Music Signal Generation and Classification

Puya Fard

Prof. Dr. Terry Sanger
University of California - Irvine

1 Introduction

The objective of this project is to do **music signal recreation.** Leveraging the trained CNN model, genre-specific feature representations are synthesized into new music signals. These features are converted into audio waveforms, demonstrating the potential of neural networks to generate realistic genre-specific samples. To validate the quality of the recreated music signals, they are **reclassified** using the same trained model, ensuring fidelity to the original genres. Finally, a **Principle Component Analysis (PCA)** will be performed on the reconstructed data for the four genres to examine and highlight the distinct differences in their variance distributions.

Furthermore, this project aims to develop and evaluate various machine learning models for **music genre classification** using computed audio features. The models tested include Dense Neural Networks, Convolutional Neural Networks (CNNs) with Conv2D layers, and hybrid architectures combining Long Short-Term Memory (LSTM) and Convolutional Neural Networks (LSTM+Conv2D). To evaluate, there will be compare and contrast analysis of the signals on the performance of these architectures.

2 Dataset

The GTZAN dataset is utilized for this project, containing 1000 .wav audio files spanning 10 distinct genres: disco, metal, reggae, blues, rock, classical, jazz, hip-hop, country, and pop. For this analysis, specific audio features are extracted and analyzed to represent each file effectively. These features include:

Chroma Feature: The mean of chroma Fourier transform.
Spectral Centroid: The mean of the spectral centroid.
Spectral Bandwidth: The mean of the spectral bandwidth.
Spectral Rolloff: The mean of the spectral rolloff.
Zero Crossing Rate: The mean of the zero-crossing rate.
MFCC Extraction: Mel Frequency Cepstral Coefficients.

Below Figure 1 is the MFCC of blues.0000.wav audio file.

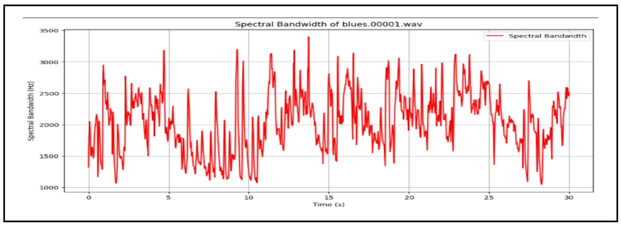

Figure 1: Spectral Bandwidth Visualization.

3 Problem Statement

Traditional approaches for music genre classification often fail to capture the complex temporal and spatial patterns in audio signals, making it challenging to distinguish between similar genres. The project aims to advance the understanding of how neural networks can effectively **analyze, classify, and generate audio data,** contributing to applications in music recommendation systems, audio synthesis, and music production.

4 Methodology

This project employs a structured approach to music genre classification and music signal recreation using Convolution (Conv2D), Dense, and Long short-term memory (LSTM) machine learning models. The methodology is divided into the following key steps:

- **Feature Extraction:** For each audio file, the key features listed in Dataset section are extracted to serve as input to the network models. These features are standardized and combined to form a feature matrix for training and testing.
- **Model Development:** Three machine learning architectures are implemented and evaluated:
  - **Dense Neural Networks (DNN)**
  - **Convolutional Neural Networks (CNN)**
  - **Hybrid Models (LSTM+CNN)**
- **Music Signal Recreation:** The trained CNN model is used to synthesize genre-specific feature representations. The genres used to recreate music signals are Classical, Pop, Metal, and Blues. These features are converted into audio waveforms using inverse transformations to recreate realistic music signals for each genre. The recreated signals are then reclassified using the trained model to validate their fidelity to the original genres.
- **PCA Analysis:** Principal Component Analysis (PCA) was performed to analyze the variance distribution of the features extracted from the four selected music genres: classical, blues, metal, and pop.

5 Findings

For each model, the performance is visualized using the training and validation mean squared error (MSE) loss curves, synthetic vs True MFCC signal, classification report, and confusion matrix.

5.1 Recreating Music Signals using dataset

The approach integrates classification and synthesis to generate feature-based audio representations, which are evaluated for genre fidelity using the same trained model. There will be two models for this section, one includes an LSTM hidden layer, one doesn't.

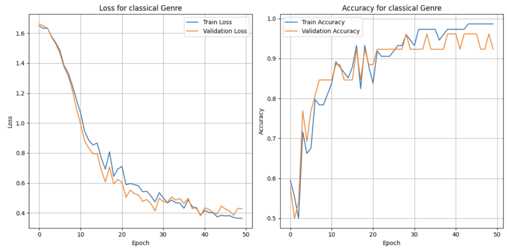

Figure 2: Validation and Loss results: Classical

Results above on Figure 2 represent the model that includes the LSTM hidden layer. The recreated classical genre performing well with LSTM model, the validation accuracy is as high as 95%. Now lets compare with the model that doesn't include.

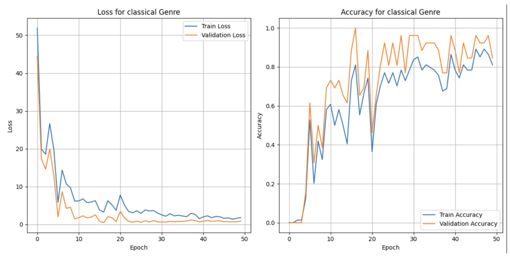

Figure 3: Validation and Loss results: Classical

The results presented in Figure 3 correspond to the model without a hidden LSTM layer. Compared to the model that includes an LSTM layer, it demonstrates lower validation accuracy, indicating the LSTM's effectiveness in capturing temporal dependencies for improved genre classification.

Now that we observed the differences in the model for classical genre, next step is observe the true vs synthethic audio file on Figure 4 for classical.0.wav, it is successfully able to replicate the signal after training process.The model was trained using the Adam optimizer with a learning rate of 0.0001 and categorical cross-entropy as the loss function. The training was conducted over 20 epochs with a batch size of 32, using an 80-20 train-test split.

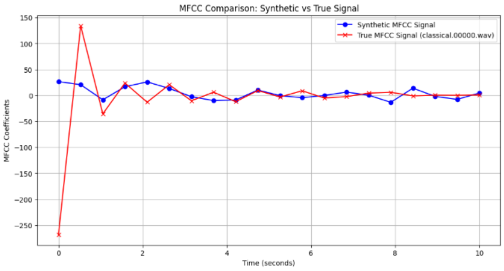

Figure 4: Synthetic vs True MFCC Signal

5.2 PCA Analysis on reconstructed music genres

Principal Component Analysis (PCA) was applied to the reconstructed music genres to explore and visualize the relationships between the genres in a reduced feature space. This visualization highlights the underlying patterns and separability of the genres based on their extracted features, such as MFCCs, chroma, and spectral centroid.

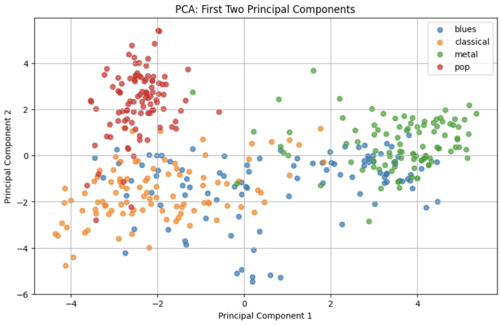

Figure 5: PCA plot

The plot shown in Figure 5 reveals distinct patterns in the reconstructed music genres. Notably, the "pop" genre is well-distinguished, forming a clear cluster in the PCA-reduced feature space. In contrast, "blues" and "classical" genres exhibit significant overlap, suggesting similarities in their extracted features, such as tonal or spectral characteristics. The "metal" genre shows partial separation, with some clustering but also areas of overlap with "blues" and "classical." These observations highlight the varying degrees of separability among the genres and suggest that certain features may better capture the unique characteristics of "pop" while struggling to differentiate between "blues" and "classical."

5.3 Music Genre Classification

To understand the concept of music classification better, and determine which models performs the best when it comes to classify the genres one from another, we have created three distinct models and architectures that will extract information from the music files, and train their corresponding network model and plot outputs. The following models were created:

- **Dense Classifier**

  This model splits the dataset into training and validation sets using 50/50 and 70/30 ratios, with a random state of 32 to ensure consistency in the training process. Following with the Adam optimizer with a learning rate of 0.0001. The model will be trained for 20, 50, and 100 epochs, using batch sizes of 16, 32, and 64, with verbose set to 1.

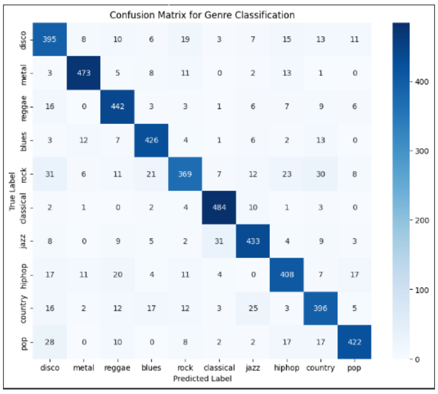

Figure 6: Confussion Matrix for Dense Classifier

From Figure 6, we can conclude that the overall accuracy of the model over the dataset trained is **85%.** It is definitely a strong percentage in terms of accuracy, however not quite the best model.

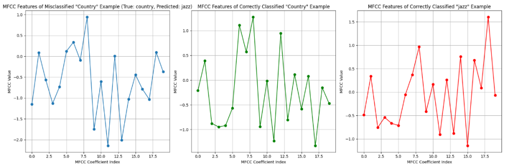

Figure 7: Miss-classification of "Country" audio-file

The figure above (Figure 7) presents an example of a misclassified signal originally from the country genre, which was incorrectly identified as "jazz." By comparing the MFCC features of the true model outputs for both the country and jazz genres, we can better understand the reasons behind this misclassification.

- **CNN Classifier**

  The CNN classification model used to train my dataset contains three Conv2D layers with increasing filter sizes (64 → 128 → 256) to extract local patterns from features with all layers using ReLU activation and the same padding.

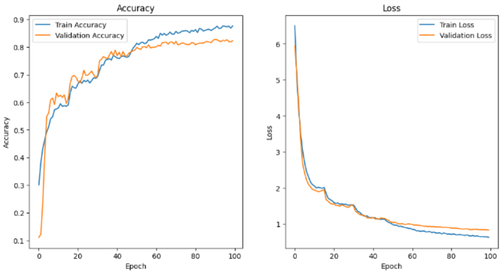

Figure 8: Accuracy results

Above Figure 8 model is trained for 100 epochs, with batch_size 32 and verbose 1 with a 40-60 split.

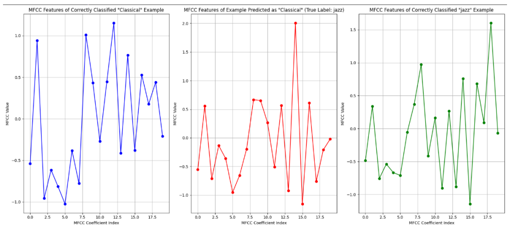

Figure 9: Miss-classification of "classical" audio-file

above Figure 9 the classification signal plot for Mfcc on a classical audio-file predicted. We can observe and understand why that happened due to certain similarities in Mfcc signal in between the two genres. Regardless, the overall accuracy of the model over the dataset trained is 82%. It has improvements on genre “Blues.

- **LSTM+CNN CLassification**

  This model primarily focuses on ConvLSTM2D for capturing spatiotemporal features, followed by batch normalization and dense layers for classification.

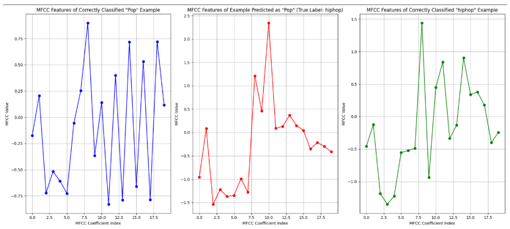

Figure 10: Miss-classification on "Pop" audio-file

From above Figure 10 we can observe the Mfcc signal output of a "pop" audio-file after the training process. This classification model showed significant improvement in "pop" genre with a 91% accuracy. Regardless, there were a few miss-classifications and one of the categories is "hip-hop."

6 Conclusion

Synthethic music recreation with LSTM hidden layer model demonstrated a 92% accuracy. The Dense Neural Network for original genre classification achieved the highest overall accuracy of 86%, with particularly strong performance in recognizing genres like "Classical," "Metal," and "Pop." The CNN model, although exhibiting higher accuracy in individual genres such as "Blues" and "Hip-hop," resulted in a slightly lower overall accuracy of 83%. The LSTM+CNN hybrid model performed the worst among all with an accuracy of 75%, suggesting that it struggled to learn meaningful temporal dependencies with the given dataset configuration.

In conclusion I have learned how to perform music recreation, music classification, and PCA analysis on output data.