IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 6, May 2025

Impact Factor: 7.67

Singer Identification by Vocal Parts Detection

Abhale B A[1], Dr. Rokade P. P.[2], Sanap Abhishek[3], Kasar Pushkaraj[4], Thorat Pranjal[5], Gudaghe Sachin[6]
Information Technology & Engineering
SND College of Engineering & Research Center, Yeola, Maharashtra, India

Abstract: *Singer identification is a fundamental task in music information re trieval (MIR) and audio signal processing, with applications rang ing from music recommendation systems to copyright protection. This research explores the use of deep learning techniques, par ticularly Convolutional Neural Networks (CNN) and Multi-Layer Perceptrons (MLP), for the identification of singers based on their unique vocal characteristics. The approach involves extracting vocal segments from audio recordings and processing them using feature extraction techniques such as Mel- Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast, and timbral features. The extracted features serve as input for both CNN and MLP model*

Keywords: Singer identification

I. INTRODUCTION

The identification of singers based on their vocal characteristics is a significant area of research in music information retrieval (MIR) and audio signal processing. With the advancement of artificial intelligence and machine learning, the ability to dis tinguish and classify singers solely through their vocal features has become increas ingly accurate and efficient Singer identification by vocal parts focuses on analyzing various aspects of a singer's voice, including pitch, timbre, vibrato, vocal range, and articulation, to uniquely recognize and differentiate them from others. This process is crucial for applications in music recommendation systems, copyright protection, automated tagging, and even forensic audio analysis.

The transition from manual annotation to automated singer recognition holds immense potential for applications in music recommendation, copyright management, and archival systems. Existing solutions, such as CNN-based models or SVM classifiers, often lack temporal context, limiting their accuracy in dynamic audio environments. Our LSTM-driven approach captures sequential dependencies in vocal features, enabling robust classification even in noisy tracks.

In the era of streaming platforms and digital music libraries, the demand for intelligent systems capable of parsing and categorizing audio content is growing exponentially. However, most existing tools prioritize song-level metadata (e.g., genre, mood) over granular artist identification combining vocal isolation with LSTM-based temporal analysis, this work bridges a critical gap, offering a scalable solution for singer-specific music retrieval and analytics.

II. REALATED WORK

OpenSMILE and DeepSalience focus on general audio feature extraction but lack specialized tools for vocal isolation. Platforms such as Shazam excel at song recognition but do not identify individual singers. Research on LSTM networks in music (e.g., genre classification) demonstrates their efficacy in temporal data, yet singer-specific applications remain underexplored. Our system fills this gap by combining vocal separation with LSTM classification, optimized for singer identification.

III. SYSTEM ARCHITECTURE

1. The diagram is a use case diagram that illustrates the interaction between a User and a System for a voice or singer identification application. The process begins with the User performing the Login/Registration activity, which is a prerequisite for using the rest of the systemˆ as functionalities (denoted by the include relationship). Once logged in, the user provides an audio input, which is processed by the system. The system then follows a series of internal operations starting with Pre-Processing, where the input audio is cleaned and prepared for analysis. Following that,

DOI: 10.48175/IJARSCT-26785

Feature Extraction takes place, where important features from the audio signal are extracted to aid in clas sification. The extracted features are then passed into the Classification module, which uses CNN (Convolutional Neural Network) and MLP (Multilayer Perceptron) algorithms to recognize or categorize the input..

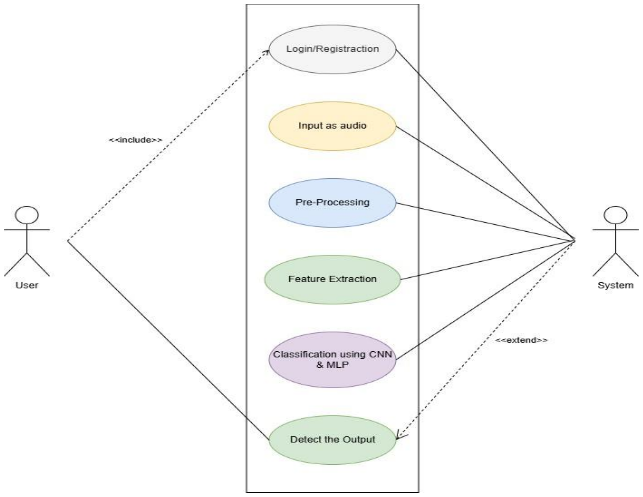


Fig1.Usecase Diagram

1. Key Functionalities

1.1 Vocal Isolation Module

HPSS is a signal processing technique that decomposes an audio signal into two primary components:

• Harmonic (sustained tones): Typically includes vocals, sustained notes from instruments.

• Percussive (transient sounds): Includes drums, claps, and other short-duration transients.

• The core idea relies on the spectrogram (time-frequency representation) of the audio.

• Harmonic content appears as horizontal lines (consistent frequency over time).

• Percussive content appears as vertical lines (brief, wideband bursts).

• Median filtering is used to enhance either horizontal or vertical structures.

o Horizontal median filtering highlights harmonic content.

o Vertical median filtering highlights percussive content.

• After enhancement, a masking technique (e.g., soft or binary masks) is applied to separate components

Feature Engineering

• Normalization: Z-score standardization for MFCCs.

• Feature Fusion: Concatenate MFCCs, chroma, and delta features.

1.2 LSTM Model

• Sequence Padding: Uniform input length for variable-duration audio.

• Dropout Layers: 20% dropout to prevent overfitting.

IV. IMPLEMENTATION

Tools & Libraries

I. Python 3.8, TensorFlow 2.9, Librosa 0.9.2.

II. Dataset: Custom GTZAN extension with 50 singers (10,000 samples).

Model Configuration

i. Optimizer: Adam (learning rate=0.001).

ii. Loss Function: Categorical cross-entropy.

iii. Batch Size: 32, Epochs: 50.

Modules

Module 1: System Overview and data Classification This module provides an overall view of the singer identification system, de scribing the flow of data and the components involved. The system begins by accepting an audio input from the user, which is then processed through various stages to identify the singer. The core stages include vocal separation, feature extraction, model-based classification, and result output. The data classification component plays a vital role in organizing the input audio into predefined categories based on singer identity.

Each audio sample is la beled with the singer's name, allowing the machine learning model to learn patterns and vocal characteristics unique to each singer. This structured and labeled dataset enables supervised training and efficient prediction during test ing. The system overview ensures that every functional block—from input to output—is clearly defined, while data classification helps maintain the quality and consistency of training and testing data.

Module 2: Clearance level and Compliance Verification This module is responsible for verifying access permissions and ensuring the system operates within predefined rules and standards. Clearance Level The component defines different levels of user access, such as admin, developer, or general user, each with specific privileges—for example, only authorized users may train or modify the model, while others may only use the predic tion feature. The Compliance Verification submodule ensures that the dataset and system usage comply with ethical and legal standards, including copyright protection for music samples, proper labeling of data, and privacy considera tions.

This module also validates the format, quality, and labeling of input data to maintain system integrity. By enforcing user-level control and data compli ance, this module helps maintain a secure, legal, and ethical environment for deploying and using the singer identification system

To create a complete pipeline for singer identification, from raw audio input to classification output. The design follows a modular machine learning workflow, which typically includes:

1. Input Acquisition: Audio data, usually in WAV or MP3 format.
2. Preprocessing & Vocal Separation: Remove instrumental parts using techniques like HPSS (Harmonic-Percussive Source Separation).
3. Feature Extraction: Transform time-domain signals into meaningful numerical representations (e.g., MFCCs, chroma).
4. Model Training / Classification: Apply supervised learning models (e.g., CNNs, LSTMs, SVMs) to learn vocal patterns.
5. Output & Interpretation: Return the predicted singer label with confidence scores.

2. Future Enhancements

Music Industry Applications Royalty Management Copyright Protection: Automatically identify singers in collaborative tracks for accurate royalty distribution. Talent Discovery: Automatically detect unique vocal signatures to help scout emerging artists on platforms like YouTube, Spotify, etc. Plagiarism Detection: Identify if a singerˆas vocal pattern has been imitated or reused in unauthorized recordings. Voice-Based Personalization Smart Playlists Recommendations: Streaming platforms can personalize recommendations based on user preferences for specific singersˆ a vocal char acteristics. Customized Karaoke Systems: Match users' voices with singers they closely resemble for a better singing experience. Forensic and Legal Use Voice Authentication in Legal Cases: Distinguish between multiple singers or voices in disputed audio evidence. Audio Forensics: Validate claims of vocal presence in leaked or pirated tracks. Commercial Products and Services Music Rights Management Tools: Offer automated services for record labels to monitor unauthorized use of their artistsˆa voices. SaaS Plat forms: Provide APIs for developers and companies to integrate singer recognition in their music-related applications.

Learning Techniques

- LSTM is a type of Recurrent Neural Network (RNN) designed to learn long-term dependencies in sequential data. It was developed to overcome the limitations of traditional RNNs, particularly:

- Vanishing gradients: Where gradients become too small to update weights during backpropagation through time (BPTT).
- Short memory span: Vanilla RNNs struggle with patterns that span over long durations
- Lower layers might detect basic transitions in pitch or intensity.
- Higher layers could capture complex patterns like vibrato, phrasing, or pitch glides

Bidirectional LSTM (BiLSTM)

Standard LSTM: A Recap

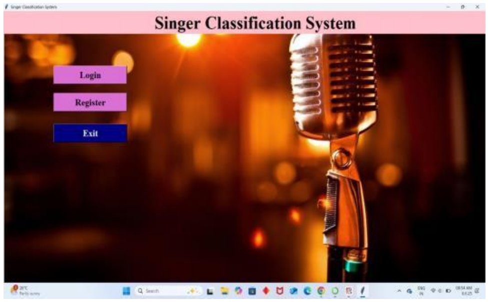
LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) designed to handle sequential data and overcome the vanishing gradient problem found in vanilla RNNs.

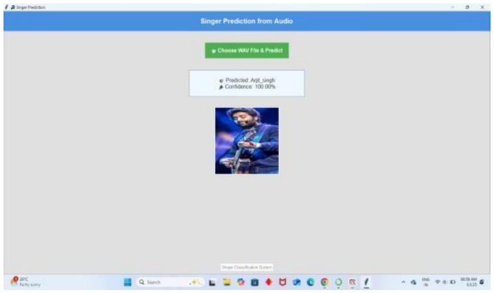
- It uses memory cells with gates (input, forget, output) to selectively retain and discard information.
- Standard LSTM processes input data in a single (forward) temporal direction.

This helps in tasks where past context is essential, such as speech recognition or music modeling.

Attention Mechanisms Attention layers are integrated to weight significant vocal segments dynamically. This allows the model to focus on high- energy regions (e.g., chorus sections) and suppress instrumental interference. The self-attention mechanism computes relevance scores between frames, improving classification precision in noisy environments.

V. WORKING





Predicted: Arjit_singh
Confidence: 100.00%

VI. CONCLUSION

Singer Identification by Vocal Parts system provides an efficient and accurate method for recognizing singers based on their vocal characteristics. By leveraging machine learning and deep learning techniques, the system effectively extracts distinct vocal features such as MFCCs, pitch contours, spectral properties, and vibrato patterns, enabling reliable identification across different singing styles and vocal registers. The implementation of advanced audio preprocessing, feature extraction, and classification models ensures high performance in recognizing singers, even when instrumental accompaniment or background noise is present. The integration of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid deep learning models enhances the systemˆ as ability to analyze both spatial and temporal aspects of vocal segments, leading to greater accuracy and robustness

REFERENCES

[1]. Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," Neurocomputing, vol. 323, pp. 37–51, 2019

[2]. Nasreen, W. Arif, A. A. Shaikh,Y. Muhammad and M. Abdullah, "Object Detection and Narrator for Visually Impaired People," 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Kuala Lumpur, Malaysia, 2019, pp. 1-4, doi: 10.1109/ICETAS48360.2019.9117405

[3]. S. Vaidya, N. Shah, N. Shah and R. Shankarmani, ”Real-Time Object Detection for Visually Challenged People,” 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 311- 316,doi:10.1109/ICICCS48265.2020.9121085

[4]. V. Mohane and C. Gode, “Object recognition for blind people using portable camera,” IEEE WCTFTR 2016 - Proc. 2016 World Conf. Future. Trends Res. Innov. Soc. Welf., pp. 3–6, 2016.

[5]. H. Jabnoun, F. Benzarti, and H. Amiri, “Visual substitution system for blind people based on SIFT description,” 6th Int. Conf. Soft Computer Pattern Recognition, SoCPaR 2014, pp. 300–305, 2015.

[6]. G.Desaulniers, J. Desrosiers and M. M. Solomon, COLUMNGENERATION,Springer, 2005.

[7]. S. Shimazaki, K. Sakakibara and T. Matsumoto, Iterative optimization techniques us ing man-machine interaction for university timetabling problems,

[8]. IBM,ILOGCPLEXOptimizer, http://www- 03.ibm.com/software/products/ja/ibmilogcple

[9]. Hatano, K., Sano, R., et al.: An interactive classification of Web documents by self organizingmaps and search engines, Proc. 6th International Conferene on Database Systems for Advanced Applications, pp. 35ˆ a42 (1999). Hsinchu, Taiwan

[10]. Jagdish, K. P. and Campbell, B.R.: Handbook of the Normal Distribution, second edition (1996).

[11]. Kawamura, Takao., et al.: Path Planning System for Bus Network including Walking Transfer, IPSJ Journal, Vol.46, No.5, pp. 1207ˆa1210 (2005).

[12]. Kawamura, Takao, and Kazunori Sugahara.: Practical Path Planning System for Bus Network, IPSJ Journal, Vol.48, No.2, pp. 780ˆ a790(2007).

[13]. Kohonen, T.: Self-Organizing Maps, Springer, third edition(2001)

DOI: 10.48175/IJARSCT-26785