REAL-TIME SIMILARITY RETRIEVAL OF MUSIC, VOCALIZATIONS, AND ARBITRARY SOUND RECORDINGS

Dana Hughes
Computer Science Dept.
College of Charleston
Charleston, SC 29424, USA
hughesd@cs.cofc.edu

Bill Manaris
Computer Science Dept.
College of Charleston
Charleston, SC 29424, USA
manaris@cs.cofc.edu

Thomas Zalonis
Computer Science Dept.
North Carolina State Univ.
Raleigh, NC 27695, USA
tgzaloni@ncsu.edu

1. INTRODUCTION

This paper focuses on late-breaking results from a multi-year research project investigating Zipf's Law in the context of music information retrieval and data mining. This research project has produced Armonique (http://armonique.org), a music similarity engine, which automatically identifies aesthetic similarities in musical content [1]. Armonique utilizes hundreds of power-law metrics to extract statistical proportions of music-theoretic and other attributes of musical pieces. Evaluation experiments with artificial neural networks (ANNs) indicate that this approach works well with music (e.g., composer identification: 93.6% - 95% accuracy; style classification: 71.5% - 96.6% accuracy; pleasantness prediction: 90.7% accuracy). Additional psychological experiments, which compare Armonique's judgments to emotional and physiological responses of human subjects, validate the aesthetic similarity of retrieved pieces from a human listener perspective [2,3].[1]

2. APPROACH

We have implemented an efficient transcription algorithm for similarity retrieval of music, vocalizations, and arbitrary sound recordings. This audio-to-MIDI transcription algorithm handles polyphony; captures harmonic, vocal and percussive instrumentation, rather than a specific instrument or class of instruments; is very efficient; and works with non-musical signals, such as bird song and sub/ultrasonic animal vocalizations. This algorithm is based on the constant Q transform [4]. Its processing efficiency is achieved by capturing specific, discrete frequencies of interest, rather than the full spectral content of the signal. Also, since the transcription output is not intended for generation of musical scores, minimal filtering and post-processing is necessary (e.g., removal of redundant information, etc.). The generated MIDI output, although not score-perfect, is clearly recognizable and very effective for real-time similarity retrieval, using power-law features automatically extracted from it.

This transcription algorithm is incorporated into the Armonique system to facilitate real-time, content-based similarity retrieval for large audio collections. This produces a Google-like search engine, where users upload actual music as search queries. Furthermore, non-musical audio recordings, such as bird song or other, arbitrary audio may be used as search queries.

3. DEMONSTRATION

We will describe the transcription algorithm utilized with the Armonique framework for real-time audio analysis. We will also demonstrate a live version of Armonique. We will show how Armonique may be used to search through audio archives, e.g., full versions of songs on Magnatune (www.magnatune.com), as well as audio clips from 7digital (www.7digital.com). We will also discuss how it may be utilized as a framework for web audio archiving, searching and classification.

4. EVALUATION

We have carried out the following experiments, to assess the effectiveness of our MIR approach.

4.1 Million Song Dataset

We conducted two classification experiments with custom multi-genre corpora from the Million Song Dataset (http://labrosa.ee.columbia.edu/millionsong/). Using the provided meta-data, we automatically extracted a corpus of 2,400 songs across 8 genres, namely ambient, blues, classic-rock, classical, country, jazz, hip-hop, and techno.

To avoid genre overlap (a common problem with standard benchmarks), songs were selected if (a) they had a genre term with very high frequency (90-100%), and (b) they had none of the other genre terms (0% frequency). For example, a blues song was selected if the term “blues” was associated 90-100% with this song, and none of the other terms (i.e., “classic-rock”, “classical”, “country”, “jazz”,

[1] This work has been funded in part by NSF grants IIS-0736480, IIS-0849499 and IIS-1049554.

"hip-hop", and "techno") were associated with this song. Excerpts of these songs (30-sec clips) were downloaded from 7digital.com.

We ran several 10-fold cross-validation ANN experiments. The ANNs were trained using 300 MIDI metrics and 72 audio metrics (these metrics are described in [3]). The MIDI metrics were extracted from a MIDI representation generated by our transcription system.

4.1.1 Classification with Eight Genres

We ran an 8-genre classification experiment consisting of all 2,400 songs in our corpus (i.e., 300 per genre). The ANN achieved a success rate of 51.3% (which is high, compared to 12.5% for random selection). The ROC area values for each of the genres were: ambient 75.4%, blues 67.8%, classical 96.1%, classic-rock 77.1%, country 85.1%, hip-hop 88.6%, jazz 81.3%, and techno 89.8%. The average ROC value was 82.7%. (The ROC value corresponds to the accuracy of a binary-classification experiment involving one genre vs. all the others.)

Three genres (ambient, blues, and classic-rock) have a high degree of misclassifications, as can be seen from the confusion matrix (see Fig. 1). It should be noted that, up to this point in the experiment, we had not listened to the songs, as not to bias the experiment. Upon listening to the songs, it became evident that, as suggested by the confusion matrix, some genre terms (e.g., "ambient" and "blues") are less well-defined than other genre terms (e.g., "classical").

4.1.2 Classification with Five Genres

We also ran a 5-genre classification experiment consisting of 1,500 songs from classical, country, jazz, hip-hop, and techno genres (i.e., excluding the three less well-defined ones). The ANN achieved a success rate of 78.5% (compared to 20.0% for random selection). The ROC area values for each of the genres were: classical 97.9%, country 94.2%, hip-hop 92.9%, jazz 89.8%, and techno 93.6%. The average ROC value was 93.6%.

```
  a   b   c   d   e   f   g   h   <-- classified as
128  13  42  11  27  18  34  27 |   a = ambient
 47  53  24  31  23  36  74  12 |   b = blues
 24   0 259   3   0   0  14   0 |   c = classical
 59  27  15  55  53  33  41  17 |   d = rock
 44  11  13  25 158  14  28   7 |   e = country
 10   5   2   9   7 213  13  41 |   f = hip_hop
 26  24  46  14  13  16 155   6 |   g = jazz
 21   2   4   5   6  41  11 210 |   h = techno
```

Figure 1. Confusion matrix for 8-genre classification.

4.2 Multi-Year Song Sparrow Recordings

The second set of experiments involved a corpus extracted from multi-year recordings of song sparrows (Melospiza melodia) in Northeastern US habitats. This corpus was provided by Dr. Melissa Hughes, an expert in the evolution and function of bird song.

The corpus consists of 13 instances of three different songs (D, F, J) generated by a single male song sparrow (M1) recorded in 1998. These songs were labeled independently by the expert near the time of the recording.

For each song, we generated a MIDI transcription using our transcription system, and extracted 300 melodic metrics. We then carried out a 4-fold cross-validation experiment, using the extracted features. The ANN achieved a success rate of 84.6% and a ROC value of 94.6%.

Given the small corpus size, we also conducted a control experiment by randomizing the classes assigned to each song instance. This time the ANN reported a success rate of 7.6% and a ROC value of 26.5%.

5. CONCLUSION

The above results (ANN classification accuracies of 82.7%, 93.6%, and 94.6%) demonstrate the value of our approach for real-time similarity retrieval and audio data mining. For both corpora (i.e., music and animal vocalizations), we were able to automatically generate classifications comparable to human expertise. Additional assessment activities are being conducted.

6. REFERENCES

[1] B. Manaris, D. Krehbiel, P. Roos, T. Zalonis, "Armonique: Experiments in Content-Based Similarity Retrieval Using Power-Law Melodic and Timbre Metrics," *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, pp. 343-348, Sep. 2008.

[2] B. Manaris, J. Romero, P. Machado, D. Krehbiel, T. Hirzel, W. Pharr, and R.B. Davis, "Zipf's Law, Music Classification and Aesthetics," *Computer Music Journal* 29(1), pp. 55-69, 2005.

[3] B. Manaris, P. Roos, D. Krehbiel, T. Zalonis, and J.R. Armstrong, "Zipf's Law, Power Laws and Music Aesthetics", in T. Li, M. Ogihara, G. Tzanetakis (eds.), *Music Data Mining*, pp. 169-216, CRC Press - Taylor & Francis, July 2011.

[4] J.C. Brown and M.S. Puckette, "An efficient algorithm for the calculation of a constant Q transform, J. Acoust. Soc. Am., 92(5):2698–2701, 1992.