International Journal of Principles and Applications of Information Science and Technology
July 2008, Vol.2, No.1

A Feature Based Music Content Recognition method using Simplified MFCC

Bokyung Sung[1], Myung-Bum Jung[1], and Ilju Ko[1]

[1]Department of Media, Soong-Sil University, Sould, Korea
{ivsinger, nzin, andy}@ssu.ac.kr

Abstract. Since large amounts of digital music began to be distributed through the internet, demand for music related content (lyrics, musician profiles and movie clips) has been increasing. Recently, music related content has been offered through online services that usually use text based technology. This technology recognizes identical content by using file name or tag information associated with the music. Recognizing identical music content is difficult if file name or tag information is wrong. We therefore need a method for recognizing identical music content that uses audio data itself, in order to overcome these problems. In this paper, we propose a feature based identical music content recognition method for music related content services. We extracted feature data that have audible characteristics from waveform data of music content and processed it using a simplified version of Mel Frequency Cepstral Coefficient (MFCC). Using this method we could identify consistent feature data, even if two waveforms used different digitizing specifications and were not exactly the same. We succeeded in all 500 recognition experiments using 1,000 randomly collected songs of the same genre to prove the validity of the proposed feature based music content recognition method. 500 digital music files were made by mixing different compression standards and sound qualities from 60 digital music files. The recognition method could find identical music content regardless of wave form changes. Our experiments prove that a simplified version of MFCC is effective in recognizing identical music content.

Keywords : Digital music service, digital music recognition, music related content, Mel Frequency Cepstral Coefficient (MFCC), Dynamic Time Warping (DTW)

1 Introduction

Recently, the direction of the music market has changed from off-line to on-line access. Music player technology has also changed from analog equipment to internet connected digital platforms including portable devices. Subsequent to these changes, many demands on music related content have arisen. Such demands range from text content, like lyrics and musician profiles, to multimedia

content such as music related pictures, music video and TV films with music. We should be able to recognize music content using the music itself to support the various related content offerings.

Commercial music services can recognize music content by using the registration numbers of CDs or text information (file names or stored tag information) from digital music files. However, text based recognition methods cannot recognize identical music if the wrong filename or tag information has been added to digital music. We therefore need a music content recognition method that uses the audio data itself to overcome these problems.

The feature based identical music content recognition method described here addresses the problems we have described. This method's structure is divided into two parts. The first part is the extraction of feature data that have audio characteristics from the waveform data of music content. The second part is the recognition of identical music content having either exactly the same wave form or a very similar waveform, by comparing feature data extracted from both input and server music. This method is independent of text information (file name, tag information) and robust across different standards and sound qualities like sampling rates, channel and bit-rates. The proposed method can recognize music content using only the waveform of the music file and then offer music related information.

The waveforms of music content are always transformed whenever they are re-sampled. The waveform is therefore not an absolute criterion even if two music files have identical music content. To extract useful feature data without reference to waveform transforming, we should use a method that can recognize audio characteristics. For the method described in this paper, Mel Frequency Cepstral Coefficient (MFCC) [1], [2], [3], [4] was used. We can use this method to acquire feature data that have audio characteristics. This is based on the human hearing model and is one of the well known methods for extracting feature data from vocal music. The waveform of music content is divided into short regular lengths as frames. Feature data are extracted from each frame by MFCC. We applied a simplified version of MFCC and extracted 13 feature data from each frame. It is hard to identify related feature data sequences that are extracted from music content through simple comparison of data values. One of the major attributes of music contents is time. The time code gap of the same sound's waveform occurs whenever music contents are re-sampled. For robust perception at these gaps, we should use dynamic methods that can be adapted to changing time axes The method used in this paper is the Dynamic Time Warping (DTW) algorithm [5], [6], [7], [8], [9]. This is an algorithm that dynamically compares each data by considering the beginning and end of time axes. It is shown to be robust when time axes change.

The structure of this paper is as follows. In section 2, we propose our feature based music recognition method. We describe how we extract audio feature data characteristics by using a simplified MFCC and a dynamic perception method based on the DTW algorithm that is shown to be robust against time axis changes. In section 3, we describe our experiments with feature based music

content recognition using various re-sampled music files. In section 4 we describe some related work, and in section 5, we provide some conclusions and suggest future work.

2 Feature Based Identical Music Content Recognition

Identical music recognition provides a means to retrieve music related content. There are two main methods; the first is to use MIDI or tag data, the second is to use extracted data from the waveform of music. MIDI data has information about the tempo, chords, instruments etc. of music. In contrast, tags have objective information about music like musicians, title and sale date. Using these methods, identifying music is straightforward, but this data may be limited or incorrectly attributed. An alternative is to extract information from music waveforms using DSP (Digital Signal Processing). Unlike the other methods, this method uses information extracted from the music waveform itself rather than external data. Therefore, this method can be applied to all music encoding standards (e.g. MP3, OGG, and WMA). Extracted information is usually pitch, timbre and harmony that take discrete value forms.

In Fig. 1, we can see the overall structure of a music content recognition method. The structure is divided into four parts. The first part is to acquire waveform audio through the decoding of music input. The second part is to extract feature data from waveform audio. The third part is to recognize identical music by measuring similarities between extracted feature data from music input and saved feature data in a music server. The fourth part is to offer the user music related content.

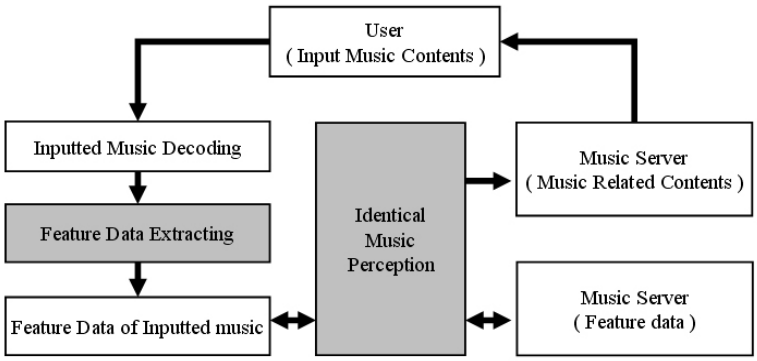
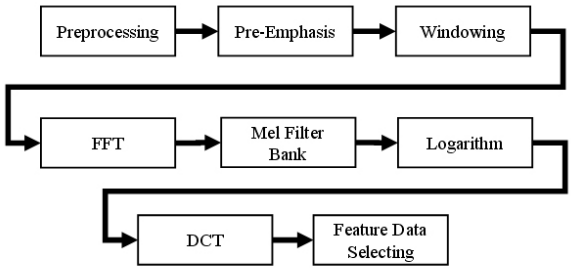


Fig. 1. The structure of a music content recognition method. Music input is decoded to a waveform, and then feature data is extracted from it. This feature data is used to identify the music so that related content can be offered.

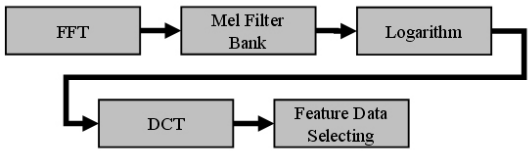
2.1 Feature Data Extraction

MFCC has been used in voice recognition to extract features from human voice input. It is based on human hearing modeling rather than on human speaking modeling. We assumed that it was possible to use MFCC to extract feature vectors not only from single channel voice but also from music data containing various instrumental sounds. In this paper we describe a simplified MFCC process for music content recognition.

Fig. 2 (a) shows standard MFCC processing and Fig. 2 (b) shows the simplified MFCC process described in this paper. In Fig. 2 (b) the preprocessing, pre-emphasis and windowing steps are skipped. Preprocessing and windowing are used to remove noise and silent areas from the continuous waveform of a human voice. Pre-emphasis is used to emphasize only the voice part in recorded sound. Unlike the voice domain, the music domain mixes various sounds of not only the human voice but also musical instruments. In addition there is also music that consists only of instrumental sounds. Because of this, we reason that the complete standard MFCC process is not required for the music domain. We therefore removed these three processing steps and evaluated a simplified form of MFCC.



(a) General MFCC process



(b) simplified MFCC process

Fig. 2. (a) shows the general MFCC process and (b) shows our simplified MFCC process.

In simplified MFCC, only the following stages are required. FFT is a step to convert the waveform signal from time domain data to frequency domain data. In the frequency domain it is easier to identify each element of the signal and to analyze and process the signal than in the time domain. The Mel Filter Bank step involves multiplying the value of the Mel-scale filter with the result of FFT

processing. This filter is modeled on the human hearing structure. The human ear is sensitive at sounds above and below 1 kHz. This processing analyzes sounds of less than 1 kHz more intensively. Log Scaling takes the log at values that run through the filter. The human ear perceives sound in a log function form. Discrete Cosine Transform (DCT) removes the correlation between the output values of the filter bank and collects features of parameters. If MFCC is applied at one frame, feature vectors as bank numbers of the Mel Filter Bank are generated. The Feature Data Selecting step involves choosing feature vectors from 0 degrees to the degree that we want to get from the feature vector as a bank number of the Mel Filter Bank. In our experiment we used 13 feature data.

2.2 Identical Music Contents Recognition

In Fig. 3, we can see some differences in the waveform between two files of the same music that are encoded by different digitizing standards. These variations occur because of differences between the various loss compression methods.

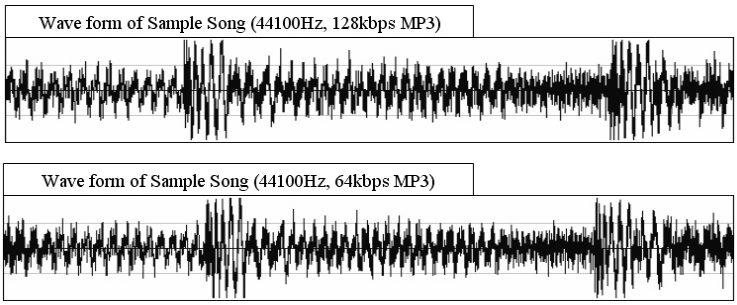


Fig. 3. These are waveforms of the same music encoded by different digitizing standards. The upper example is a waveform of a sample MP3 file that is encoded at 44100 Hz and 128Kbps. The lower example is a waveform of a sample MP3 file that is encoded at 44100Hz and 64Kbps.

Despite having the same music contents, matching errors can occur if we match only the same time code frames when comparing music files. There are no differences between these two feature data sequences of the same music if they use the same digitizing standard. However, there are differences in waveform between two feature data sequences of the same music if they use different digitizing standards, so feature data values will also be different. Therefore we need to compare the areas around a feature when measuring the similarities between two feature data sequences of digital music that use different digitizing standards. Similarity is measured by the absolute distance between two feature data. The smaller the value, the higher the similarity.

Fig. 4 shows a graph comparing the 'Difference Value' with the 'DTW value', which is the result of applying the DTW algorithm. The Difference Value is the

distance difference between the feature data of two music files in a frame sequence. The Difference Value method gives a perfect match in the case of music that has the same digitizing standard. On the other hand, it shows a big difference value in cases where the same music is encoded using different digitizing standards. In contrast, the 'DTW Value' method shows consistent similarity even where music has different digitizing standards. Therefore the similarity between digital music can be measured by applying the DTW algorithm regardless of digitizing standards.

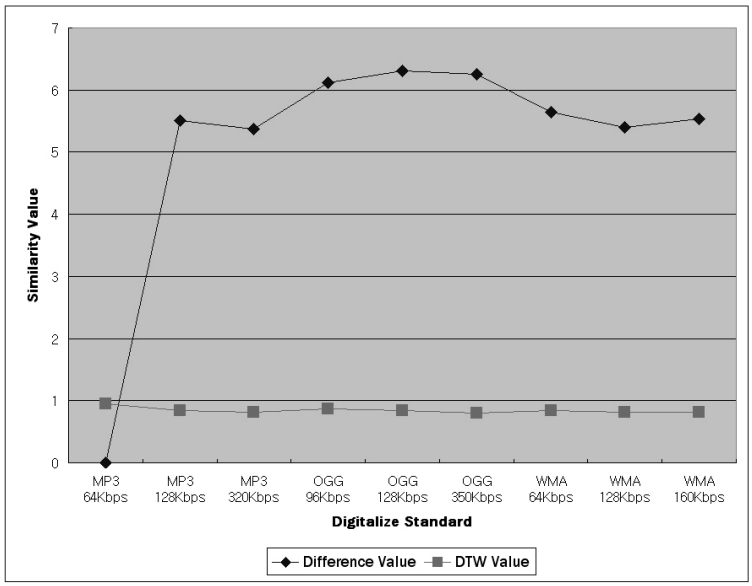


Fig. 4. Graph comparing similarity values derived from difference values and DTW values.

Figure 5 shows the result of applying DTW to compare input music and server music. To find the best warping path for each frame, we consider not only the same sequence frame but also the surrounding part frames.

Formula (1) is the computation formula of DTW. In D (A, B), 'A' and 'B' are two music inputs. Function D computes the distance difference. For similar degree measurement, two input songs are able to be expressed as a line of feature vectors. If the lengths of the two songs are different, they should be warped at one time axis. The function that does this is the Warping Function, 'c(j)'. K means the point number of the song that is expressed as a line of feature vectors from the warping function. 'ω(j)' is the weight coefficient. It is used to provide flexibility to the warping function and to find a fitting warping function. The denominator of the DTW function is the weight sum of warping functions. In the function, it used to compensate for the influence of K that is the point number.

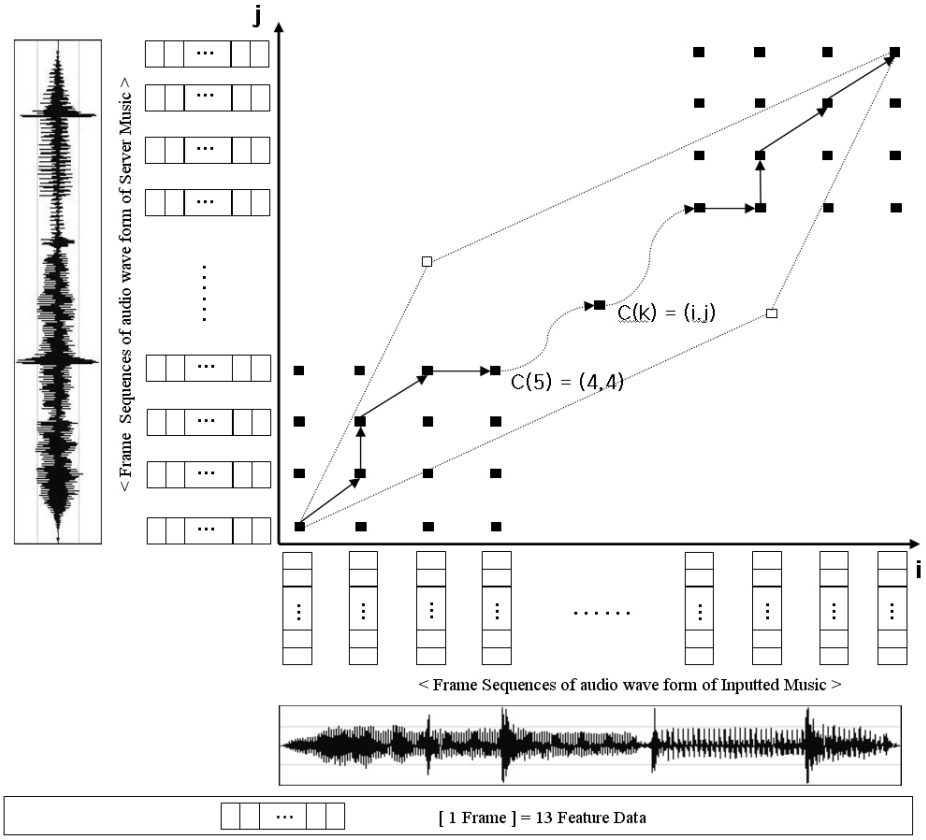


Fig. 5. This graph shows the DTW algorithm. Feature data of input music and server music are dynamically warped.

$$D(A,B) = \min_{F}\left[\frac{\sum_{j=1}^{k} d(c(j)\cdot\omega(j))}{\sum_{j=1}^{k}\omega(j)}\right] \tag{1}$$

3 Experimental Results

The experimental objective was to find music in the music server that was identical to input music. The music server contained 1,000 songs in random order. All of them were sampled at 22050Hz and quantized 16bit.

The input music data was 60 songs that were also present in the music server. 500 digital music files were made by mixing different compression standards and

sound qualities from these 60 songs. The nine standard specifications used were MP3 (64kbps, 128kbps, 320kbps), OGG (96kbps, 128kbps, 350kbps) and WMA (64kbps, 128kbps, 160kbps) (Fig. 6).

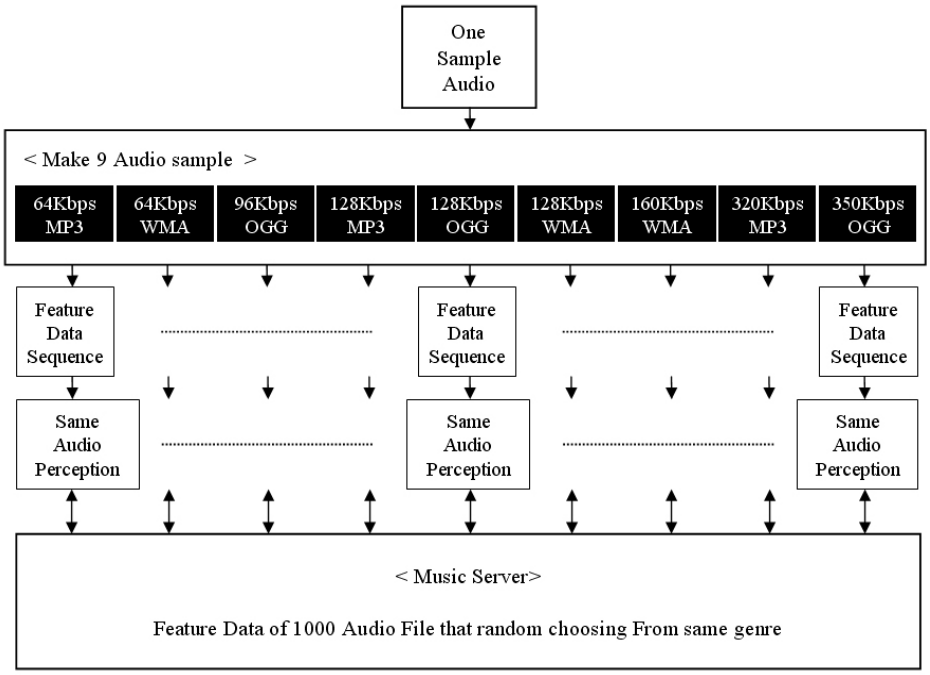


Fig. 6. The structure of the experimental method. An input music sample is converted 9 audio samples that use different digitizing standards. Then the 9 audio samples are used to locate identical audio in the music server.

Fig. 7 shows nine result graphs of similarity measurements between one music query and all the music in the music server. The nine audio queries and music sample number 581 in the music server are same song. We can see that the key similarity value for song 581 is clearly smaller than the others in the graphs. The vertical axis of each graph is the similarity value. Similarity means the distance difference between the feature data of two music files. If the similarity value is close to zero, it means a high similarity. The similarity value is close to zero on the 581 scale of each graph. The similarity values do not reach zero because a dense comparison between two audio files was not made to improve speed. We performed the retrieval experiment 500 times with 60 songs using the method described above. In these recognition experiments, we achieved 100% correctness

In similarity measurement between input music and server music, each music file on the music server is compared with the input music. The input music is usually in a short length clip form. Therefore, the lengths of the input music clip

and the server music are usually different. In the comparison method, every moment of comparison gives a similarity value. From all these similarity values, the smallest value is the Key Similarity Value, and the second and third smallest values are the Second Similarity Value and Third Similarity Value respectively. Figure 8 is a result graph of similarity measurements between nine standard music samples that were generated from one input music file and the music on the server. The value for music that has Key similarity is plainly smaller than values of music that have Second and Third similarity. As a result of this, we can see that the Key similarity value is a discrimination value. This graph also shows that we can find identical music by using samples that use different digitizing standards.

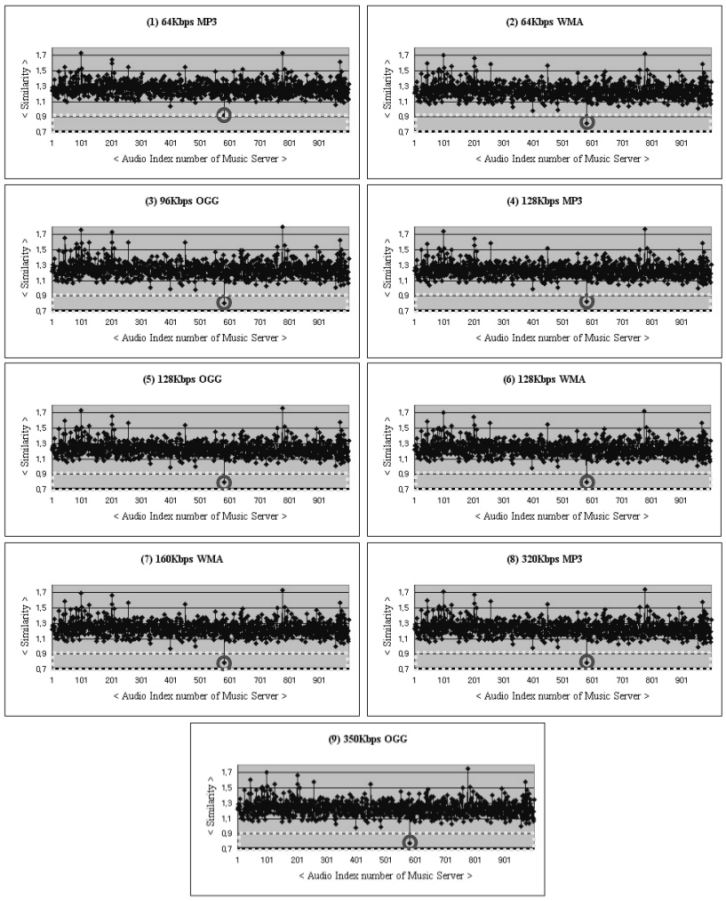


Fig. 7. Nine result graphs of similarity measurements between one music query and all the music in the music server. The various cases are 64Kbps MP3, 64Kbps WMA, 96Kbps OGG, 128kbps MP3, 128kbps OGG, 128kbps WMA, 160Kbps WMA, 320Kbps MP3 and 350Kbps OGG.

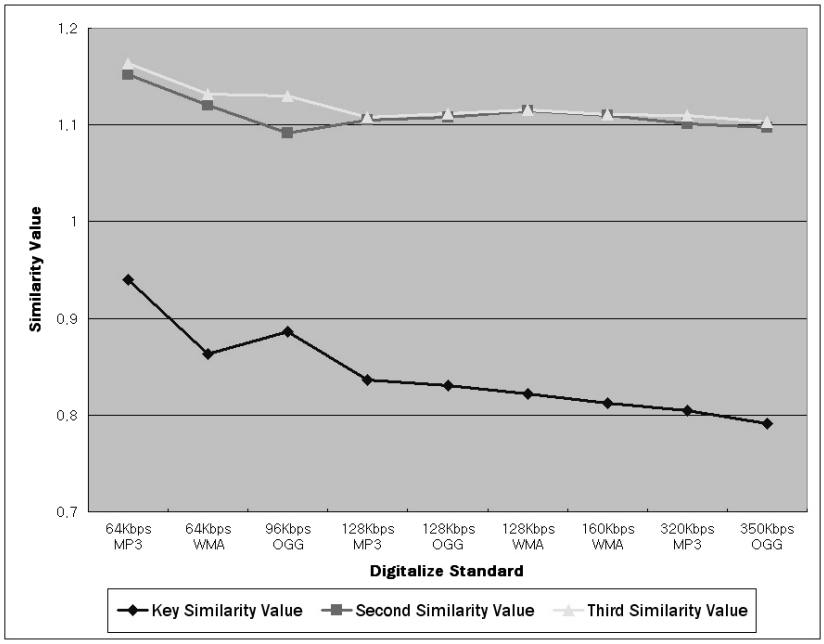


Fig. 8. This graph compares Key, Second and Third Similarity values of 581 audio samples using nine different digitalizing standards.

4 Related Work

MFCC and DTW are methods that have been used in many fields. MFCC [1], [2], [3], [4] has been applied widely from voice recognition to acoustic sound analysis. DTW [5], [6], [7], [8], [9] has been used in both the vision and voice recognition fields. In the voice recognition field, MFCC and DTW are usually both used together [10], [11]. However, even if both voice and music are usually represented as a waveform signal, these are considered to be different domains. Music waveforms consist of various sounds that are unlike voice waveforms. In the music domain, the transformation rate of the time axis is no higher than for voice. Therefore, applying DTW without any transform can be a reasonable approach. However, the music feature extraction method described here should be a form that is considered more appropriate to processing the attributes of music waveforms.

5 Conclusion

In this paper, we proposed a feature based identical music content recognition

method using a simplified version of MFCC. We used DTW to measure the similarities between samples in order to recognize identical music. Our proposed method successfully found identical music content, regardless of the transformation of the wave forms in the samples. We therefore proved that identical music content recognition is possible using simplified MFCC.

In future work we will apply our proposed method to music retrieval and web service applications for offering music content. Our results indicate that we will be able retrieve music from music clip queries without further user action and to serve the correct music related content.

References

1. Z. Jun, S. Kwong, W. Gang, Q.Hong.: Using Mel-Frequency Cepstral Coefficients in Missing Data Technique. EURASIP Journal on Applied Signal Processing, Vol.2004, No. 3 (2004) 340-346.
2. M. Xu, NC. Maddage, C. Xu, M. Kankanhalli, Q. Tian.: Creating audio keywords for event detection in soccer video. Multimedia and Expo. ICME03 Proceedings, Vol.2, No.2 (2003) 281-284.
3. Shannon B.J., Paliwal K.K.: A comparative study of filter bank spacing for speech recognition. International Microelectronic engineering research conference, Brisbane, Australia, (2003).
4. Ziyou Xiong, R. Radhakrishnan, A. Divakaran, T.S. Huang.: Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. 2003 Multimedia and Expo. ICME '03 Proceedings. 2003 International Conference on, (2003) 397-400.
5. JC. Brown, A. Hodgins-Davis, PJO. Miller.: Classification of vocalizations of killer whales using dynamic time warping. The Journal of the Acoustical Society of America, Vol. 119, (2006) EL34-EL40.
6. C.A. Ratanamahatana, E. Keogh.: Three myths about dynamic time warping data mining. Proceedings of SIAM International Conference on Data Mining, (2005)
7. A.M. Youssef, T.K. Abdel-Galil, E.F. El-Saadany, M.M.A. Salama.: Disturbance classification utilizing dynamic time warping classifier. IEEE Transactions on Power Delivery, Vol. 19, No. 1, (2004) 272-278.
8. A. Pikrakis, S. Theodoridis, D. Kamarotos.: Recognition of isolated musical patterns using context dependent dynamic time warping. IEEE. Speech and Audio Processing, Vol. 11, (2003) 175-183.
9. EJ. Keogh, MJ. Pazzani.: Computer Derivative Dynamic Time Warping. First SIAM international Conference on Data Mining (2001)
10. Kai Sze Hong, Sh-Hussain Salleh.: An Education Software in Teaching Automatic Speech recognition (ASR). 7th International Conference on Spoken language Processing (2002) 625-628.
11. Jingwei Liu.: A DTW-Based DAG Technique for Speech and Speaker Feature Analysis. 8th European Conference on Speech Communication and Technology (2003) 473-476.