CONTENT-BASED MUSIC SIMILARITY SEARCH AND EMOTION DETECTION

Tao Li, Mitsunori Ogihara

Department of Computer Science
University of Rochester
Rochester, NY 14627-0226

ABSTRACT

This paper investigates the use of acoustic based features for music information retrieval. Two specific problems are studied: **similarity search** (searching for music sound files similar to a given music sound file) and **emotion detection** (detection of emotion in music sounds). The Daubechies Wavelet Coefficient Histograms (proposed by Li, Ogihara, and Li), which consist of moments of the coefficients calculated by applying the Db8 wavelet filter, are combined with the timbral features extracted using the MARSYAS system of Tzanetakis and Cook, to generate compact music features. For similarity search, the distance between two sound files is defined to be the Euclidean distance of their normalized representations. Based on the distance measure the closest sound files to an input sound file is obtained. Experiments on Jazz vocal and Classical sound files achieve a very high level of accuracy. Emotion detection is cast as a multiclass classification problem, decomposed as a multiple binary classification problem, and is resolved with the use of Support Vector Machines trained on the extracted features. Our experiments on emotion detection achieved reasonably accurate performance and provided some insights on future work.

1. INTRODUCTION

Music is not only for entertainment or pleasure. Social and psychological effects of music have been studied extensively for decades. At the beginning of the 21st century the world is witnessing ever-increasing growth of the on-line music. Efficient and accurate automatic music information processing (in particular, accessing and retrieval) will be extremely important research issues of this century. Traditionally musical information has been retrieved and/or classified based on standard reference information, such as the name of the composer, the title of the work, the album title, the style of music, and so on. These basic pieces of information will remain essential, but information retrieval based purely on these is far from satisfactory. In [1] Huron points out that since the preeminent functions of music are social and psychological, the most useful characterization would be based on four types of information: the style, emotion, genre, and similarity

Of these four types, there has been a considerable amount of work in extracting features for speech recognition and music-speech discrimination, but much less work has been reported on the development of descriptive features specifically for music signals. To the best of our knowledge, currently the most influential approach to direct modeling of music signals for automatic genre classification is due to Tsanetakis and Cook [2], where the timbral texture, rhythm, and pitch content features are explicitly developed. The accuracy of classification based on these features, however, is only 61% on their ten-genre sound dataset. In [3] we proposed a new feature extraction method based on wavelet coefficients histogram, *DWCH*. The DWCH features are computed from histograms on Daubechies wavelet coefficients at different frequency sub-bands with different resolutions and seem to represent both local and global information very well. In [3] it is shown that the use of DWCH together with advanced machine learning techniques, accuracy of music genre classification can be significantly improved. On the ten-genre dataset of [2], the accuracy of classification has been increased to almost 80%. On genre specific classification, i.e., distinguish one genre from the rest, the accuracy can be as high as 98%. The paper [3] shows that the best way of using DWCH is to combined it with the timbral features. These results seem to suggest that genre classification can now be efficiently done using such feature sets as DWCH and motivate us to study the three remaining types of information as proposed by Huron. Of the three, style can be thought as a problem residing between similarity and genre, and thus, similarity and emotion are the most urgent issues in music information retrieval.

The objective of similarity search is to find music sound files similar to a given music sound file given as input. Music classification based on genre and style is naturally the form of a hierarchy. Similarity can be used to group sounds together at any node in the hierarchies. The use of sound signals for similarity is justified by an observation that audio signals (digital or analog) of music belonging to the same genre share certain characteristics, because they are composed of similar types of instruments, having similar rhythmic patterns, and similar pitch distributions [4].

Relationship between musical sounds and their impact on the emotion of the listeners has been well studied for decades. The celebrated paper of Hevner [5] studied this relation through experiments in which the listeners are asked to write adjectives that came to their minds as the most descriptive of the music played. The experiments substantiated a hypothesis that music inherently carries emotional meaning. Hevner discovered the existence of clusters of descriptive adjectives and laid them out (there were eight of them) in a circle. She also discovered that the labeling is consistent within a group having a similar cultural background. The Hevner adjectives were refined and regrouped into ten adjective groups by Farnsworth [6]. Our goal is to use treat the emotion detection problem as a multiclass classification problem.

The similarity search processes can be divided into *feature extraction* and *query processing* while the process of emotion contains *feature extraction* and *multi-label classification*. In the feature extraction step, we extract from the music signals information

Supported in part by NSF grants EIA-0080124, DUE-9980943, and EIA-0205061, and by an NIH grant P30-AG18254.

representing the music. The features extraction should be *comprehensive* (representing the music very well), *compact* (requiring a small amount of storage), and *effective* (not requiring much computation for extraction). To meet the first requirement the design has to be made so that the both low-level and high-level information of the music is included. In the second step, we build an efficient mechanism (an algorithm and/or a mathematical model) for processing the queries or classification based on the features.

2. RELATED WORK

The content-based acoustic features are classified into timbral texture features, rhythmic content features, and pitch content features [2]. Timbral features are mostly originated from traditional speech recognition techniques. They are usually calculated for every short-time frame of sound based on the Short Time Fourier Transform (STFT) [7]. Typical timbral features include Spectral Centroid, Spectral Rolloff, Spectral Flux, Energy, Zero Crossings, Linear Prediction Coefficients, and Mel-Frequency Cepstral Coefficients (MFCCs) (see [7] for more detail). Among these timbral features MFCCs have been dominantly used in speech recognition. Logan [8] examines MFCCs for music modeling and music/speech discrimination. Rhythmic content features contains information about the regularity of the rhythm, the beat and tempo information. Tempo and beat tracking from acoustic musical signals has been explored in [9, 10]. Foote and Uchihashi [11] use the beat spectrum to represent rhythm. Pitch content features deals with the frequency information of the music bands and are obtained using various pitch detection techniques.

There has been much work on music style recognition, genre categorization, and similarity. Dannenberg, Thom, and Watson [12] demonstrate that machine learning can be used to build effective style classifiers for interactive performance systems. Kuo and Shan [13] present a personalized content-based music filtering system to support music recommendation based on user's preference of melody style. Chai and Velcoe [14] use the Hidden Markov Model to classify songs based on their monophonic melodies. As to genre classification, Tsanetakis and Cook [2] propose a comprehensive set of features to directly model music signal and explored the use of those features for musical genre classification using the K-Nearest Neighbor Model and the Gaussian Mixture Model. Lambrou et al. [15] use statistical features in the temporal domain as well as three different wavelet transform domains to classify music into rock, piano and jazz. Deshpande, Singh, and Nam [16] use Gaussian Mixtures, Support vector machines and Nearest Neighbors to classify the music into rock, piano and jazz based on timbral features.

The problem of finding sound files similar to a given sound files has been studied in the past [17, 18, 19, 20, 21]. Logan and Salomon propose the use of MFCC to define similarity [18]. Nam and Berger propose the use of timbral features (spectral centroids, short-term energy function, and zero-crossing) for similarity testing [17]. Cooper and Foote study the use of self-similarity to summary music signals [19]. Foote, Cooper, and Nam use this summarization for retrieving music files [20]. Rauber, Pampalk, and Merkl study a hierarchical approach in retrieving similar music sounds [21].

While there has been much work on emotion recognition from speech [22, 23], there has been little work on automatic music emotion detection.

3. FEATURE EXTRACTION

Our extracted feature contains traditional sound features including MFCC and other timbral features and DWCHs.

3.1. Mel-Frequency Cesptral Coefficients (MFCC)

MFCC is designed to capture short-term spectral-based features. After taking the logarithm of the amplitude spectrum based on short-term Fourier transform for each frame, the frequency bins are grouped and smoothed according to Mel-frequency scaling, which is design to agree with perception. MFCC features are generated by decorrelating the Mel-spectral vectors using discrete cosine transform.

3.2. Other Timbral Features

Spectral Centroid is the centroid of the magnitude spectrum of short-term Fourier transform and is a measure of spectral brightness. *Spectral Rolloff* is the frequency below which 85% of the magnitude distribution is concentrated. It measures the spectral shape. *Spectral Flux* is the squared difference between the normalized magnitudes of successive spectral distributions. It measures the amount of local spectral change. *Zero Crossings* is the number of time domain zero crossings of the signal. It measures noisiness of the signal. *Low Energy* is the percentage of frames that have energy less than the average energy over the whole signal. It measures amplitude distribution of the signal.

3.3. DWCH

There are many kinds of wavelet filters, including Daubechies wavelet filter, Gabor filter etc. Daubechies wavelet filters are the one commonly in image retrieval (more details on wavelet filter can be found in [24]. In our work, we use Daubechies wavelet filter Db8 with seven levels of decomposition. After the decomposition, we construct the histogram of the wavelet coefficients at each subband. The coefficient histogram provides a good approximation of the waveform variations at each subband. From probability theory, a probability distribution is uniquely characterized by its moments. Hence, if we interpret the waveform distribution as a probability distribution, then it can be characterized by its moments. To characterize the waveform distribution, the first three moments of a histogram is used [25]. The first three moments are the average, the variance and the skewness of each subband. In addition, we also compute the subband energy, defined as the mean of the absolute value of coefficients, for each subband. In addition, our final DWCH feature set also includes the tradition timbral features for speech recognition.

Our DWCH feature set contains four features for each of seven frequency subbands along with nineteen traditional timbral features. However, we found that not all the frequency subbands are informative and we only use four subbands. The total number of features is 35. More details can be found in [3].

4. SIMILARITY SEARCH

4.1. Method Description

After feature extraction, we represent each music track M_i by a 35-dimension vector $V_i = (V_{i1}, \cdots, V_{i35})$. We normalize each dimension of the vector by subtracting the mean of that dimension

across all the tracks and then dividing the standard deviation. The normalized 35-dimensional representation vector is

$$\hat{V}_i = (\hat{V}_{i1}, \cdots, \hat{V}_{i35}),$$

where $\hat{V}_{ij} = \frac{V_{ij} - mean(V_{.j})}{std(V_{.j})}, 1 \leq j \leq 35$. After normalization, we compute the Euclidean distance between the normalized representation and the distance serve as similarity (in fact, dissimilarity) measure for our purpose. We then return the tracks with shortest distances to the given query as our similarity search result.

4.2. Experiments

4.2.1. Jazz Vocal Music

We created a collection of 250 Jazz vocal sounds files, covering 18 vocalists and 35 albums. For each music file, its first 30 seconds of the music were profiled into features using the approach described earlier.

Next, 60 tracks were selected from the collection as queries. For each query the nine closest matches were found, which were ranked in the increasing order of their Euclidean distance to the input sounds. Of the 60, 28 queries (46.7%) had a track from the same album as the closest match, 38 queries (63.3%) had at least one track from the same album in the top three matches, and 54 queries (90.0%) had at least one track from the same album in the top nine.

For each of the 22 queries for which the system selected no tracks belonging to the same album, at least one of the top three choices had sounds very close to the query. For example, the system selected a segment from a ballad with a low-range female voice (S. Vaughan) accompanied by a piano trio as the most similar to a ballad with a low-range male voice (J. Hartman) accompanied by a piano trio; the system found the husky voice of B. Dearie to be similar to the husky voice of K. Krog.

4.2.2. Classical Music

We created a collection of 288 Classical sound files, covering 72 albums (15 orchestral, 10 chamber, six songs and lieder, ten instrumental, ten string solo and ensemble, seven choral, six opera, and eight concerto albums). We selected a track from each album to obtain a list of nine closest sound files in the entire collection. For 33 queries (45.3%) the top two selections contained a track from the same album, for 29 of the remaining 39 (41.3% of the total), at least three out of top five were of the same format and from the same period (one of baroque, classical–romantic, and contemporary). Thus, for a total of 62 out of 72 (86%), the tracks identified were highly satisfactory.

5. EMOTION DETECTION

We cast the emotion detection problem as a *multi-label classification problem*, where the music sounds are classified into multiple classes simultaneously. That is a single music sound may be characterized by more than one label, e.g. both "dreamy" and "cheerful."

5.1. The Multi-label Classification Method

We resort to the scarcity of literature in multi-label classification by decomposing the problem into a set of binary classification problems. In this approach, for each binary problem a classifier is developed using the projection of the training data to the binary problem. To determine labels of a test data, the binary classifiers thus develop are run individually on the data and every label for which the output of the classifier exceeds a predetermined threshold is selected as a label of the data. See [26] for similar treatments in the text classification domain.

To build classifiers we used Support Vector Machines [27] (SVM for short) Based on the theory of structural risk minimization, SVMs are designed to learn a decision boundary between two classes by mapping the training examples onto a higher dimensional space and then determining the optimal separating hyperplanes between that space. SVMs have shown superb performance on binary classification tasks and has been widely used in many applications. Our SVM implementation is based on the LIBSVM [28], a library for support vector classification and regression.

5.2. The Dataset and Emotional Labeling

A collection of 235 Jazz sound files was created from 80 Jazz instrumental albums as follows: From each album the first four tracks were chosen (some albums had less than three music tracks in the first four). Then from each music track the sound signals over a period of 30 seconds after the initial 30 seconds were extracted in MP3.

The files were labeled independently by two subjects: a 39 year old male (subject 1) and a 25 year old male (subject 2). Each track was labeled using a scale ranging from -4 to $+4$ on each of three bipolar adjective pairs: (Cheerful, Depressing), (Relaxing, Exciting), and (Comforting, Disturbing), where 0 is thought of as neutral. Our early work on emotion labeling [29] used binary label (existence versus non-existence) based on the ten adjective groups of Farnsworth. The classification accuracy was not impressive was not very high (around 60%). This low performance may be due to the fact that there were so many labels to choose from. The recent experiments conducted by Moleants and his group [30] using scales on ten bipolar adjective pairs produced suggest that variations in emotional labeling can be approximated using only spanned three major principal components, which are hard to name. With these results in mind we decided to generate three bipolar adjective pairs based on the eight adjective groups of Hevner.

5.3. Experiments

The accuracy of the performance is presented in Table 1. Here the accuracy measure is the Hamming accuracy, that is, the ratio of the number of True Positives and True Negative against the total number of inputs. In each measure, the tracks labeled 0 are altogether put on either the positive side or the negative side. It is clear that the accuracy of detection was always at least 70% and sometimes more than 80%. Also, there is a large gap in the performance between the two subjects on the first two measures. We observe that this difference is coming from the difference in the cultural background of the subjects. To deal with labeling of a much larger group of listeners one should cluster them into groups depending on their labeling and train the emotion detection system for each group.

Accuracy	Cheerful	Relaxing	Comforting
Subject 1	0.833 (0.080)	0.704 (0.099)	0.724 (0.051)
Subject 2	0.696 (0.100)	0.837 (0.073)	0.709 (0.091)

Table 1. Experimental Results. The quantity within the parentheses is the standard deviation of the corresponding labeling.

6. CONCLUSIONS

This paper studied the problem of finding similar music sound files and detecting emotions based on the acoustic features calculated from 30 seconds of music signals using FFT and Wavelet transforms. For similarity search, the preliminary experiments conducted on Jazz vocal tracks and on classical tracks achieved more than 86% of perceived accuracy, in the sense that tracks sounding very similar to listeners are found. For emotion detection, our experiments show that emotion detection is harder than similarity, the accuracy values ranging between 70% and 83%. Our future goals are: to carefully create a data collection, to include emotional contents and lyrics in similarity search, and to mix these pieces of information to obtain better accuracy.

7. REFERENCES

[1] D. Huron, "Perceptual and cognitive applications in music information retrieval," in *International Symposium on Music Information Retrieval*, 2000.

[2] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–298, July 2002.

[3] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *SIGIR'03*. 2003, pp. 282–289, ACM Press.

[4] W. J. Dowling and D. L. Harwood, *Music Cognition*, Academic Press, Inc, 1986.

[5] K. Hevner, "Experimental studies of the elements of expression in music," *American Journal of Psychology*, vol. 48, pp. 246–268, 1936.

[6] P. R. Farnsworth, *The social psychology of music*, The Dryden Press, 1958.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, NJ, 1993.

[8] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symposium on Music Information RetrievalISMIR*, 2000.

[9] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," in *ACM Multimedia*, 1994, pp. 365–372.

[10] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of the Acoustical Society of America*, vol. 103, no. 1, 1998.

[11] J. Foote and S. Uchihashi, "The beat spectrum: a new approach to rhythm analysis," in *IEEE International Conference on Multimedia & Expo 2001*, 2001.

[12] R. Dannenberg, B. Thom, and D. Watson, "A machine learning approach to musical style recognition," in *International Computer Music Conference*, 1997, pp. 344–347.

[13] F. Kuo and M. Shan, "A personalized music filtering system based on melody style classification," in *IEEE International Conference on Data Mining*, 2002, pp. 649–652.

[14] W. Chai and B. Vercoe, "Folk music classification using hidden Markov models," in *International conference on Artificial Intelligence*, 2001.

[15] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing (ICASSP-98)*, 1998, vol. 6, pp. 3621–3624.

[16] H. Deshpande, R. Singh, and U. Nam, "Classification of music signals in the visual domain," in *Proceedings of the COST-G6 Conference on Digital Audio Effects*, 2001.

[17] U. Nam and J. Berger, "Addressing the same but different-different but similar problem in automatic music classification," 2001.

[18] B. Logan and A. Salomon, "A content-based music similarity function," Tech. Rep. CRL 2001/02, Cambridge Research Laboratory, June 2001.

[19] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, 2002, pp. 81–85.

[20] J. Foote, M. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, 2002, pp. 265–266.

[21] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarities," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, 2002, pp. 71–79.

[22] V. A. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of the Artificial Neural Networks In Engineering '99*, 1999.

[23] T. Polzin and A. Waibel, "Detecting emotions in speech," in *Proceedings of the CMC*, 1998.

[24] I. Daubechies, *Ten lectures on wavelets*, SIAM, Philadelphia, 1992.

[25] A. David and S. Panchanathan, "Wavelet-histogram method for face recognition," *Journal of Electronic Imaging*, vol. 9, no. 2, pp. 217–225, 2000.

[26] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[27] V. N. Vapnik, *Statistical learning theory*, John Wiley & Sons, New York, 1998.

[28] C. Chang and C. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[29] T. Li and M. Ogihara, "Detecting emotion in music," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, 2003, To appear.

[30] D. Moleants, 2003, Personal communication.