An Audio Recommendation System Based on Audio Signature Description Scheme in MPEG-7 Audio

Yao-Chang Huang and Shyh-Kang Jeng
Graduate Institute of Communication Engineering
National Taiwan University, Taipei 10617, Taiwan, R.O.C.
Taipei,Taiwan, ROC
skjeng@ew.ee.ntu.edu.tw

Abstract

In this paper, we propose an audio recommendation system based on audio signature in MPEG-7 Audio. The procedure of our system is as follows. First, the user assigns each song a rating value in his collected songs and the system extracts the audio signature as the music features from audio data. The system then uses LBG Vector Quantization approach to rate a new music. Experiments show that the audio signature is a useful music feature for audio music recommendation.

1. Introduction

In recent years, the Automatic Recommendation System (ARS) about multimedia has become more and more important. Since the traditional text-based ARS is of limited use for multimedia database, the content-based search of ARS multimedia is becoming popular.

The ARS broadly falls into three categories, content-based filtering, collaborative filtering and link-based system [1]. Our system is a content-based filtering system, which uses machine learning techniques to analyze the data content, attempting to find out the user's preference. The personal profile will be built by the system, and the system will take this characteristic to rate other music. Related works on the ARS for music have been discussed in [2], [3].

Midi data are often used in ARS [2], [3] because the midi format contains lots of musical features, such as pitch, duration and velocity, which can be easily extracted. However, for most music consumers, wave and mp3 format are also common. In addition, midi loses more music information than Wav/ MP3 format. One music piece with midi format may not be seen as the same with the one with wav format. Thus an ARS based on features extracting from music signals directly is also necessary and crucial.

Our system analyzes the audio data by the scalable audio signature defined in MPEG-7 Audio [4], which calculates the flatness of the audio spectrum. This approach has the following advantages.

First, MPEG-7 is an international standard, and it provides a unified interface for all audio format. Besides, it is based on XML-Schema, and is to be widely adopted widespread in Internet in the future.

Second, the audio signature has the flexibility between the similarity matching speed and precision, depending on the system needed [5].

In the following section, we explain why we use audio signature as an audio feature. Section 3 illustrates the basic structure of our recommendation system, and Section 4 presents the experiment results. The last section draws conclusions.

2. Audio signature

MPEG-7, as implied in its formal name, "Multimedia Content Description Interface", standardizes the metadata accompanied with media content. In MPEG-7, two elements are defined: The audio-visual features extracted from metadata are defined in a standard and these features are called "Descriptor" (D). The other is high level Description Schemes (DS), which are associated with one descriptor or a set of descriptors and description schemes, such as Melody, Timbre, Audio Signature DS, and so on.

The work was supported by the Ministry of Education, Taiwan, R.O.C., MOE Program for Promoting Academic Excellence of Universities under the grant number 89-E-FA06-2-4

0-7803-8603-5/04/$20.00 ©2004 IEEE

We found that the melody part (Melodycontour and Melodysequence DS) is not convenient because we will need a satisfactory polyphonic automatic transcription program, which is not easy to find. Most of the descriptors in timbre part (HarmonicTimbre and PercussiveTimbre DS) only describe the spectral characteristics of a single note and is not suitable for our purpose. For an AudioSignature DS, its AudioSpectruemFlatness descriptor can describe the characteristics of the spectrum, which fits to analyze the signal spectrum for a long period and can be utilized in nowadays-audio format. Besides, in MPEG-7, the scalable series property [6] can be arbitrarily adjusted according to the need of temporal resolution and spectral bandwidth [5].

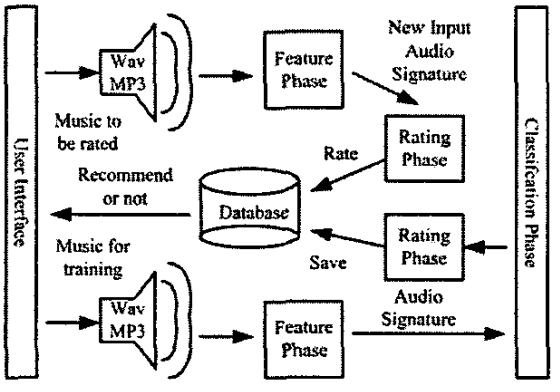

Figure 1.Overview of the system

3. System Architecture

The basic structure of the whole system flowchart is illustrated in Figure 1. The goal of the system is to analyze the listening habits of a certain individual and use it to rate other music data. Using vector quantization, we can determine the rating value of a new music by the most common rating values in group of audio data.

We divide the system flow into three phases, feature extraction phase, classification phase and rating phase. The feature exaction phase receives input audio and starts to generate MPEG-7 Audio Signature DS. The classification phase handles the corresponding audio signatures and compares them with the other signatures in the database. In the rating phase, the new music is rated.

3.1. The Feature Extraction Phase

The major work of this phase is to extract the audio signature from the audio input according to the spectrum flatness of the specified frequency subbands. Every frame has equal parts of log-frequency bands and the frequency components derived from short time Fourier transform (STFT) will be needed to compute the flatness of the frequency subband as the arithmetic mean divided by geometry mean. A flatness close to 1 means that in this subband, the frequency components are nearly the same, and representing a "noise-like" signals; otherwise, a value close to 0 means these components are "tone-like" signals. From the flatness, we can proceed to design a similarity measure algorithm that can classify these audio signatures. In our system, we extract two kinds of features. One is the audio signature raw data, and the other is the mean and variance of audio signatures of each subband over the total numbers of frame.

3.2. The Classification Phase

The classification phase is to match the input audio signature and the audio data in the database. We use the LBG vector quantization method [7] for classification. Each quantization group after classification will have a quantization vector averaged by audio signature members in the group. The distances between an input audio signature and these quantization vectors are calculated as the minimum of the mean square errors made by the input to the other audio data in the database.

Beside, in order to reduce the complexity, we adopt the partial distortion elimination (PDE) matching algorithm [7] that has been used in motion estimation of video processing. In the beginning, our system sets a minimum mean-square-error (mmse) threshold and the system will update the mmse when the mse is smaller than mmse. The system may leave the matching loop when the accumulated mse has been larger than mmse. Finally, we choose the quantization group which has mmse of all distances.

3.3. The Rating Phase

For the propose of constructing the user's preference, a user must assign the songs for training with personal rankings ranging from one to five before the feature extraction phase. This helps us to classify the categories of these songs.

After having established the user's preference, the rating phase starts when the user inputs a new song. The system will choose the most common rating value in the quantization group as the rating of the new song. The system performance can then be examined by

checking if his rating is the same as the one recommended by system.

In addition, we give a tolerance to the results. If the difference between the rating value by the system and the one by personal assignment is plus/minus 1, we may regard the rating value as an acceptable answer. This is based on the assumption that people may have similar feeling of preference if the difference of the ratings is small.

4. Experiment results

4.1. Experiment Design

Experiments were conducted on a standard PC, with a Pentium IV 2.4 GHz CPU Windows XP professional OS. The music database is composed of 561 44.1KHz, 16-bits resolution, mono wave and mp3 clips. Each audio clip was represented by a 30-second signal extracted from the beginning. We collected classic, jazz and pop Mandarin and Western songs as samples.

The 571 samples are divided into two parts. Two-thirds of the samples (391) are used for training the database, while the remaining one-third (180) are used as test queries. Then we continue to divide this two parts manually into three categories, classic, jazz, and pop.

The rating value on each song is already known for examination. Besides, we take two users' rating data for the experiment. Table 1 shows the overall score-distribution in the experiment database. The number enclosed in the parentheses represent number of pieces belonging to classic, jazz, and pop categories respectively.

User1/2 Score	1	2	3	4	5
Total #	69(30/21/18)	94(41/30/23)	81(38/23/20)	78(41/12/25)	71(28/12/31)
Training #	41(19/12/10)	44(24/9/8)	38(20/8/10)	33(18/7/8)	25(13/3/9)
Total #	45(18/4/23)	92(38/12/42)	128(67/30/31)	79(41/23/15)	49(14/29/6)
Training #	24(10/1/13)	40(22/6/12)	52(29/10/13)	43(24/13/6)	19(9/9/1)

Table 1 The score distribution in the experiment database

In addition to our proposed system, we also applied two other classification algorithms for comparison.

1. Five Arrays Average (FAR): By intuition, we choose five representative audio signatures from each score group, and each of them is obtained by the average audio signature of the score group. Then the system will decide the rating of the input by selecting the minmse distances between input and representative audio signature.

2. K- Nearest Neighborhood classifier (KNN): The rating of the new song is decided by the most common rating numbers in the k-nearest neighbors.

Table 2 shows our experiments on the testing systems varied between classification algorithms and types of features.

Feature / Classifier	VQ	KNN	FAR
Audio Signature Raw Data	I	II	III
Mean and Variance of Audio Signature	IV	V	VI

Table 2 Testing Systems numbers

The abbreviations used in the following are:
ACR : The Average Correct Recommendation ratio
ST : Standard deviation

4.2. Experiment Results

Table 3 shows the relationship between the music categories and recommendation ratio. the effect of the number of frequency bands. We find that different users can have different performance on each music categories, and it may be good to classify music pieces into their own music categories first. For the following tables, suffices I ,II ,III etc. represents the system numbers for testing in Table 2,3,and 4.

System I		Classic	Jazz	Pop	Total
User1	ACR I (%)	61.7	71.8	55.5	60.1
	ST I	1.68	1.57	1.97	1.7
User2	ACR I (%)	72.3	66.6	75.5	72.5
	ST I	1.30	1.47	1.23	1.4
System IV		Classic	Jazz	Pop	Total
User1	ACR IV (%)	62.8	71.8	53.3	62.4
	ST IV	1.80	1.56	1.94	1.84
User2	ACR IV (%)	76.6	53.8	68.9	73.6
	ST IV	1.21	1.69	1.36	1.36

Table 3 ACR and ST of our proposed system (frequency number = 16)

Table 4 shows the effect of the number of frequency bands. We tested 4 cases. The number of frequency bands in each case is 2,4,8 and 16, and each case is calculated only once. From Table 4 we can

observe that the number of frequency bands have less effect on our proposed system

System I		2	4	8	16
User1	ACR I (%)	56.7	61.8	60.7	60.1
	ST I	1.73	1.66	1.97	1.7
User2	ACR I (%)	73.0	74.2	69.6	72.5
	ST I	1.32	1.33	1.44	1.4
System IV		2	4	8	16
User1	ACR IV (%)	49.4	59.0	59.5	62.4
	ST IV	1.88	1.67	1.85	1.84
User2	ACR IV (%)	72.5	70.2	73.0	73.6
	ST IV	1.32	1.42	1.31	1.38

Table 4 ACR and ST on frequency subband numbers

Table 5 illustrates the performances of all six systems with different features and classifiers. We discovered that LBG vector quantization method has a better performance than KNN and FAR classification algorithm. It is also amazing that the feature of audio signature mean and variance seems perform better than audio signature raw data. It is with fewer data and can speed up the comparison procedure. However, it needs more experiments for justification.

		I	II	III	IV	V	VI
User 1	ACR (%)	60.1	57.8	46.1	62.4	52.2	46.1
	ST	1.70	1.81	2.24	1.84	1.94	2.23
User 2	ACR (%)	72.5	62.9	35.4	73.6	56.7	35.5
	ST	1.40	1.68	2.30	1.37	1.76	2.30

Table 5 ACR and ST on our system and testing systems (frequency subband number = 16, KNN : K value =5)

5. Conclusions

We have proposed an audio recommendation system based on audio signatures in MPEG-7 Audio. The recommendation system using audio signatures is simple and easy to implement. The performance of the system with tolerance is satisfactory(around 75%). Other MPEG-7 descriptors and other powerful classifiers like support vector machine and k-medians algorithm will be tested and compared in the near future.

6. References

[1] D.W. McDonald, "Ubiquitous Recommendation Systems," *IEEE Computer*, 2003.

[2] F. F. Kuo and M. K. Shan, "A Personal Music Filtering System Based on Melody Style Classification," *IEEE Proc. Data Mining*, 2002.

[3] H. C. Chen and L. P. Chen, "A Music Recommendation System Based on Music Data Grouping and User Interests," *IEEE Proc. User Modeling*, 1999.

[4] S. Quackenbush and A. Lindsay, "Overview of the MPEG-7 standard," *IEEE Trans. Circuit and Systems for Video Technology*, vol.11, no.6, pp. 725-729,2001.

[5] J. Herre, O. Hellmuth and M. Cremer, "Scable Robust Audio Fingerprinting Using MPEG-7 Content Descirption," *IEEE Proc. Workshop on Multimedia Signal Processing*, 2002.

[6] *Information Technology-Multimedia Content Description Interface*-Part4: Audio, ISO/IEC 15938-4,2001.

[7] S. H. Chen and W. M. Hsieh, "Fast algorithm for VQ codebook design," *IEE Proc. Communication, Speech and Vision*, 1991.