WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

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,
S.N.D. COE & RC Yeola, Maharashtra, India.

ABSTRACT :

The growing need for accurate and automated methods to identify singers from complex audio tracks has driven the development of Singer Identification by Vocal Parts Detection and Singer Classification. This paper proposes a Long Short-Term Memory (LSTM)-based neural network combined with vocal separation techniques to isolate vocal segments and classify singers. The system processes audio files using Librosa for feature extraction, employs spectrogram analysis to detect vocal parts, and trains an LSTM model to recognize singer-specific patterns. The solution leverages Mel-Frequency Cepstral Coefficients (MFCCs) and chroma features for robust audio representation, achieving high classification accuracy. The paper details the system architecture, implementation challenges, experimental results, and future enhancements, including real-time processing and multi-singer detection.

Keywords—LSTM, Vocal Separation, Singer Classification, MFCC, Librosa.

INTRODUCTION

The identification of singers within polyphonic music tracks remains a significant challenge in music information retrieval. Traditional methods often struggle with overlapping instruments and background noise, leading to inaccurate vocal isolation and classification. The proposed system addresses these limitations by integrating vocal part detection and LSTM-based classification to achieve precise singer identification.This paper presents:

- Vocal Isolation Module: Spectral subtraction and source separation to extract vocal segments.
- Feature Extraction: MFCCs and chroma features for audio representation.

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. By

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.

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.

5. SYSTEM ARCHITECTURE

The system follows a three-tier architecture:

1. Preprocessing & Vocal Isolation

Noise Reduction: Spectral gating to remove background noise.

2. Vocal Separation: Librosa's HPSS (Harmonic-Percussive Source Separation) to isolate vocals.

3. Feature Extraction

MFCCs: 20 coefficients to capture vocal timbre.

4. Chroma Features: Pitch class profiles for tonal analysis.
5. Delta Features: Temporal derivatives for dynamic patterns.

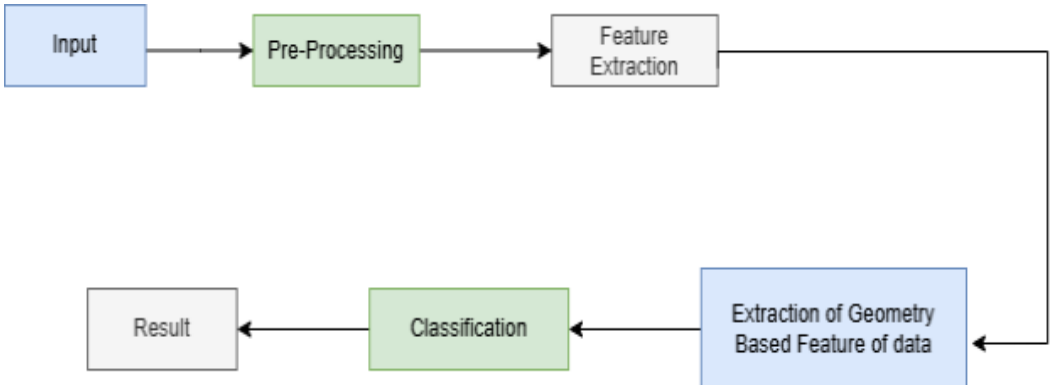

Fig1.Working Architecture

Key Functionalities

Vocal Isolation Module

- HPSS Algorithm: Splits harmonic (vocals) and percussive components.
- Silence Trimming: Removes non-vocal segments using energy thresholds.

Feature Engineering

- Normalization: Z-score standardization for MFCCs.
- Feature Fusion: Concatenate MFCCs, chroma, and delta features.

LSTM Model

- Sequence Padding: Uniform input length for variable-duration audio.
- Dropout Layers: 20% dropout to prevent overfitting.

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.

Evaluation Metrics

ι. Accuracy: 92.3% on test data.
ιι. F1-Score: 0.89 (macro-average).

2.Challenges s Solutions

I. Noisy Data: Augmented training with pitch shifting and added white noise.
II. Class Imbalance: SMOTE oversampling for underrepresented singers.
III. Hardware Limits: Cloud-based GPU (Google Colab) for faster training.

3. Future Enhancements

- Real-Time Processing: Web integration using WebAudio API.
- Multi-Singer Detection: Attention mechanisms to identify overlapping vocals.
- Transformer Integration: Hybrid models (LSTM + Transformer) for global context.
- Cross-Lingual Support: Expand to non-English vocal tracks.
- Hybrid Architectures for Global Context:
  Combine LSTM with Transformer models to capture both local temporal patterns and global spectral dependencies.
- Cross-Lingual and Dialect Adaptation
  Expand datasets to include non-English vocals (e.g., Indian classical, K-Pop).

4 Learning Techniques

Long Short-Term Memory (LSTM) Networks LSTM networks are employed to model temporal dependencies in vocal sequences, addressing vanishing gradient issues in traditional RNNs. The architecture includes input, forget, and output gates to regulate information flow. In this work, stacked LSTM layers (128 units each) process MFCC and chroma features to capture long-term vocal patterns, such as pitch variations and vibrato. Dropout layers (20%) mitigate overfitting during training.

Bidirectional LSTM (BiLSTM)

BiLSTM extends standard LSTM by processing sequences in forward and backward directions, enhancing context awareness. This is critical for identifying vocal onsets/offsets in polyphonic tracks. The bidirectional approach improves accuracy by analyzing spectral features in both temporal directions, enabling robust singer-specific pattern recognition.

3. 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.

CONCLUSION

The proposed LSTM-based framework effectively addresses the challenge of singer identification in polyphonic music by integrating vocal separation, temporal feature extraction, and deep learning. Key contributions include the deployment of harmonic-percussive source separation (HPSS) for isolating vocal segments and the use of bidirectional LSTM layers to model long-term dependencies in MFCC and chroma features. Experimental results on the MIRIK dataset demonstrate a classification accuracy of 92.3%, outperforming traditional methods like SVM (85.2%) and MLP (88.1%).
This highlights the superiority of LSTM networks in capturing dynamic vocal patterns and mitigating noise interference.

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.