

A brief overview of music similarity analysis techniques

Charlotte Curtis

May 26, 2008

Introduction

A digital music collection stored on a personal computer or a central server has several advantages over its analogue counterpart. One of these advantages is the ability to organize and retrieve music based on a variety of features, such as metadata or web-based social music services such as last.fm. However, these methods depend on external factors such as accurate metadata or an internet connection.

The field of Music Information Retrieval (MIR) has emerged in the past few years to develop new methods of interacting with music. By analyzing the audio data itself to extract various features, retrieval methods such as “Query-by-Humming” are possible (identification of a song through humming the melody) [1]. Another interesting application of audio-based MIR is the detection of musical similarity for the purposes of generating mood-based playlists, or for discovering new artists that match a certain style.

Musical similarity is an intuitively simple task yet considerably more difficult to define mathematically. The feature most frequently associated with similarity is timbre, but other features such as rhythm and harmony can also be used [2, 3]. These features can be combined to generate playlists to suit the user’s mood, maintain a beat at a party, or simply rediscover music hidden in a large collection.

Preprocessing

Since the majority of digital music collections are in a compressed data format such as mp3, ogg, or aac, the audio must first be converted to a raw data format such as Pulse Code Modulation (PCM) [4]. In ad-

dition, since the music must be simply recognizable for feature extraction rather than high quality, it has been determined that mixing stereo signals to mono and down-sampling to 11 kHz does not significantly affect accuracy, yet greatly increases performance [4].

Feature Extraction

As research in the field of Music Information Retrieval has advanced, several general algorithms have emerged as popular and functional techniques.

Timbre

The basic algorithm for extracting a timbre feature vector is as follows [2]:

1. Split the audio signal into short overlapping frames
2. Compute a vector of Mel Frequency cepstrum Coefficients (MFCCs) for each frame
3. Compute a statistical model of MFCC distribution

This general algorithm leaves a lot of room for parameter variation, such as the length and amount of overlap of the frames, the length of the MFCC vectors, and the method of statistical modeling [2].

Rhythm Patterns

Rhythm is a difficult feature to quantify mathematically, yet is intuitively obvious to discriminating listeners [5]. In work done by Lidy et Al. in [6, 4], an

expanded definition of rhythm patterns was developed to describe amplitude modulations in various frequencies. The basic algorithm is as follows [6]:

1. Split the track into short segments
2. Use a short time Fast Fourier Transform to compute the energy per frequency bin (spectrum) to generate a spectrogram
3. Sum the frequency bins of the spectrogram to 24 critical bands of the Bark scale
4. Transform the data into the decibel scale to obtain a measure of specific loudness for each frequency band
5. Apply another Fourier Transform to compute the spectrum of the modulation signal and remove the time dependence
6. Apply a band-pass filter to remove features at frequencies that are not relevant to humans' sensation of rhythm
7. Average the feature vectors from each 6-second segment

While parameters from this algorithm may be varied, the work done in [6, 4] uses a segment length of 6 seconds (using every 3rd sample), cut-off frequencies of 0.168 and 10 Hz, and accentuating values which were determined to be most important to human rhythmic sensation (around 4 Hz). The final feature vector for each song contained 1440 elements.

Similarity Calculations

The feature vectors produced for each song represent a vast amount of data. In order for these data to be compared on-the-fly for dynamic playlist generation, efficient comparison algorithms must be employed. Various techniques are adopted, most commonly Earth Mover's Distance (EMD) [2], asymptotic likelihood approximation [3], or Monte Carlo sampling [7].

Optimization

Each step in the process of determining musical similarity requires a significant amount of computational resources. A variety of parameters can be varied in order to achieve increased performance without significantly affecting accuracy. In [8], the reconstruction of the PCM signal was skipped when extracting features from mp3's, resulting in a faster preprocessing step. In [7], each of the parameters in the process of extracting and comparing timbre information was varied and timed, resulting in an order of magnitude decrease in computation time.

Speed is a key component to the MIREX conference (<http://www.music-ir.org>); in order for a system to be appealing to the user, it must be both accurate and efficient. However, the speed of one component may be more important than others; for example, at MIREX 2006, the fastest algorithm for feature extraction was developed by [9], whereas the fastest comparison algorithm was developed by [10]. In most cases, speed of comparison may be most important, as feature extraction is only performed once.

Evaluation

In order to evaluate the effectiveness of any algorithm, a correct value, or gold standard, should be used. Unfortunately, it is difficult to determine a gold standard for music similarity, as it is a subjective measure. That said, empirical studies suggest that there is little variation between subjects when asked to evaluate the similarity of two songs [9]. These results indicate that subjective evaluation may be a reasonable indicator of similarity; indeed, more accurate than assuming that a given artist will consistently produce similar music, as is used in [11].

Nomenclature

Bark scale a psychoacoustic scale covering the human audible range

EMD measures the cost of moving one distribution to another

harmony combination of notes forming a characteristic chord

Mel scale a psychoacoustic scale covering the human audible range

MFCC coefficients obtained after warping the cepstrum (inverse Fourier transform of the log-spectrum) onto the Mel-frequency scale

timbre character/quality of musical sound

References

- [1] R. Typke, F. Wiering, and R. C. Veltkamp, “A survey of music information retrieval systems,” *Proceedings of the International Conference on Music Information Retrieval*, pp. 153–160, 2005.
- [2] J. J. Aucouturier and F. Pachet, “Improving timbre similarity: How high is the sky,” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, pp. 1–13, 2004.
- [3] N. Scaringella, G. Zoia, and D. Mlynek, “Automatic genre classification of music content,” *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 133–141, 2006.
- [4] A. Rauber, E. Pampalk, and D. Merkl, “The som-enhanced jukebox: Organization and visualization of music collections based on perceptual models,” *Journal of New Music Research*, vol. 32, no. 2, pp. 193–210, 2003.
- [5] I. Mierswa and K. Morik, “Automatic feature extraction for classifying audio data,” *Machine Learning*, vol. 58, no. 2, pp. 127–149, 2005.
- [6] T. Lidy, G. Polzlbauer, and A. Rauber, “Sound resynthesis from rhythm pattern features—audible insight into a music feature extraction process,” *Proceedings of the International Computer Music Conference*, pp. 93–96, 2005.
- [7] E. Pampalk, “Speeding up music similarity,” *2nd Annual Music Information Retrieval eXchange, London*, 2005.
- [8] D. Schnitzer, “mirage,” Master’s thesis, Technische Universität Wien, 2007.
- [9] E. Pampalk, “Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns,” tech. rep., 2006.
- [10] T. Lidy and A. Rauber, “Mirex 2006 computing statistical spectrum descriptors for audio music similarity and retrieval,” 2006.
- [11] B. Logan, “Music recommendation from song sets,” *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pp. 425–428, 2004.