Application in Musical Instrument Identification and Transcription Through Machine Learning

Aaron Luna
Texas State University
IEEE
San Marcos, TX
a_l523@txstate.edu

Jessica A. Wright
Texas State University
IEEE
San Marcos, TX
pvp9@txstate.edu

***Abstract*— This document explores the use and process of machine learning to identify instruments from recordings. Using the IRMAS dataset, audio samples of western instruments are trained and tested through a long short-term memory layer and convolutional neural network machine learning model. The results provide mixed accuracies by instrument. The overall highest accuracy scores were awarded to the instruments which have the most universal sound, the organ. This is due to the nature of the sound actuating mechanism and the fact that it is not solely dependent on human breathing. Wind-blown instruments had much less accuracy in testing because a wide variety of sounds, timbres, styles, articulation, and tone exist in any given audio sample.**

I. Introduction

A. Executive Summary Introduction

Musical transcription is a taxing practice, requiring years of aural and composition experience. With the use of synthesizers and MIDI files, it is possible to play an electronic instrument (synthesizer) and have the collected MIDI file converted into notation software. However, the same cannot be done with a manual wind-blown or percussive instrument. Some applications have been created to accomplish this but fail when it comes to large ensemble audio files. Classical instruments are not all transcribed the same way; different clefs, octaves, and transpositions depend on the traditionally accepted notation of each 'voice.' The need arises for a transcription service to separate the sound of multiple instruments in an audio file with the correct composition techniques, especially in the case where no transcription yet exists (folk and gospel music, improvisation, damaged/lost records, etc.).

B. Music Understandment

Understanding music involves various components, and one way to effectively organize it is through music extraction. Going into the realm of music analysis, extracting essential components plays a pivotal role in unraveling its intricate layers. With the number of available music songs and genres in today's world, the organization type can be whatever is desired. The methodology can be fit to suit specific choices. It can be exploited successfully depending on how much application is put into it. The efficacy of this approach greatly hinges on the depth of application and the precision of extraction techniques used.

An instrumentation depiction is another extraction in which aids in determining music genre. It not only helps in genre identification but also provides a nuanced understanding of musical nuances. The ability to discern and dissect the presence of various instruments within a composition forms the bedrock for comprehensive music analysis. Having deeper insights into the presence of different musical instruments is crucial for this project. This insightful process is facilitated by leveraging sophisticated separation algorithms explicitly designed for audio, enabling the isolation of individual instruments within a musical piece. This music source separation is a great form of constituting an audio mixture with heavy computational implementations. However, it's important to note that the accuracies of these algorithms might not always be very high. Despite the advancements so far, the inherent limitation lies in the variable accuracies of these algorithms new or old, indicating the continuous quest for enhancement and refinement in real-world applications.

Music source separation represents an innovative avenue where intricate audio blends are meticulously disentangled, often demanding substantial computational resources. This computational complexity underscores the challenges in achieving consistently high accuracies, emphasizing the ongoing need for iterative improvements and advancements in the field.

Navigating these complexities requires a systematic approach that initially involves the reduction of instruments detected within music files. This gradual improvement paves the way for incremental accuracy enhancement, fostering a more comprehensive comprehension and analysis of a composition's instrumentation. By following this methodical pathway, a deeper understanding of music emerges. It is like unveiling its underlying composition and genre intricacies, thereby enriching the overall musical landscape.

II. Background

A. Executive Summary Background

A 2012 study, "A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals (Bosch, Janer, Fuhrmann, Herrera)," sought to automate instrument recognition as a step towards the goal of transcription. However, this study did not implement the myriad applications of this ability. Because the same instrument

may sound different when played by another instrumentalist (technique, experience, airflow), the most important attributes are in the frequency range, voice part (bass, harmony, melody, rhythm), and overtones. Creating a more accurate instrumentation recognition system will produce a transcription service with more capabilities and higher accuracy. In the context of this project, instrumental identification itself is useful in organizing a music library by ensemble or instrument type. Alternatively, a musician may isolate their part from the recorded audio to listen and practice alongside the track. Implementations for design engineers include taking the data after machine learning to build specialized headphones designed for specific music genres. An audio engineer may be most thankful for this type of filtering to avoid multiple microphone sound set-ups and the time that comes with mixing each track individually.

B. Background for Neural Network type

LSTM, which stands for Long Short Term Memory, constitutes a crucial component of Recurrent Neural Networks (RNNs) and excels in learning extended dependencies present within sequential data. Stacking LSTMs allows the creation of deep LSTM networks, enhancing their capacity to process and understand complex sequences effectively.

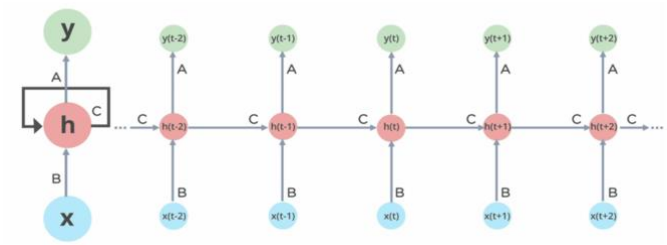

Fig. 1. Example of RNN

RNNs primarily excel in handling sequence and time series data due to their accuracy in modeling sequential information, despite their tendency for slower computations. A distinctive feature of RNNs lies in their short-term memory functionality, enabling the storage of only a limited span of information, thereby influencing their ability to retain and process long-term dependencies.

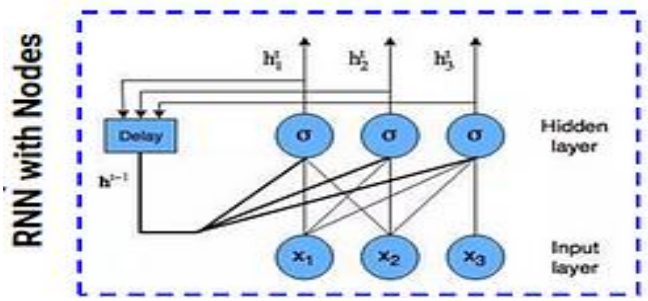

Fig. 2. Implementation of RNN with hidden layer

III. DATASET

TABLE I. FEATURE OPTIONS TO CATEGORIZE AUDIO FILES, DEPENDING ON NEED.

Frequency Domain Features	Time Domain Features	Statistical Features
Spectogram	Root Mean Square Energy	Mean
MFCC	Zero Crossing Rate	Standard Deviation
Spectral: Contrast, Bandwidth, Roll-off	On-set	Skew
Chromogram	Off-set	Kurtosis

A. Training Data

Within the training dataset, an intricate organizational structure is established with individual folders meticulously tailored to house and distinguish various instrument types. Each dedicated folder encapsulates a rich assortment of music or sounds exclusively attributed to a specific instrument, fostering a focused environment for honing the model's understanding of distinct musical elements. This granular segregation not only streamlines the training process but also enables the machine learning algorithm to grasp the nuanced characteristics unique to each instrument type.

The diversity within this curated repository manifests in the variance of available files attributed to each instrument. Some categories exhibit a plethora of files, representing a wide spectrum of musical compositions or sound samples, whereas others might have a more limited collection. This variability in file quantity adds a layer of complexity, demanding the model's adaptability to learn from diverse data distributions, ensuring a robust comprehension of various musical nuances.

Moreover, embedded within these files lies a spectrum of audio quality. The sounds captured in these files exhibit a spectrum of fidelity, where certain audio recordings boast exceptional clarity and crispness, presenting a pristine representation of the instrument's timbre and tonal characteristics. Conversely, other recordings might display variations in quality, with nuances of background noise or fluctuations in clarity, presenting a broader range of acoustic scenarios for the model to discern and decipher.

The multifaceted nature of this training dataset, encompassing a mosaic of instrument types, variable file

quantities, and a spectrum of audio qualities, serves as a rich foundation for the machine learning model. It fosters a comprehensive learning environment, instilling adaptability within the model to discern and differentiate between diverse musical elements, thereby enhancing its accuracy and robustness in instrument classification tasks.

B. Testing Data

Contrary to the intricate structure of the training dataset, the testing data introduces a different dimension in the form of complete songs, showcasing a diverse array of musical compositions. However, within this collection, disparities in song duration are evident, presenting varying lengths among the files. Some testing files encompass extended song segments, while others contain shorter excerpts, contributing to a heterogeneous compilation that challenges the model's adaptability to varying song lengths.

Adding to the complexity, each of these testing data files is accompanied by a .txt files meticulously detailing the instruments featured within specific segments of the song. These text annotations serve as valuable guides, offering comprehensive insights into the instrumentation present at distinct junctures of the song. This breakdown, segment by segment, furnishes the model with a detailed roadmap. It is facilitating its ability to associate instruments with corresponding segments of the music, further refining its understanding of multi-instrumental arrangements.

The bundle of complete songs with varying durations, coupled with detailed textual annotations specifying instrument inclusion, presents a multifaceted challenge to the testing phase. These variations challenge the model's adaptability, necessitating its capacity to navigate through disparate song lengths and accurately associate instruments with corresponding segments, thereby ensuring a robust and comprehensive evaluation of the model's classification capabilities.

This nature of the testing data, characterized by diverse song lengths and comprehensive instrument annotations, serves as a rigorous evaluation framework for the machine learning model. It demands adaptability and precision from the model, compelling it to accurately discern and classify instruments within varying segments of music, ultimately fortifying its efficacy in real-world applications where songs exhibit diverse compositions and durations.

C. Equations

$$P_1 = \frac{tp_1}{tp_1 + fp_1} \tag{1}$$

The equation above is used for precision score and the one below is used for recall. We calculate: true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn).

$$R_1 = \frac{tp_1}{tp_1 + fn_1} \tag{2}$$

The equation below is for F1 score and is the harmonic mean between precision and recall.

$$F_1 = \frac{2P_1R_1}{P_1 + R_1} \tag{3}$$

IV. Procedures

Using the Librosa python library, the audio files are converted to the frequency domain to extract features in Table 1. Specific subsets of the .librosa archive includes .decompose, .beat, .feature., .filters, .onset to transfer these files. The data is then returned in an array with labels. Each feature is normalized and stored in a matrix. A combination of machine learning methods is necessary. The Convolutional Neural Network (CNN) model handles spatial features (i.e., spectrogram) and then feeds into the Long Short-Term Memory (LSTM) layer for temporal features (i.e., rhythm). Additional layers which may prove necessary are batch normalization and dropout for convergence and regularization, respectively. Overfitting is prevented by model checkpointing and early stopping.

The training and testing procedures necessitate the isolation of each instrument into distinct files, requiring separate training and testing for every instrument category.

Initially, the provided code was adjusted for machine learning techniques, distinguishing between training and testing data. Multiple iterations were conducted trying to optimize hyperparameters. Like altering batch size, learning rate changes, and dropout rate changes, with specific numerical values yielding superior performance in the model. Furthermore, modifications were made by transitioning between Conv1D and Conv2D models, along with the integration of callbacks for enhanced functionality. Creating an ensemble to combine predictions from multiple models to improve overall performance. An example of how we created an ensemble of LSTM models was using a voting approach for classification tasks. Cross validation to employ techniques like K-fold cross-validation to evaluate model performance more reliably and prevent overfitting.

Refinements were introduced to the standard scalar method, optimizing its role in preprocessing the data. Training the model for an optimal number of epochs. Too few epochs may lead to underfitting, while too many epochs can cause overfitting. Notably, the inclusion of a confusion matrix provided a visual representation of the model's performance. Throughout these iterations, variations in metrics such as accuracy, precision, recall, and F1 score were observed, fluctuating with different implementations and configurations.

V. Results

The IRMAS dataset was split 80% for training and 20% for testing. Test documentation for model efficiency will include, but is not limited to, precision, accuracy, recall, and F1 scores. Upon these findings, a confusion matrix will be used to troubleshoot classification errors.

TABLE II. ACCURACY, PRECISION, RECALL, AND F1 SCORES FOR TRAINING DATA

Music Type	# of files in folder	Accuracy	Precision	Recall	F1 Score
Cello	388	38.46%	0.26	0.38	0.30
Clarinet	505	41.18%	0.29	0.41	0.33
Flute	451	36.96%	0.19	0.37	0.24
Acoustic Guitar	637	50.00%	0.31	0.50	0.38
Electric Guitar	760	39.47%	0.20	0.39	0.26
Organ	682	78.00%	0.78	0.78	0.77
Piano	721	49.32%	0.39	0.49	0.44
Saxophone	626	46.03%	0.28	0.46	0.29
Trumpet	577	55.17%	0.40	0.55	0.47
Violin	580	38.00%	0.23	0.38	0.29
Singing Voice	778	36.17%	0.24	0.36	0.29

Organ instrument recognition did the best, most likely because organs have a machine providing a steady flow of air. The same passage performed on multiple organs will have the most similar qualities of the trained instruments because the air actuator is universal.

Saxophone instrument recognition earned the lowest accuracy score. Out of the trained instruments, the saxophone samples varied the most in musical style. For example, a jazz solo saxophone performance typically uses a higher range of pitches (frequencies) than classical. Articulation (pitch onset) is harsher, tone is breathier, the resonance chamber of the player is smaller. These aspects are not identifiable by a computer; however, they can be measured by volume (amplitude), noise, and present overtones, respectively. When the performer can control the instrument in less-common ways to achieve a style, that audio sample becomes more of an outlier.

Despite modifications in various fields, the accuracies across the rest of the data remained consistently low. Surprisingly, the accuracies reported earlier were the best outcomes achieved. These persistently low scores raise doubts about the applicability of this dataset to engineering aspects due to the considerable challenges in achieving higher accuracy levels.

Similarities Across Various Songs from all 3 testing data folders:

1. **Instrumental Variation:** Multiple songs showcase a wide range of instrumental variety, including electric guitar, piano, saxophone, trumpet, violin, and acoustic guitar, among others.
2. **Vocal Emphasis:** Many songs emphasize vocals, sometimes accompanied by various instruments in the background, ranging from clear and crisp singing to foreign language vocals.
3. **Audio Quality Variation:** Throughout the files, variations in audio quality are noticeable, with instances of clear, crisp sounds, fuzzy background noises, and occasional difficulty in distinguishing instrument types due to audio quality issues.
4. **Genre Influences:** There are prevalent genre influences, notably in jazz, blues, and classical styles, influencing the instrumentation and overall musical composition across different songs.
5. **Song Segment Structure:** Some songs consist of multiple preview files broken down from longer compositions, showcasing segments that vary in length and composition.
6. **Challenges in Instrument Recognition:** Certain songs present challenges in accurately recognizing instruments due to softer or quieter segments, rapid tempo changes, or chaotic and heavy music compositions.
7. **Variability in Genre Recognition:** Different genres are represented within the testing data, with varying degrees of ease in distinguishing and recognizing instruments based on the genre's characteristics.

These commonalities provide a glimpse into the diverse and complex nature of the testing dataset, highlighting recurring patterns and challenges that the machine learning model may encounter in identifying and distinguishing instruments within music segments.

VI. CONCLUSION

Long Short-Term Memory (LSTM) proves to be a valuable tool for handling this data due to its ability to retain information over extended sequences. It is very crucial and specific for music data since it one of the most popular and effective ways to process it cleanly. It serves well in applications in Neural Networks handling large amounts of data and can be specified to a specific subcategory called RNN. Recurrent Neural Networks (RNNs) require a substantial amount of data to perform effectively, impacting their performance significantly.

Music, being incredibly diverse worldwide, showcases vast variations in styles, instruments, and compositions across different cultures and regions. This extensive diversity in music entails that the same instruments

can be utilized across multiple genres, exhibiting distinctive sounds based on their applications. Deciphering all components used within complete songs can present challenges during implementation due to the complexity and layering of musical elements.

Not all music data is considered useful or beneficial, as certain datasets may contain inconsistencies or inadequacies affecting the analysis. Certain musical instruments demonstrate superior adaptability for conversion into the frequency domain, enabling easier interpretation and analysis. Despite attempts, the achieved accuracies fell short of meeting the criteria for implementation in our intended applications, necessitating further refinement in both code implementation and research efforts.

To improve outcomes, additional code refinement and research exploration are essential, along with the exploration and adaptation of diverse methodologies. Additionally, a refined dataset with more considerations may improve accuracies. Audio samples could be accompanied with a more detailed text file including genre, ensemble, performer experience level, etc. A completely new dataset could easily be created and catered to the purpose of the project with considerations in recording quality or making sure that only one performer plays any given instrument in one selected style.

VII. Future Work

Music has proved itself to be one of the concepts that has outlived a lot of other creations or even technological advances. Some can argue it has been around for as long as thousands of years since music is just a composition of sounds put together. Sounds from nature fall into music genres and even outer space sounds. Technology will continue to advance over the next couple of decades and music will naturally advance with it. Some immediate small concepts for future work can be music data being converted into a different file type for easier storage or interaction. This conversion can also allow for different Machine Learning techniques to arise to hopefully build a neural network with strong qualities for testing and training.

As far as future work that could happen immediately, this would revolve around testing and training on more data. Adding into the mix of this current data more music from different instruments. This could allow for more learning within Librosa library to occur and be able to have sounds from all kinds of musical instrument items. This can better apply to creating technology-based devices for specific music wants like discussed in the executive summary of building those specialized headphones. Also with helping out in a recording studio for making new music with the help of AI.

A different approach can be taken with the help of FASST. It provides a platform to experiment with various source separation algorithms, including BDMO, to separate audio sources mixed within multichannel recordings.

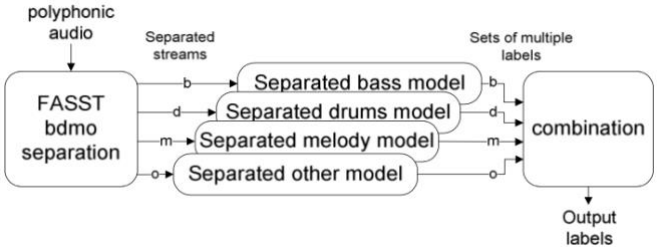

Fig. 3. FASST separation combined with intrument recognition using models

References

[1] Juan J. Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera, "A Comparison of Sound Segregation Techniques For Predominant Instrument Recognition In Musical Audio Signals", in Proc. ISMIR (pp. 559-564), 2012.

[2] Juan J. Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera, "A Comparison of Sound Segregation Techniques For Predominant Instrument Recognition In Musical Audio Signals", in Proc. ISMIR (pp. 559-564), 2012

[3] Dominica Jay, Long short-term memory (LSTM) RNN in Tensorflow, 2023, Web.

[4] Aakarsha Chugh, Deep Learning | Introduction to Long Short Term Memory, 2023, Web.

[5] Damian Valles, "Camvas Lecture Slides", Texas State University, 2023