Music Genre Classification of Audio Signals

Hakan Tekgul
University of Illinois Urbana-Champaign
Urbana, Illinois
tekgul2@illinois.edu

Raimi Shah
University of Illinois Urbana-Champaign
Urbana, Illinois
rsshah2@illinois.edu

ABSTRACT

ABSTRACT HERE!

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; **Reliability**;

KEYWORDS

Fault Tolerance, Reliability, Instruction Criticality, Embedded Systems

ACM Reference Format:
Hakan Tekgul and Raimi Shah. 2018. Music Genre Classification of Audio Signals. In *Proceedings of Machine Learning for Signal Processing Conference (CS 598 PS Fall 18')*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the past decade, with the introduction of technology that can store huge amounts of data, a large amount of musical data is increasingly available to public on different application platforms such as Spotify. As the number of musical data in our phones and Internet keeps increasing, there is a need to characterize each music track so that finding a specific song in a large archive of music would not be a problem. Musical genres are commonly used to describe and characterize songs for music information retrieval. Pachet suggests that genre of music can be the best general information for music content description [1]. Hence, a system that can classify musical genres can solve the problem of locating a specific sound track on any device.

The only problem with musical genre classification is the fact that the definition of genre is very subjective by its nature and there exists thousands of genres or sub-genres. It is also important to note that, the definition of music genre tends to change with time, as what we call Rock song today is very different from the rock songs twenty years ago. Even though musical genres are subjective, there are certain features that can easily distinguish between different genres. By using features such as distribution of frequency or the number of beats, it is possible to classify main genres of music. For classification of musical genres, various approaches have been proposed. Unfortunately, most of these approaches have been proven to show accuracy around 60-70%. Therefore, new approaches that can maximize classification accuracy must be considered.

Hence, we try to improve the classification accuracy of music genre classification of audio signals in this work. Specifically, we use a wide range of machine learning algorithms, including k-Nearest Neighbor (k-NN) [12], k-Means Clustering [14], Support Vector Machines [9], Gaussian Mixture Models [1] and different types of Neural Networks to classify the following 5 genres: metal, classical, blues, pop, country.

Our main goal in this study is to maximize the classification accuracy of 5 genres and compare different methods of machine learning for classification of audio signals. We use state-of-the-art machine learning platforms such as PyTorch [11] to introduce deep learning into our project. We experiment with different neural network architectures and types of neural network. Moreover, we use Mel Frequency Cepstral Coefficients (MFCC) [3] to extract useful information from musical data as recommended by past work in this field. To summarize, we make three main contributions in this paper:

- We experiment with a wide range of machine learning algorithms and state their classification accuracy for 5 different genres.
- We propose a method for feature extraction and audio processing that is dependent on both MFCC and PCA. We also discuss the significance of such methods.
- We report experimental data that describe the overall effectiveness of our classification methods by including confusion matrices.

——»> ADD a paragraph that states experimental results briefly!!!! «<———

2 RELATED WORK

The development of music genre classification has been increasing rapidly in the past decade. Many approaches have been proposed that build different models for genre classification. Some approaches concentrate on the processing of audio signals, whereas some approaches try to combine audio signals with lyrics from each musical track to increase accuracy. Some of the related work to our project is presented below.

Firstly, Tzanetakis and Cook [13] introduced different feautres to organize musical tracks into a genre by using k-NN and Gaussian Mixture Model (GMM) methods. Three different feature sets for speaking to tumbrel surface, rythmic substance and pitch substance of music signse were suggested. They also introduced a dataset for music genre classification (GTZAN Dataset [13]), which is widely used today in many project, including ours.

Furthermore, Aucouturier and Pachet [1] used GMM and utilized Monte Carlo procedures for evaluation of KL divergence, which

CS 598 PS Fall 18', December 2018, Champaign

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

was used in a k-NN classifier. They conveyed some significant component sets for musical information retrieval that we use in our work, specifically the MFCC.

Apart from models such as GMM or k-NN, Feng [2] proposed an approach that uses Restricted Boltzmann machine algorithm to build deep belief neural networks. By generating more dataset from the original limited music tracks, he shows great improvement in the classification accuracy and describes the significance of neural networks for music genre classification.

Xing et. al. [16] proposes a similar approach that uses convolutional neural networks. By combining max and average pooling to provide more more statistical information to higher neaural networks and applying residual connections, Xing et. al. [16] improves the classification accuracy on the GTZAN data set greatly. Li, Chan and Chun [7] recommend a very similar technique to concentrate musical example included in audio signals by using convolutional neural networks. They present their revelation of the perfect parameter set and best work on CNN for music genre classification.

Finally, Smaragdis and Whitman [15] presents a very interesting musical style identification scheme based on simultaneous classification of auditory and textual data. They combine musical and cultural features of audio tracks for intelligent style detection. They suggest that addition of cultural attributes in feature space improves the proper classification of acoustically dissimilar music within the same style.

–» Add a paragraph that compares our work to above «– As compared to these works,...

3 PROPOSED APPROACH

3.1 Musical Dataset

For musical data, Marsyas is an open source software framework for Music Information Retrieval with the GTZAN Genre Collection Database, which has 10 genres and each genre has 100 30-second audio tracks. All the tracks are 22050 Hz Mono 16-bit audio files in .au format.

For this project, we chose five distinct genres; classical, metal, blues, pop, country. Hence, our dataset was 500 songs total, from which we used 80% for training and 20% for testing. We chose five very distinct genres as previous works [7] suggest more than 5 genres can decrease accuracy a lot and introduces many problems.

3.2 Feature Extraction: Mel Frequency Cepstral Coefficients (MFCC)

Previous works [3] on music classification and processing of audio signals directed us to use MFCC (Mel Frequency Cepstral Coefficients) as a method for feature extraction so that time domain waveforms can be represented in the frequency domain in a mel-scale. For the process of MFCC, we first computed the spectogram of each waveform by using Fast Fourier Transform and a Hamming Window. Then, we mapped each frequency to mel scale, as mel scale is the best scale for human ears. The mel spectogram of a song from each genre is shown on Figures 1 through 5, so that the difference between each genre can be visualized. After computing the mel-spectograms of each song, we applied discrete cosine transform (DCT) and then removed the very high frequency values from our data. At the end, we had an MFCC array of each song, where

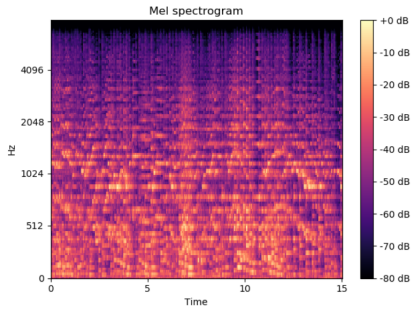

Figure 1: Mel-spectogram of classical genre.

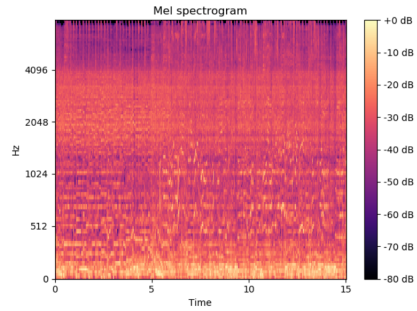

Figure 2: Mel-spectogram of metal genre.

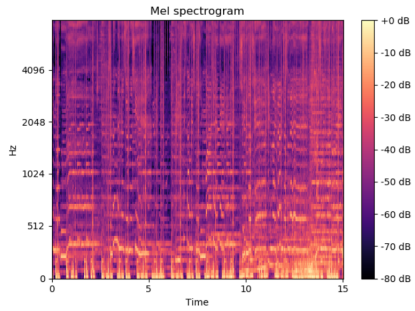

Figure 3: Mel-spectogram of pop genre.

we stacked all them together, created appropriate labels for each genre and constructed our training and testing datasets. As stated, we used 80% for training and 20% for testing our classifiers.

3.3 Dimensionality Reduction with Principal Component Analysis (PCA)

After feature extraction and construction of final dataset, we thought of using dimensionality reduction before putting out data through classifiers. Since Principal Component Analysis (PCA) is a well-known and effective method for reduction of dimensions, we used PCA on our dataset. A realistic choice for number of reduced dimensions is to visualize the data with different PCA values and then pick the minimum dimension that can keep at least 95% of the

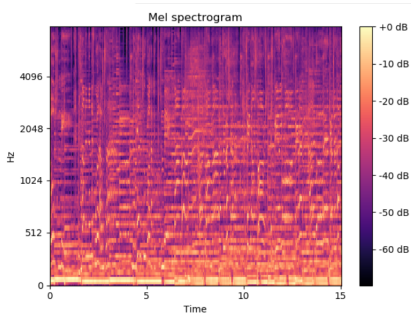

Figure 4: Mel-spectogram of country genre.

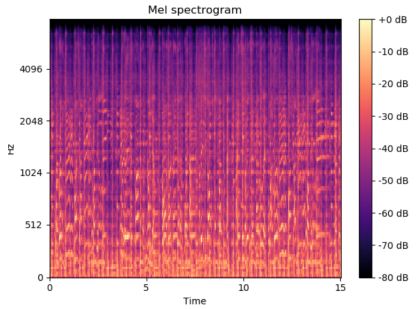

Figure 5: Mel-spectogram of blues genre.

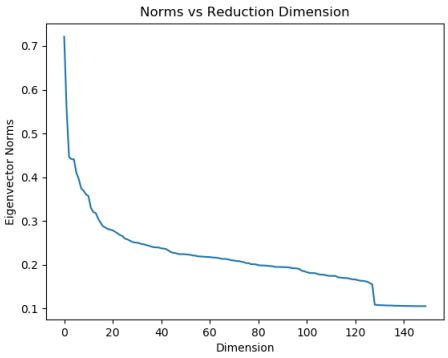

Figure 6: Norms of eigenvectors of our data plotted with respect to PCA dimensions. Note that we only want to keep the most significant components.

significant components. The visualization of our data with respect to different PCA dimensions and the corresponding eigenvectors is shown on Figure 6. Note that figures 7 and 8 also present a visualization of each genre in 2 and 3 dimensions. After extensive analysis of the visualization, and experimenting with out classifiers, we reduced the dimensionality to 16. Even though 16 dimensions performed very well on classifiers such as k-NN or SVM, we had to use much bigger dimensions for our neural network, since neural networks need much more data in practice.

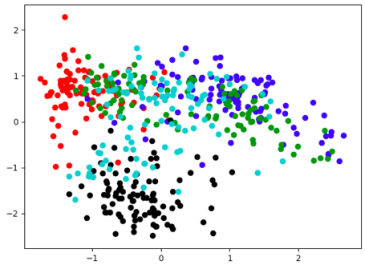

Figure 7: 2-dimensional scatter plot of our data with different genres.

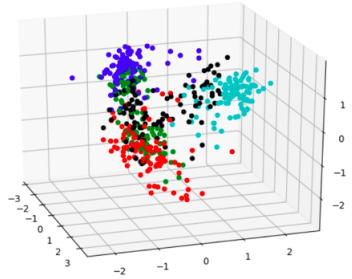

Figure 8: 3-dimensional scatter plot of our data with different genres.

3.4 Machine Learning Algorithms

3.4.1 ***K-Nearest Neighbor (K-NN).*** The first algorithm we used is the very famous and effective k-closest neighbors algorithm. k-NN is a non-linear algorithm that can detect direct or indirect spread of data. It is very effective for huge amounts of data. One downside of k-NN is the fact that it makes hard decisions and might produce low classification accuracy. Other than that, k-NN is computationally expensive since it does not learn any data, and it has to compute the distance for every point in prediction. We used Euclidean distance for k-NN, which produced good results. After our experiments, we also found that k=5 produced the best results.

3.4.2 ***Support Vector Machines (SVM).*** The second technique we used is the support vector machine (SVM), which is a directed organization method that discovers the extreme boundary splitting two classes of information [9]. The idea behind the algorithm is to project data onto a higher dimensional space in order to separate classes in a better way. We used Radial Basis Function (RBF) for our kernel for SVM and change the penalty to three.

3.4.3 ***K-Means Clustering.*** We attempted to use clustering by Kmeans which produces a hard assignment to a cluster for each data point. We initialized this algorithm randomly and assigned each point to the nearest cluster by using euclidean distance, and updated each cluster mean. This algorithm proceeds iteratively until the cluster means do not change. We found that reducing the data to 16 dimensions produced the best results. One challenge we had was evaluating clustering methods. We found that we could use Fowlkes-Mallows score to give an accuracy.

3.4.4 ***Gaussian Mixture Models (GMM).*** After KMeans clustering, we attempted to improve the clustering accuracy by implementing Gaussian Mixture Models. Each gaussian cluster has a mean and associated covariance, and each data point has a probability associated with each cluster. This gives soft assignments, which are usually better to deal with. We used sklearn's GMM class and experimented with different initialization techniques and covariance.

3.4.5 ***Simple 3-layer Neural Network.*** After implementing different wll-known classifiers, we wanted to experiment with neural networks since they generally produce promising results in machine learning applications. Firstly, we used PyTorch to process our dataset and constructed a 3-layer neural network that uses ReLU for nonlinearity. Then, we experimented with different PCA dimensions and different hidden layer sizes to produce the best accuracy. The architecture of the network that gives the best classification results is shown in Figure 9.

3.4.6 ***Convolutional Neural Network (CNN).*** We also wanted to experiment with Convolutional Neural Networks, since they can be much more effective than a simple 3-layer neural network or any other classifier. The downside of CNN is the fact that it requires a great deal of hyperparameter tuning. We experimented with different PCA dimensions, convolution kernel sizes, nonlinearity functions and the amount of dropout that we have to add. After testing our classifier with many different parameters, we were able to get a very good accuracy for music genre classification. The architecture that produced the best accuracy is shown on Figure 10.

3.4.7 ***Super-Classifier (SC).*** After completing a few other methods, we decided to try an ensemble method where we took 4 classifiers (GMM, k-NN, SVM, 3-layer Neural Network) and for each data point in the testing set, ran each classifier and took the most common label as the prediction. This produced better results than each of the methods individually. The motivation behind this implementation of the super classifier was the fact that each individual classifier had different least and most accurate genres.

4 EXPERIMENTAL RESULTS

4.1 Experiment Setup

The experimental setup for computation of classification accuracies were quite simple. After extracting features of our data through MFCC and applying PCA, we saved our training and testing datasets into a file, so that we do not have to do all the computations again. Then, we wrote a script that loads the training and testing datasets, and then puts the training data as an input to each of our classifiers with the corresponding genre labels. After training on each classifier, we computed predictions of each song and compared each label with its ground truth. Finally, we outputted the classification accuracy of each classifier and their confusion matrices, which are shown in Figure 11 through 18.

4.2 Classification Accuracy

5 CONCLUSION

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

Table 1: Classification results of each classifier

Classifier Type	*Accuracy*	*Most Accurate Genre*	*Least Accurate Genre*
K-NN	81%	Metal	Classical
SVM	84%	Classical	Country
GMM	85%	Metal	Blues
K-Means	-%	-	
3-layer NN	88%	Classical	Pop
CNN	-%	-	-
SC	- %	-	-

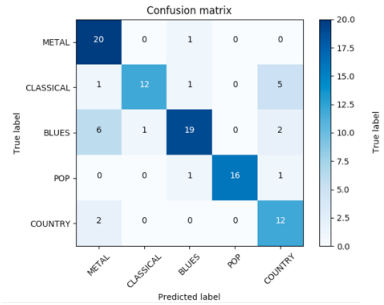

Figure 9: Confusion matrix for k-Nearest Neighbors.

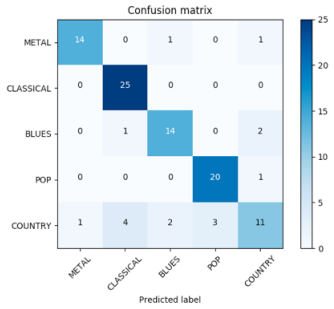

Figure 10: Confusion matrix for Support Vector Machines.

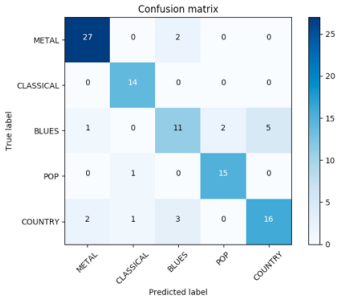

Figure 11: Confusion matrix for Gaussian Mixture Models.

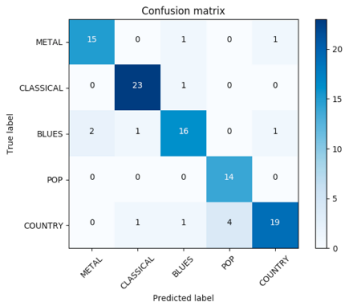

Figure 12: Confusion matrix for 3-layer Neural Network.

REFERENCES

[1] Jean-Julien Aucouturier. 2003. Representing Musical Genre: A State of the Art. *Journal of New Music Research* 32 (03 2003), 83–93. https://doi.org/10.1076/jnmr.32.1.83.16801

[2] Tao Feng. 2016. Deep learning for music genre classification. *Pattern Recognition Class Paper* (2016).

[3] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. 2011. A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia* 13, 2 (April 2011), 303–319. https://doi.org/10.1109/TMM.2010.2098858

[4] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown. 2001. MiBench: A free, commercially representative embedded benchmark suite. In *Proceedings of the Fourth Annual IEEE International Workshop on Workload Characterization. WWC-4 (Cat. No.01EX538)*. 3–14. https://doi.org/10.1109/WWC.2001.990739

[5] Chris Lattner and Vikram Adve. 2004. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-directed and Runtime Optimization (CGO '04)*. IEEE Computer Society, Washington, DC, USA, 75–. http://dl.acm.org/citation.cfm?id=977395.977673

[6] Chunho Lee, M. Potkonjak, and W. H. Mangione-Smith. 1997. MediaBench: a tool for evaluating and synthesizing multimedia and communications systems. In *Proceedings of 30th Annual International Symposium on Microarchitecture*. 330–335. https://doi.org/10.1109/MICRO.1997.645830

[7] Tom L. H. Li, Antoni B. Chan, and Andy HW. Chun. 2010. Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network.

[8] Q. Lu, M. Farahani, J. Wei, A. Thomas, and K. Pattabiraman. 2015. LLFI: An Intermediate Code-Level Fault Injection Tool for Hardware Faults. In *2015 IEEE International Conference on Software Quality, Reliability and Security*. 11–16. https://doi.org/10.1109/QRS.2015.13

[9] Michael I. Mandel, Graham E. Poliner, and Daniel P. W. Ellis. 2006. Support vector machine active learning for music retrieval. *Multimedia Systems* 12 (2006), 3–13.

[10] N. Oh, P. P. Shirvani, and E. J. McCluskey. 2002. Error detection by duplicated instructions in super-scalar processors. *IEEE Transactions on Reliability* 51, 1 (Mar 2002), 63–75. https://doi.org/10.1109/24.994913

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

[12] L. E. Peterson. 2009. K-nearest neighbor. *Scholarpedia* 4, 2 (2009), 1883. https://doi.org/10.4249/scholarpedia.1883 revision #137311.

[13] G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10, 5 (July 2002), 293–302. https://doi.org/10.1109/TSA.2002.800560

[14] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. 2001. Constrained K-means Clustering with Background Knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 577–584. http://dl.acm.org/citation.cfm?id=645530.655669

[15] Brian Whitman and Paris Smaragdis. 2002. Combining Musical and Cultural Features for Intelligent Style Detection. In *ISMIR*.

[16] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing. 2016. Improved Music Genre Classification with Convolutional Neural Networks. In *INTERSPEECH*.