates the potential of transformer-based models like Wav2Vec2 in audio classification, particularly for singer identification. By integrating pre-trained models, automated feature extraction, and advanced data augmentation, the approach enhances accuracy and robustness. Its adaptability allows for broader applications in audio recognition. Future improvements may involve expanding the dataset, optimizing real-time processing, and exploring alternative deep-learning architectures to further refine performance and efficiency.

VII. REFERENCES

[1] “Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations" by Alexei Baevski et al.

[2] “Attention Is All You Need" by Vaswani et al.

[3] “Deep Speech: Scaling up End-to-End Speech Recognition" by Hannun et al.

[4] “Audiomentations: A Python Library for Audio Data Augmentation" by Kevin Francksen et.

[5] “Transfer Learning with Pretrained Audio Models" by Kong et al.