Music Visualization Using Audio Features and Tags

Tzu-Chun Lin, Wei Lee Woon, Jimmy C. Peng

Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
{tlin, wwoon, jpeng}@masdar.ac.ae

Abstract. Two methods for clustering and visualizing a collection of music are implemented. The first method utilized audio feature sets while the second method was based on user-defined tags from *last.fm*. The focus of this paper is to find under what conditions the latter method can best match the result of the former. The Sammon Stress function was used to compare the two visualizations. A good correspondence was observed between the two schemes provided certain conditions are met: (i) selection of distinct musical genres and (ii) removal of spurious tags.

Keywords: Feature Extraction, Music Information Retrieval, User-defined Tags

1 Introduction

1.1 Novelty and Motivation

Broadband internet access and content sharing platforms have accelerated the dissemination of new music and stimulated research in music information retrieval. Music is often analyzed using audio features and listener defined "tags". To elucidate the relative strengths and weaknesses of these two methods, this paper presents a comparative study of visualizations of music tracks generated using audio features and tags collected from *last.fm*.

1.2 Definitions and Related Work

Audio features are numerical indicators extracted from a track using signal processing techniques. A deep discussion is beyond the scope of this paper but see for e.g. the seminal work in [1] and other examples presented in [2].

The analysis of music using tags is comparatively less well researched. We know of only one study focusing on the use of tags for music classification [3], though using tags for general clustering is more common [4]. However, tags are used extensively in commercial recommendation systems (including non-music related examples such as *Flickr*).

In the present context visualization refers to the use of machine learning techniques to generate intuitive and low-dimensional representations of

In MML 2013: International Workshop on Machine Learning and Music, ECML/PKDD, Prague, 2013.

complex data. Examples of applications of visualization in music can be found in [5,6].

2 Methodology

2.1 Data Collection

Last.fm is a popular music recommendation platform which supports tagging of tracks, artistes and albums. Importantly, an officially released API has been provided to support data collection. Near 4000 tracks from the "free downloads" section were selected for our study, spanning a range of genres.

2.2 Tag Similarity

The popularity of a track is strongly correlated to the number of tags it is labeled with. In our study, the 520 most popular songs with at least ten tags from thirty genres were chosen. The tags were then sorted based on the number of "clicks" received. The Jaccard index was then used to calculate tag similarity:

$$\mathrm{J}(\mathrm{A},\mathrm{B}) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Here, A and B represent the tag sets corresponding to each of the two songs.

2.3 Audio Feature Similarity

Two feature sets were extracted using 23ms windows (512 samples) with 50% overlap:

1. *jAudio* [7] - which is the feature extraction component of *jMIR* . A 79-dimension vector was extracted including RMS, spectral flux and MFCC.
2. *Musicminer* [8] – uses timbre distance based on frequency analysis to produce a 20-dimensional feature vector which included FFT, autocorrelation and phase domain transform.

Once extracted, the features were statistically standardized; the Euclidean distance measure was then applied to measure the similarity between tracks.

3 Results

One important assumption is that tags and audio features represent the same underlying "music space". This can be tested using the Sammon Stress function, which is often used for comparing topographical representations:

$$\mathrm{E} = \frac{1}{\sum_i d_{ij}^t} \sum_{j<i} \frac{\left(d_{ij}^t - d_{ij}^a\right)^2}{d_{ij}^t} \quad (2)$$

Where d_{ij}^t and d_{ij}^a represent tag and feature based similarities between tracks i and j respectively. The results of these calculations are shown in Tables 1 and 2 (lower numbers indicate a better match).

Table 1. Sammon Stress function for different number of top tags

	10 tags	9 tags	8 tags	7 tags	6 tags
jAudio	24.30	24.24	24.11	24.88	25.27
Musicminer	31.55	31.50	31.33	32.00	32.53

Table 2. Sammon Stress function for different genres (Only some are listed)

	Rock	Electronic	Classical	Metal	Latin	Hip-Hop
jAudio	32.48	27.55	1.54	28.83	7.37	17.74
Musicminer	28.90	26.52	2.33	29.99	15.34	12.53

The best performance was obtained using the *jAudio* feature set with the top 8 tags. This combination is used to generate a force directed visualization of songs from three genres: classical, Latin and hip-hop. To ensure that the sizes of the visualizations are manageable, only ten most popular tracks from each selected genre are included.

As can be seen (Fig. 1), tracks from the three genres form three distinct clusters. There were broad correspondences between the two visualizations. For example

1. In both cases there were overlaps between Latin and Hip-Hop while the classical tracks were better separated (classical music is relatively distinct from the other two genres).
2. While not identical, many local neighborhood relationships were preserved, for e.g. (L4,H5), (L8,H3), (C4,C1).

4 Conclusions and Future Work

It was shown that under appropriate conditions, both audio and tag information produced approximately compatible representations, indicating that both methods of analyzing music were compatible with a common underlying "music space". This is an important finding which will support a

variety of applications including automatic recommendation and music browsing systems.

There are many interesting avenues for future work but specific examples include extending the method to a larger range of genres and more optimal selection of feature subsets.

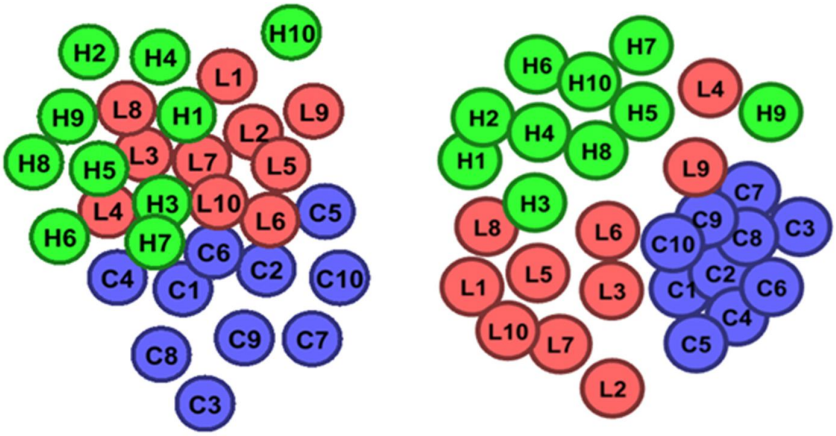

Fig. 1. Audio (left), Tag (right). Force-directed graph of: Classical(C), Latin(L), Hip-Hop(H)

References

1. Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5), 293-302.
2. Herrera, P., Bello, J., Widmer, G., Sandler, M., Celma, Ò.,Vignoli, F., ... & Serra, X. (2005, November). SIMAC: Semantic interaction with music audio contents. In *Integration of Knowledge, Semantics and Digital Media Technology*, 2005. EWIMT 2005. The 2nd European Workshop on the (Ref. No. 2005/11099) (pp. 399-406). IET.
3. Lehwark, P., Risi, S., &Ultsch, A. (2008). Visualization and clustering of tagged music data. In *Data Analysis, Machine Learning and Applications* (pp. 673-680). Springer
4. Begelman, G., Keller, P., &Smadja, F. (2006, May). Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland (pp. 15-33).
5. vanGulik, R., Vignoli, F., & van de Wetering, H. (2004). Mapping music in the palm of your hand, explore and discover your collection. In Proc. *5th International Conference on Music Information Retrieval*.
6. Vembu, S., & Baumann, S. (2005). A self-organizing map based knowledge discovery for music recommendation systems. In *Computer music modeling and retrieval* (pp. 119-129). Springer Berlin Heidelberg.
7. McKay, C., Fujinaga, I., &Depalle, P. (2005). *jAudio: A feature extraction library. In Proceedings of the International Conference on Music Information Retrieval* (pp. 600-3).
8. Moerchen, F., Ultsch, A., Thies, M., Loehken, I., Noecker, M., Stamm, C.,Kuemmerer, M. (2004). Musicminer: Visualizing perceptual distances of music as topograpical maps. Tech. Rep., Dept. Mathematics and Comp. Sci., University of Marburg, Germany