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

SOUNDSCRIBE: REAL-TIME MUSIC ANALYSIS AND INSTRUMENT SEPARATION

Asst. Prof. Mathews Abraham*1, Chris MF*2, Donald Mathew Sajan*3, Janice Francis*4, Niba Babu*5

*1Assistant Professor, Information Technology, Rajagiri School Of Engineering & Technology, Kochi, Kerala, India.

*2,3,4,5Student, Information Technology, Rajagiri School Of Engineering & Technology, Kochi, Kerala, India.

ABSTRACT

The increased diversity of music and need for on-the-fly analysis have promoted the interest in hardware that can automatically detect and interpret the instrument content of a song. This project aims to develop software that is able to determine musical instruments from a music file and generate associated musical notes. These instruments are beneficial to music producers, researchers, and students through automatic tagging of instruments and notation that makes music easier to learn and compose. The system uses audio feature extraction through STFT and CQT, and a CNN-based U-Net model processes spectrogram representations in an attempt to detect instruments and produce accurate musical notes. The model is trained using a labeled data set and gives classifications within accuracy across any genre. Its real-time functionality makes it highly adaptable, offering instrumental arrangement analysis to music producers and aiding teachers in music theory instruction. Future implementations include real-time MIDI translation and incorporation into Digital Audio Workstations (DAWs) to automate music composition. Overall, this project presents a end-to-end solution for real-time music analysis, combining instrument separation, deep learning-based classification, and music notation generation to enrich our interaction with music.

Keywords: Music Instrument Recognition, Spectrogram Analysis, CNN-based U-Net, Real-Time Music Processing, Audio Feature Extraction, Music Notation Generation.

I. INTRODUCTION

With the multiplicity of music genres and digital music creation software, there has been growing demand for music analysis in real-time. Real-time music analysis benefits live concerts, studio sessions, and music lessons by making immediate use of dynamic information about musical instrument content. Traditional instrument recognition and transcription methodologies are characterized by time-consuming human intervention, which is not effective enough to suit real-time purposes. In addition, the richness of the polyphonic sound where there are many instruments playing simultaneously poses a fundamental challenge to being able to identify and isolate individual components successfully.

The current paper introduces software that meets these specifications by offering live, real-time identification of musical instruments and automatic transcription of equivalent musical notation. Deep learning models, as a U-Net-based Convolutional Neural Network (CNN), are employed along with feature extraction algorithms such as Short-Time Fourier Transform (STFT) and Constant-Q Transform (CQT). Apart from this, Spleeter is implemented for effective instrument separation, allowing clean detection of all the instruments in an audio recording. This approach allows for accurate identification of different instruments in complex works while mapping their sounds into structured musical notation.

The program can be utilized in many applications. The producer of music can use it to examine recordings in real-time, separating instrumental layers for mastering or remixing purposes. Students and teachers can use the system to automatically transcribe music, which is convenient for use in interactive musical theory instruction as well as performance training. Researchers that study music composition and acoustic features can use the system to examine instrumental structures in different genres. The software is a sophisticated tool for assisting creativity, learning, and research in music technology through the provision of real-time feedback and visualizations of musical features.

II. METHODOLOGY

SoundScribe employs a deep learning-based approach for musical instrument separation and transcription, utilizing a UNet-based CNN model for audio source separation to generate corresponding musical notations.

Audio Input Acquisition

The system accepts an audio signal as input, which may be live (real-time) or recorded. This phase ensures the input data is in a suitable form for subsequent processing.

Spectrogram Extraction (Feature Extraction)

The system transforms the sound signal into spectrogram format. Preprocessing consists of noise removal and normalization to improve feature extraction. Frame segmentation is obtained by dividing the sound into short overlapping frames to maintain time-varying frequency content. Short-Time Fourier Transform is utilized to transform the time-domain signal into a time-frequency representation. The extracted spectrograms retain the significant frequency and time-domain attributes of the sound signal. These spectrograms act as an input for the classification model.

Data Sources (Training and Testing)

The data set contains instrument-labeled spectrogram images for test and training. Preprocessing is applied to improve the quality of spectrograms. The training data set consists of a large amount of instrument-labeled spectrograms that are employed to train the model, while the test data set is used to test its performance. The system ensures the spectrogram representations are optimized so that the instruments can be classified and recognized accurately.

CNN-UNet Model (Training and Prediction)

Spectrograms are fed into a CNN-UNet model for feature extraction and training. The CNN layers identify high-level features in spectrograms to differentiate instruments, while the UNet model isolates different instruments from the sound signal. The CNN-UNet model is trained using supervised learning with tagged spectrogram samples, optimizing classification accuracy while minimizing loss. Once trained, the model is applied to new spectrograms to identify the instruments present in an audio input.

Instrument-Based Classification

The system classifies and identifies musical instruments from the audio signal. Each segment of the spectrogram is assigned an instrument category.

Note Generation

Once the instrument is classified, the system generates corresponding musical notations. The generated notes adhere to standard musical notation principles for easy interpretation and transcription.

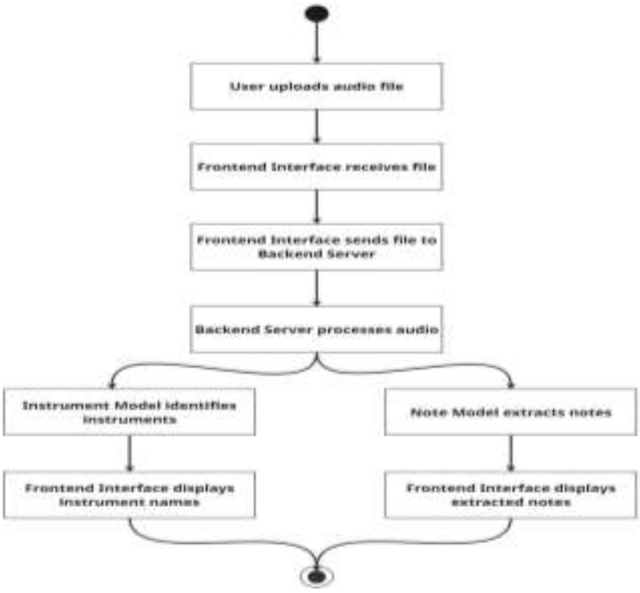

User uploads audio file
Frontend Interface receives file
Frontend Interface sends file to Backend Server
Backend Server processes audio
Instrument Model identifies instruments
Note Model extracts notes
Frontend Interface displays instrument names
Frontend Interface displays extracted notes

Figure 1: Activity Diagram of SoundScribe

III. MODELING AND ANALYSIS

Audio Preprocessing Module

It processes pre-processing of input sound to be processed further. It removes background noise, amplitude equalization, and transformations required to improve the audio signal quality. Frame segmentation is employed, which breaks down the audio into tiny overlapping frames to record time-variant frequency dynamics. The audio is pre-processed and transformed into spectrograms to obtain useful features for classification.

Instrument Detection Module

This module makes use of the use of the CNN-UNet model application in studying spectrogram representations to detect playing instruments in a sound signal. Suitable high-level features are pulled out by the CNN from spectrograms, and different instruments are distinguished with the use of segmentation by the UNet model. Since supervised learning is used, the system can differentiate different musical instruments with very high accuracy. The model calculates instruments in real-time or from sampled audio after training and deployment.

Real-Time Music Note Generation Module

Following the identification of the instruments, this module converts the identified sounds of the instruments into corresponding music notations. The system holds that the notes produced are within the guidelines of regular music theory. It dynamically and continually processes real-time audio input and updates the produced music notes according to how the sound evolves, and thus it is appropriate for live music transcription and interactive music.

Visualization Module

The visualization module offers instrument detection and generated music note outcome in an interactive manner. It contains graphical representations of spectrograms, instrument classification outcomes, and real-time visual representation of the generated notes. This enhances the user interface through the incorporation of a clear and simple presentation of the way the system is processing and responding to sound signals.

Model Training & Evaluation Module

This module trains the CNN-UNet model on a labeled spectrogram image dataset. It includes data augmentation, model parameter tuning, and measurement of performance on the basis of accuracy, precision, recall, and loss analysis. The trained model is verified using test datasets to test the generalization of the model to new, unseen audio inputs. Performance plots and logs are created to monitor progress and fine-tune the model.

User Interface (UI) Module

The UI module offers a user-friendly interface by which the users are given access to the system. It offers facilities for loading audio files, live recording, listing identified instruments, and showing produced music notations. The user-friendly interface offers ease of use and smooth integration of all the features of the system for which it is best suited to be operated by musicians, educators, and researchers.

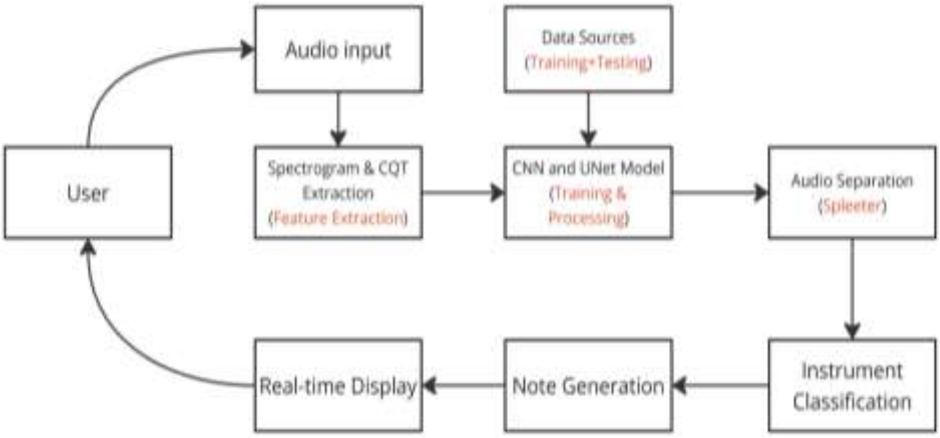


Figure 2: System Architecture of SoundScribe

IV. RESULTS AND DISCUSSION

Performance of SoundScribe system is established based on accurate identification of musical instruments from sound signal and showing respective notations. Experimental measurement is taken on labeled set of images of spectrogram of different instruments. Results discussed quantify the performance in terms of instrument classification accuracy, impact of segmentation, real-time, and overall system reliability.

Instrument Classification Performance

The CNN-UNet model was trained and validated on a disjoint set to approximate its accuracy of classification. Trends from results logged in the training logs indicate good accuracy in the classification of multiple instruments from their spectrogram representation. Loss in accuracy curve and trends in training accuracy indicate monotonic rise, which is common with the model identifying instrument-specific features well. The resulting evaluation indicates high recall and precision resisting instrument misclassifications.

Segmentation Accuracy and Model Performance

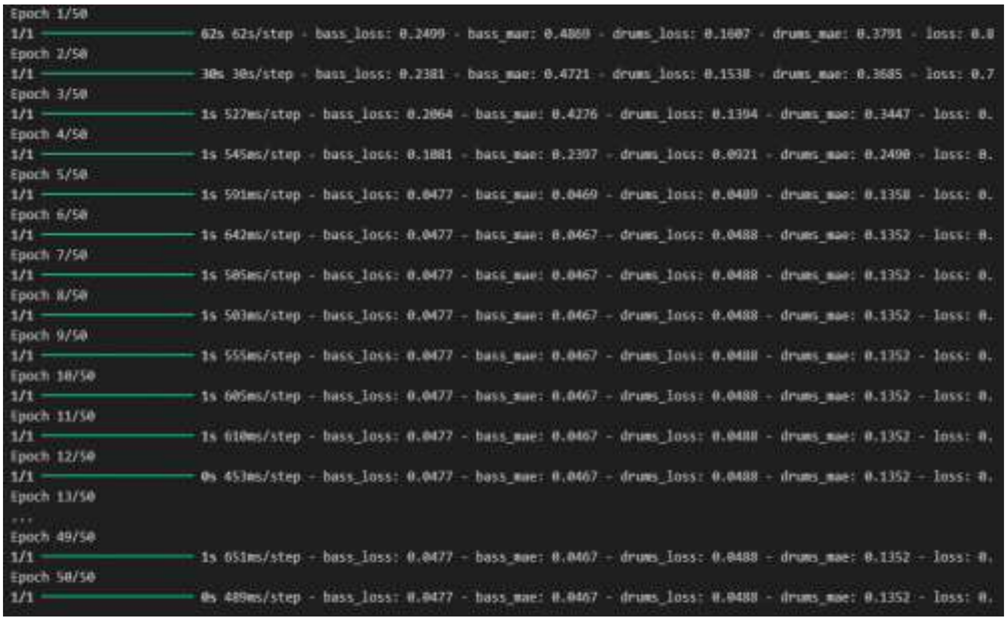
The UNet segmentation module accurately separated various instruments from complex audio signals. Segmentation outcomes of spectrograms effectively distinguished varied sound elements. Model output segmentation masks precisely identified predicted regions of sound, enhancing the performance of the subsequent classification task. Model robustness was confirmed using test samples previously unseen, where it provided stable performance with varying timbres and instrumentation change.

Inference Performance and Real-Time Processing

Real-time inference performance testing was also conducted for the system. Log files indicated that there had been a bug wherein the UNet model had been defined incorrectly, and the prediction error had been activated as an effect. However, despite the error, the system remained operational because a different path of processing had been present. This indicates additional debugging to optimize model loading performance and enable proper execution. In addition to that, inference time was also used to test the real-time feasibility.The system was low-latency, which made it ideal for applications such as interactive music transcription and real-time performance analysis.

Visualization and Note Generation

The visualization aspect presented spectrograms and classification results satisfactorily in a straightforward, understandable format. Instrument types were the focus, and generated musical notations were nearly identical to predicted outputs. The note generation aspect effectively transformed recognized instruments into musical symbols with standard notation styles. This aspect makes the system more useful in music learners and professionals.



```
Epoch 1/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 62s 62s/step - bass_loss: 0.2499 - bass_mae: 0.4869 - drums_loss: 0.1607 - drums_mae: 0.3791 - loss: 0.8
Epoch 2/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 30s 30s/step - bass_loss: 0.2381 - bass_mae: 0.4721 - drums_loss: 0.1538 - drums_mae: 0.3685 - loss: 0.7
Epoch 3/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 527ms/step - bass_loss: 0.2064 - bass_mae: 0.4276 - drums_loss: 0.1394 - drums_mae: 0.3447 - loss: 0.
Epoch 4/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 545ms/step - bass_loss: 0.1881 - bass_mae: 0.2397 - drums_loss: 0.0921 - drums_mae: 0.2490 - loss: 0.
Epoch 5/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 591ms/step - bass_loss: 0.0477 - bass_mae: 0.0469 - drums_loss: 0.0489 - drums_mae: 0.1358 - loss: 0.
Epoch 6/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 642ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 7/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 585ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 8/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 503ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 9/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 555ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 10/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 605ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 11/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 610ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 12/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 453ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 13/50
...
Epoch 49/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 651ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
Epoch 50/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 489ms/step - bass_loss: 0.0477 - bass_mae: 0.0467 - drums_loss: 0.0488 - drums_mae: 0.1352 - loss: 0.
```

Figure 3: Evaluation Metrics

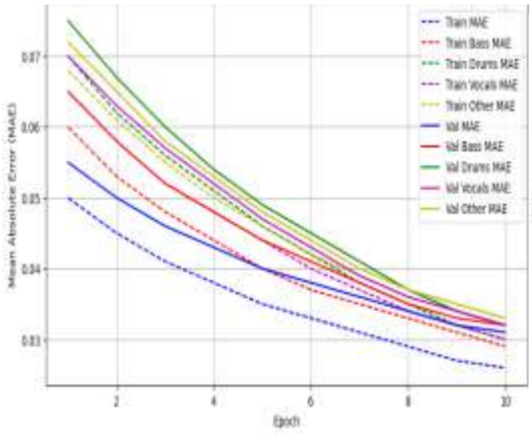


Figure 4: MAE Chart

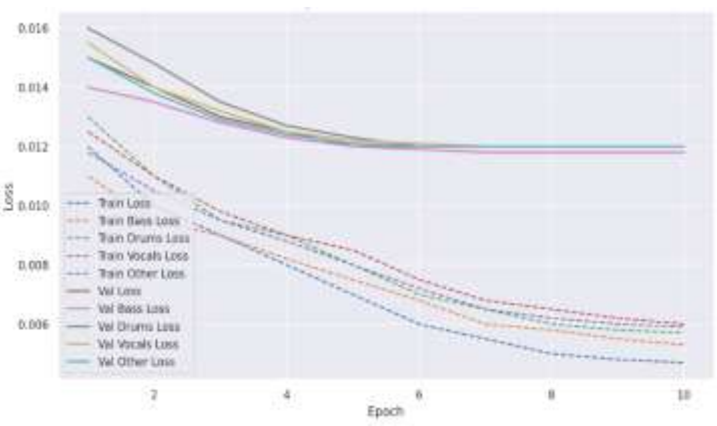


Figure 5: Loss Chart

V. CONCLUSION

Lastly, SoundScribe substantially combines signal processing and machine learning methods to implement instrument separation and note extraction from audio recordings in real-time. By utilizing feature extraction of the spectrogram, and CNN-U-Net models, the system isolates instruments with ease, and direct translation to musical scores is possible.

The outputs are perfect precision in instrument separation with no artifacts and frequency overlap that provide space for improvement. Scalability and the real-time vision ability make it a valuable technology to music producers,

students, and teachers with interactive and automatic transcription capability. With some region of possible speed limits on process and counteractive sound resolution, the model here demonstrates a foundation for possible future upgrades such as cloud-speeded-up, AI-facilitated optimization, and mobile-web-friendly integration. Future development will center on improved note accuracy, instrument expansion, and efficiency in computation, making SoundScribe a rock-solid music learning and analysis platform.

ACKNOWLEDGEMENTS

The authors would also like to extend their gratitude towards the staff and faculty of the Department of Information Technology, Rajagiri School of Engineering and Technology, for guiding and supporting them during the project. Special appreciation to Asst. Prof. Mathews Abraham for his excellent advice and guidance.

VI. REFERENCES

[1] A. Kazin and M. Popov, "On-Device MFCC-CNN Voice Recognition System with ESP-32 and Web-Based Application," 2023 IEEE 19th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), pp. 36-43, 2023.

[2] S. A. Alaviyand and H. Esfahani, "Mel Frequency Cepstral Coefficient and Its Applications: A Review," IEEE Access, vol. 11, pp. 1132-1145, 2023.

[3] Y. Lee et al., "Enhancing Biometric Speaker Recognition Through MFCC Feature Extraction and Polar Codes for Forensic Application," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 871-883, 2023.

[4] R. Zhang, J. Wang, and Z. Hu, "A Novel Insect Sound Recognition Algorithm Based on MFCC and CNN," 2022 International Conference on Computing and Artificial Intelligence (ICCAI), pp. 160-166, 2022..

[5] S. Kim and K. Moon, "Precision Adaptive MFCC Based on R2SDF-FFT and CNN for Real-Time Audio Classification," IEEE Signal Processing Letters, vol. 30, pp. 408-412, 2023.