**Music Genre Classification Based on   
Spectrogram Image Analysis**

David Chen

Harvard University  
51 Brattle St.  
Cambridge, MA 02138  
Telephone number, incl. country code

yuc995@g.harvard.edu

Cindy Liu

Harvard University  
51 Brattle St.  
Cambridge, MA 02138  
+1 (617) 237-0548

cindyliu@post.harvard.edu

**ABSTRACT**

In this paper, we introduce a methodology for the classification of audio data that leverages the many classification tools that currently exist for non-audio data. The classification of images and text data is well-established in the field of machine learning and many libraries and frameworks have been developed for such analysis. Audio classification, on the other hand, has lagged behind its atemporal counterparts, largely due to inherent difficulties in working with audio data. Our methodology allows researchers to apply an image classification tool to audio data by analyzing spectrograms, opening up many avenues in audio analysis and, we hope, accelerating progress in the development of tools for audio classification and analysis.

**Keywords**

Music genre classification; spectrogram; convolutional neural networks; image classification; feature extraction; MFCC; FFT; deep learning; machine learning.

# INTRODUCTION

Music in today’s world is ubiquitous. It can remind us of specific products and brands, thoughts and ideas, and past and imagined events, and can affect our mood significantly. When we hear a piece of music, we almost unconsciously categorize it in our minds – a pop song on the radio, electronic dance music in a nightclub, classical masterpieces at the symphony – and in the vast majority of cases, we have accomplished this labeling within the first 30 seconds of hearing the music. Therefore, there must be certain recognizable characteristics of different musical styles or genres themselves that allow us to classify what we hear so quickly, rather than through analysis of the complete musical work after experiencing it in its entirety.

Machine learning offers a way of giving us some insight into the problem. If a model can predict the genre of a piece of music with only data from the first 30 seconds, it will have achieved the same feat – and in an automatable way. The automation of music genre classification has become a busy area of research in recent years due to the demands of the music entertainment industry. To expand, a music service/business must amass a growing collection of musical works, and the manual annotation and organization of these ever larger and larger music collections has quickly become a wholly unscalable process. The search for efficient ways of automating the classification of music by genre (and many other labels) is thus quickly becoming an essential task in the music industry.

Unfortunately, in the field of machine learning, audio analysis has reached what some researchers call the “bottleneck of music genre classification” [3]. This is partially due to the lack of established conventions surrounding the format of audio data. Currently available audio datasets sometimes consist of raw audio files in varying formats, but more frequently researchers have chosen to reduce the size of the problem by training their models on a data set of derived features or annotations of the raw data rather than the raw audio itself. This has achieved some success – for example, Tzanetakis, *et. al.*, was able to reach a ~60% accuracy rate using standard classifiers and feature sets based on timbre (musical texture), rhythm, and pitch [4] – but many of these feature sets are determined based on previous research in speech recognition, which does not necessarily translate well to the much more complex feature space of musical data. Primarily, these feature sets fail to capture *musical intent*, that is, the inherent patterns underlying the composition of the music, rather than qualities of a particular recording or sample. Musical intent is a key characteristic of music that is not found in speech (the closest example might be spoken word performances), and which is not capturable by investigating pitch, rhythm, or timbre alone.

While much of the research on music genre classification has gone in the direction of feature set analysis, some intrepid researchers have set forth to do battle with raw audio data. Li, *et. al.*, showed promising results when applying CNN methodology directly to audio data (84% accuracy) using a fascinating modification to the image-based CNN in which they took windows of adjacent, overlapping audio clips along the duration of the full piece (paralleling the windows and stride parameters in image analysis) [3]. However, they only achieved 30% accuracy on their test dataset, which they acknowledge is due to overfitting. They additionally note that their method is highly sensitive to variations in musical timbre and other features (an artifact of their feature extraction method, MFCC), and performs less well when trained on data containing more than 4 labels and/or falling into genres that are generally considered to be similar to each other.

Thus a major problem remains of reliably classifying large music datasets that contain numerous labels of varying degrees of “similarity”. The size of the datasets renders it infeasible to employ manual classification to annotate each track; but significantly, it also means that manual extraction of features from the data (currently the only way of capturing musical intent) is also impractical. Classification techniques that do not require manual intervention are therefore at a high premium. We propose a method of spectrogram analysis that not only does not require manual intervention, but also preserves musical intent, generates a reduced feature space of manageable dimensions, and opens up musical analysis to the wide world of available image processing tools, allowing researchers to work around the overfitting problems experienced by previous researchers seeking to utilize image-processing tools directly on audio files.

# METHODOLOGY

## Dataset Selection

While the availability of music datasets is nowhere near that of text, image, or even speech datasets, several options currently exist.

### Million Song Dataset

The largest of these options by far is the Million Song Dataset (MDS) [1], containing metadata and derived features for 1 million Western contemporary music tracks (i.e., it lacks world music and is underrepresented in the classical genre and somewhat overrepresented in the pop, rock, and indie genres [2]). The MDS genre labels are sourced from a website called MusicBrainz.org which allows users to tag songs – each track

## Feature Selection

The

The

# RESULTS

## DDD

Ple

## DDD

The

## DDD

Please

## DDD

For

Table . Table captions should be placed above the table

|  |  |  |  |
| --- | --- | --- | --- |
| **Graphics** | **Top** | **In-between** | **Bottom** |
| Tables | End | Last | First |
| Figures | Good | Similar | Very well |

# CONCLUSIONS

Footnotes

## DDD

Do

# SUMMARY & FUTURE WORK

subsection

## Subsections

The

### Subsubsections

The

#### Subsubsections

The

#### Subsubsections

The

# ACKNOWLEDGMENTS

Our

# REFERENCES

1. Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. 2011. The Million Song Dataset. In *Proceedings of the 12th International Society of Music Information Retrieval Conference* (ISMIR 2011). http://labrosa.ee.columbia.edu/millionsong/
2. Dawen, L., Gu, H., and O’Connor, B. 2011. *Music Genre Classification with the Million Song Dataset*. 15-826 Final Report. Carnegie Mellon University.
3. Li, T. L. H., Chan, A. B., and Chun, A. H. W. 2010. Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*  (IMECS 2010, Hong Kong, March 17-19, 2010). http://www.iaeng.org/publication/IMECS2010/IMECS2010\_pp546-550.pdf
4. Tzanetakis, G., and Cook, P. 2002. Music Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*. Vol 10, No 5 (July 2002). https://www.researchgate.net/publication/3333877
5. Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= <http://doi.acm.org/10.1145/161468.16147>.
6. Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.
7. Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM, New York, NY, 526-531. DOI= <http://doi.acm.org/10.1145/332040.332491>.
8. Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.
9. Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.
10. Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.
11. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= <http://doi.acm.org/10.1145/964696.964697>.
12. Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= <http://dx.doi.org/10.1016/j.jss.2005.05.030>.
13. Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= <http://doi.acm.org/10.1145/90417.90738>.

Columns on Last Page Should Be Made As Close As Possible to Equal Length