Unsupervised Classification of Music Genre Using Hidden Markov Model
(Content Based Music Structure Analysis)

Xi Shao#*, Changsheng Xu#, Mohan S Kankanhalli*

#Institute for Infocomm Research, 21 Heng Mui Keng Terrace Singapore 119613
{shaoxi,xucs}@i2r.a-star.edu.sg
*School of Computing, National University of Singapore
mohan@comp.nus.edu.sg

Abstract

Music genre classification can be of great utility to musical database management. Most of current classification methods are supervised and tend to be based on contrived taxonomies. However, due to the ambiguities and inconsistencies in the chosen taxonomies, these methods are not applicable for much larger database. In this paper, we proposed an unsupervised clustering method based on a given measure of similarity which can be provided by Hidden Markov Models. In addition, in order to better characterize music content, a novel segmentation scheme is proposed based on music intrinsic rhythmic structure analysis and features are extracted based on these segments. The performance of this feature segmentation scheme performs better than the traditional fixed-length method according to experimental results. Our preliminary results also suggest that proposed method is comparable to supervised classification method.

1. Introduction

The ever increasing wealth of digitalized music on the Internet calls for an automated organization of music material. Music genre is an important description that can be used to classify and characterize music from different sources such as music shops, broadcasts and Internet. It is very useful for music indexing and content-based music retrieval. For human being, it is not difficult to classify music into different genres. Although to let computers understand and classify music genre is a challenging task, there are still perceptual criteria related to the melody, tempo, texture, instrumentation and rhythmic structure that can be used to characterize and discriminate different music genres.
A number of methods have been proposed to discriminate music, speech, silence, and environment sound [1]. It is extremely more difficult to discriminate music genres than discriminate music, speech and other sounds. Several researches focus music genre classification on MIDI. Chai and Varcoe used HMMs to classify four different symbolic representations of folk music from three Western European countries [2]. However, MIDI is a structured format, so it is easy to extract features according to its structure. Actual sounds such as wav and mp3 are different from MIDI, thus MIDI style classification is not practical in real applications. Soltau [3] classified music into rock, pop, techno and classic using HMM and ETMNN to extract the temporal structure from the sequence of cepstral coefficients. Pye [4] used Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture model (GMM) to classify music into six types of blues, easy listening, classic, opera, dance and rock. Tzanetakis[5] explored timbral texture, rhythmic and pitch feature extraction. GMM and K-nearest neighbor (KNN) classifiers were used on the extracted features. All above classification methods are supervised. To classify music genres, generally a large number of training examples for each genre must be collected and labeled. This is a labor-intensive and error-prone process and is only feasible for a limited set of genres. Therefore, unsupervised music genre classification method needs to be investigated. Pachet [6] suggested using similarity measures based on audio signals as well as on cultural similarity gleaned from the application of data mining techniques to text documents, which differed with previous method in emerging classifications according to some similarity measure. But it works only for title and artist name appearing in the music sources.

To our best knowledge, there is no unsupervised music genre classification method proposed so far. In this paper, we present a novel unsupervised approach for automatic music genre classification. Our approach contains two steps. In the first step, as Figure 1 shows, every individual music piece is segmented into clips, and each clip is further segmented according to its intrinsic rhythmic structure. Features are extracted based on these segments. Then we train a Hidden Markov Model (HMM) for this music piece based on these features. In the second step, we embed the distance between every pair of music pieces (HMMs) into a distance matrix and perform clustering to generate desired clusters.

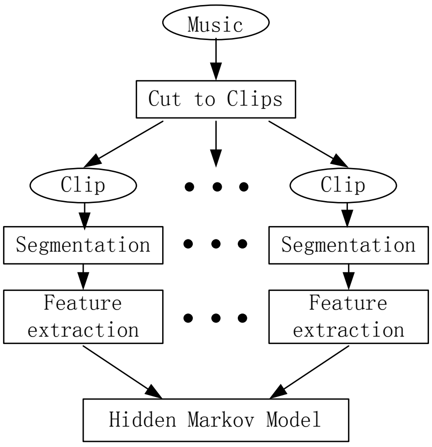
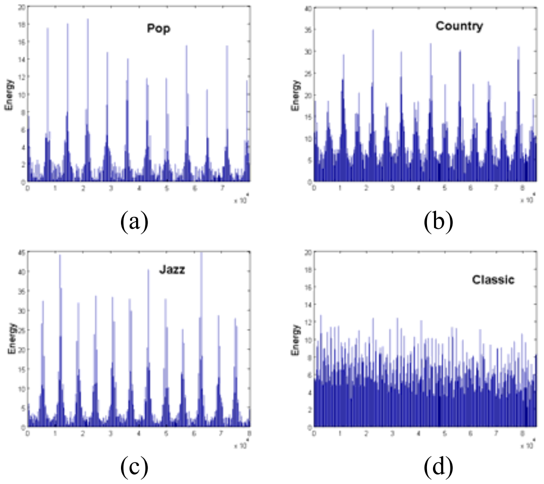


Figure 1. HMM training for individual music piece

2. Feature Selection

In order to better discriminate the different genres of music, we consider segmenting the music clip according to its intrinsic rhythmic structure. There are two reasons for using this segmentation scheme. Firstly, compared with the fixed length segmentation for music clips, segmenting music clips according to its intrinsic rhythm captures the natural structure of music genres better. Secondly, rhythmic structure characterizes the movement of music piece over time and contains such information as the regularity of the rhythm, beat, tempo, and time signature. These salient periodicities contain obvious time-sequential information which can be readily modeled by the HMMs. The different rhythmic structures for different genres are illustrated in Figure 2. The horizontal axis represents the sample index and the vertical axis represents the energy after autocorrelation. It can be seen that Pop, Country and Jazz are highly structured music, and the inter-beat-interval, which is defined as the temporal difference between two successive beats, is almost a constant for a particular piece of music. However, the rhythmic structure varies for these three genres. As for Classic music, it is not a so highly structured music and the inter-beat-interval varies from time to time, which distinguishes it from other three genres. Our proposed rhythm tracking and extraction approach can be found in [7].



(a) (b)

(c) (d)

Figure 2. Rhythmic structures for different genres

After the music clips have been segmented according to inter-beat-interval, three types of the features are extracted for each segment.

2.1. Mel-frequency cepstral coefficient

The mel-frequency scale was originally developed in phonetics to help model the non-linear nature of human auditory system and has proven to be highly effective in recognizing structure of music signals. The mel-cepstral features can be illustrated by the Mel-Frequency Cepstral Coefficients (MFCCs) [8] and it have also been proven useful in music genre classification [4].

2.2. Linear prediction coefficients (LPC) derived cepstrum coefficient

Linear predictive analysis is based on a model of the vocal tract as an all-pole filter [8]. The LPC coefficients are a short time measure of speech signal as the output of this all-pole filter. Although it has been designed to model speech production, it is also partially valid for musical instruments. In this case, the filter embodies the effect of resonating body of instrument, namely, its timbre.

An alternative feature for LPC coefficient is LPC derived cepstra coefficient, which can be computed simply as:

$$c_n = -a_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i c_{n-i} \qquad (1)$$

where a_n is the LPC coefficient.

The principal advantage of LPC derived cepstral coefficients is that they are generally decorrelated and this allows diagonal co-variances to be used in the HMMs.

2.3. Delta and acceleration

Delta Values [Δ ($\boldsymbol{V}_i$)] and acceleration values [acc ($\boldsymbol{V}_i$)] can be appended to any feature vector $\boldsymbol{V}_i$. They are computed as $\Delta(\boldsymbol{V}_i) = \boldsymbol{V}_i - \boldsymbol{V}_{i-1}$ and $\mathrm{acc}(\boldsymbol{V}_i) = \Delta(\boldsymbol{V}_i) - \Delta(\boldsymbol{V}_{i-1})$. Delta and acceleration values are very important improvements in feature extraction for HMMs because they effectively increase the state definition to include first and second order memory of past states.

3. Clustering by Hidden Markov Models

Our task is to classify observed low-level audio features into different music categories. Unlike most classic pattern classification problems, the data to be classified here are time series data. To handle this problem, Hidden Markov Model [9] [10] is used. It can be completely defined by the number of hidden states, a static state transition probability distribution **A**, the observation symbol probability distribution **B** and the initial state distribution $\boldsymbol{\pi}$. We can define one HMM model as $\lambda = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$.

Once the model topology and observation (training) vectors are determined, parameter estimation for the HMM is done using Baum-welch algorithm [8].

As for clustering, an important issue is how to measure the similarity of samples. HMMs provide a proper distance metric for sample comparison. The distance between two samples is defined as:

$$D(O^{(1)}, O^{(2)}) = \frac{\frac{1}{N_1}[\log P(O^{(1)} \mid \lambda_1) - \log P(O^{(2)} \mid \lambda_1)]}{2} + \frac{\frac{1}{N_2}[\log P(O^{(2)} \mid \lambda_2) - \log P(O^{(1)} \mid \lambda_2)]}{2} \quad (2)$$

where $O^{(1)}$ =($o_1 o_2 \cdots o_{N_1}$) is a sequence of observations generated by HMM model λ_1 and $O^{(2)}$ is generated by HMM model λ_2 . N_1 and N_2 are the length of $O^{(1)}$ and $O^{(2)}$ separately. The detail interpretations of Eq.(2) can be found in [8].

In our experiment, initially, we build a HMM model for each music piece. Assume there are N pieces of music in the database, then the distance between two music pieces can be calculated by Eq.(2) and the distance matrix **D** is $N \times N$ dimension. Given a distance matrix **D**, many clustering methods can be used. In this paper, we use *Agglomerative Hierarchical Clustering* [11] to generate clusters. This method does bottom-up clustering. It starts with N singleton clusters and forms a sequence of clusters by successive merging. For the scenario where the number of desired clusters is known (denoted as C), the merging process will be stopped when the C clusters are generated. While for the number of desired clusters is unknown, the merging process will be stopped when the distance between two nearest clusters is above a threshold.

4. Experimental Results

To test our proposed approach, we carried out an experiment on various genres of music pieces.

4.1. Data collection

The music dataset for each genre contains 50 music pieces. The genres are Pop, Country, Jazz and Classic. They are collected from music CDs and internet. All data are 44.1k Hz sample rate, mono channels and 16 bits per sample.

4.2. Classification results

As mentioned in previous section, for each music piece, we split it to 30 second clips. Using these clips as training data, a continuous-input HMM template is created for each music piece with random initial parameters. Each state's observation distribution is modeled by a single Gaussian with 36 dimensional mean and 36 by 36 diagonal variance for MFCC (6) and LPC cepstra (6) features supplemented by delta and acceleration values. Hidden state number is varied between 3, 4, 5 states. In our experiment, we found that the number of hidden states did not have dramatic impact on the system in terms of classification accuracy.

Table 1 illustrates the classification results using proposed method with 5-state HMMs. The column titles represent actual genre, while the row titles represent classification assigned by the system.

Table 1. 5-state HMM Classification Results

	Pop	Country	Jazz	Classic
Pop	**88%**	0%	12%	0%
Country	0%	**92%**	8%	0%
Jazz	20%	4%	**76%**	0%
Classic	0%	0%	0%	**100%**

It can be seen that some types of music have proven to be more difficult to classify than others. In particular, Jazz has proven to be difficult to distinguish from Pop music. It probably results from the fact that jazz music usually comprises the improvisation of the musicians, producing variations in most of the parts, which makes it similar to Pop music. Classic music has proven to be the easiest to classify. This makes intuitive sense because Classic is most different from the other genres.

For comparison, we use a fixed-length segmentation scheme with 20 ms time window and 50% overlapping to segment the music clips. As Table 2 shows, the average classification accuracy is 75% using the same datasets and HMM topology, which is far below that of our proposed segmentation scheme.

We also compare the performance of proposed method with other supervised learning classification method such as SVM classifier, as described in our previous paper [12]. It was adopted because it yielded the best classification results among all supervised learning classifier. On the same dataset, as Table 2 shows, our proposed method is comparable to the SVM classifier. However, for SVM classifier, two problems make it not applicable to the real world application. Firstly, from the music data point of view, SVM classifier is based on contrived taxonomies. It is not applicable to very large databases due to the ambiguities and inconsistencies in the chosen taxonomies. Secondly, from the classifier point of view, addition of new genre necessitates retraining all SVM classifiers. It is time-consuming work due to the slow training speed of SVM, especially when the genre hierarchy grows large.

Table 2. Comparison Result

	Proposed Method	Fix-length Segmentation	SVM Classifier
Average Accuracy	89%	75%	93%

5. Conclusions and Future Work

We have presented how to use HMMs for unsupervised classification of music genres. We have also illustrated that segmenting music sample according to its intrinsic rhythm structure outperforms the fixed-length segmentation scheme, at least in training the HMM.

There are two directions that need to be investigated in the future. First, we need to explore more music features that can be used to characterize music contents because the better feature set can improve the performance dramatically. The second direction is to scale-up unsupervised classification to real world application. We will explore the possibility of combining our classification method with SVM, to use the strengths of both. For example, our proposed method could be employed to initially classify into broad, strongly different categories, and SVM could then be employed to classify finely narrow subcategories. This would partially solve the problem of fuzzy boundaries between genres and could lead to better overall results.

6. References

[1] L. Lu, H. Jiang and H. J. Zhang, “A Robust Audio Classification and Segmentation Method”, In *Proc. ACM Multimedia 2001*, Ottawa, Canada, 2001.

[2] W. Chai and B. Vercoe, “Folk Music Classification Using Hidden Markov Models”, In *Proc. of IC-AI01*, 2001.

[3] H. Soltau, T. Schultz, M. Westphal and A. Waibel, “Recognition of Music Types”, In *Proc. of IEEE ICASSP98*, pp.1137-1140, 1998.

[4] D. Pye, “Content-Based Methods for the Management of Digital Music”, In *Proc. of IEEE ICASSP00*, pp.2437-2440, 2000.

[5] G. Tzanetakis, P. Cook, “Musical Genre Classification of Audio Signals”, *IEEE Transactions on Speech and Audio Processing*, Vol.10, No.5, July 2002.

[6] F. Pachet,., G. Westermann, & D. Laigre. Musical Datamining for EMD. In *Proceedings of Wedel Music Conference*, Italy, 2001.

[7] N. C. Maddage, A. Shenoy, C. S. Xu and Y. Wang, “Semantic Region Detection in Acoustic Music Signal”, submitted to *International Conference of Acoustic Speech Signal Processing*(*ICASSP*) 2004.

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

[9] L. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE,* 77:2(257-286), 1989.

[10] S. Young, etc. *The HTK Book* (*for HTK Version 3.2*).*http://htk.eng.cam.edu/.* Cambridge University Engineering Department, December 2002.

[11] R. O. Duda, P. E. Hart and D .G . Stork, *Pattern Classification*, *second edition*; A Wiley-Interscience Publication, 2000.

[12] C. Xu, N. C. Maddage, X. Shao, etc, "Musical Genre Classification Using Support Vector Machines", In *Proc. of IEEE ICASSP03*, pp.V429-V432, 2003.