Detection of Chorus from an Audio Clip using Dynamic Time Warping Algorithm

Garima Vyas & Malay Kishore Dutta
Dept.of Electronics and Communication Engineering
Amity School of Engineering & Technology
Noida (U.P)-201303, India
gvyas@amity.edu, mkdutta@amity.edu

Hicham Atassi & Radim Burget
Dept. of Telecommunications
Brno University of Technology
Technick`a 12,612 00, Brno,Czech Republic
atassi@feec.vutbr.cz, burgetrm@feec.vutbr.cz

Abstract— **This paper describes a method to detect repeating segments in an audio signal by using dynamic time warping algorithm. The proposed framework extracts features from frames of the audio by Mel frequency cepstral coefficients. The features extracted from the audio clip of the chorus were matched against the features of the whole clip by dynamic time warping. The number of matches found was determined by self similarity matrix. The experimental results indicate that the minimum distance matches between query and reference clip is successfully achieved. The proposed scheme was tested in a database of audio signals and the experimental results are encouraging. The proposed scheme was implemented and tested using a database of audio signals with accuracy up to 98%.**

Keywords— Dynamic Time Warping; Mel Frequency Cepstral Coefficients; Repeating Segment; Audio Signal.

I. INTRODUCTION

Human beings have amazing ability to distinguish different types of audio. Given any audio piece, one can instantly identify the type of audio (e.g., human voice, music or noise), speed (fast or slow), the mood (happy, sad, relaxing etc.), and determine its similarity to another piece of audio. However, a computer perceives a piece of audio as a sequence of sample values. A lot of research is going on how can we classify music on the basis of its genre (E.g. classic, pop, folk music, Hip-hop etc)? And how can the system automatically recommend music to the listener? Any two audio clips are generally multi-variant time series which means the *i*-th part of one clip is not similar to *i*-th part of second clip. Hence for classification proposes some decisive feature points [1], [2] have to be extracted.

Songs generally have some repeating segments; these segments are called as chorus and can be used as audio thumbnails. The feature points from the small segment of audio are generated. These feature points are generally known as audio fingerprint. Audio features can belong many domains like temporal domain, frequency domain or cepstral domain [4], [5]. A feature set to be chosen is entirely dependent on application. For music information retrieval and classification cepstral features [1], [4], [5] seem to give best results [8], [10], [11]. In this paper the Mel Frequency Cepstral Coefficients are extracted. The existing technique for feature extraction in cepstral domain is Linear Prediction Cepstral (LPC) coefficient. MFCC [1], [2], [4], [5] has some advantage over LPC as it uses Mel Scale instead of normal frequency scale. The Mel scale is used to mimic how human auditory system responds to sound. Mel scales spectral resolution becomes lower as the frequency increases. Hence, information at higher frequencies is down-sampled. Another advantage of MFCC is that is uses discrete cosine transform which provides excellent energy compaction means that the majority of the information is concentrated at lower bits and less information is stored at higher bits. With these consideration of these points MFCC feature extraction was chosen for this work.

The matching is done by dynamic time warping (DTW).The main aim of the DTW [6] is to match or to find similarity between two Multi-Variate time dependent signals under given boundaries. The amount of similarity can be described in the term of distance between two audio fingerprints. Hence, the output of DTW is a distance vector. The less is the distance between two fingerprints, more is the similarity between them and vice -versa. The warping, also known as elastic shifting, of the signals along the time axis is done in a non-linear fashion. The two sequences may be out of time phase. The older approach to measure similarity between two multi-variate time series signal is “Euclidean Distance”. DTW is preferred over Euclidean distance [9] in the case of DTW the indexing of time series among two different time phase signals is more accurate and more reliable [6].

Hence, the objective is to train proposed system to determine similar type [3], [11] of audio chorus in a same audio clip or in different audio clips. The proposed method may be very useful for music classification [10], indexing [8] and music information retrieval. The detection of chorus can help in browsing music from a collection of songs. Hence it will save time of the user.

The rest of the paper is organised as follows. Section 2 describes the proposed method to be followed which includes the feature extraction method by using MFCCs and also describes the matching mechanism of the query

978-1-4799-2291-8/14/$31.00 ©2014 IEEE

with reference by using DTW algorithm. In Section 3 we represent experimental results illustrating the efficiency of the proposed framework. The last section is about the conclusion of the work and possible future work in this area.

II. PROPOSED METHOD

To achieve the above mentioned goals the methodology that is followed is explained by following flow chart shown in Fig. 1:

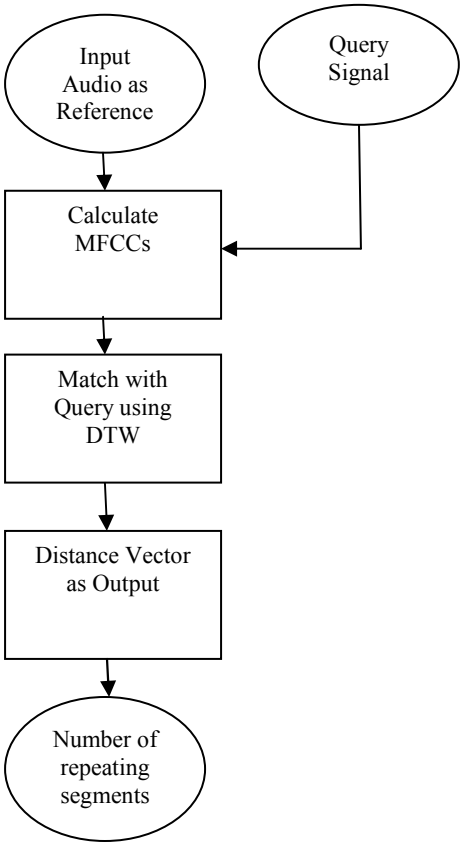


Fig. 1: Proposed Method for finding number of repeating segments in the audio clip.

Firstly, we calculate feature points from a reference audio by using MFCCs. These feature points are matched with the feature points of the query signal by adopting dynamic time warping algorithm. The output of the DTW is a distance vector which shows the distance between the query and reference. Less distance reflects more similarity between two audio clips and vice-versa. Hence, the number of chorus can be calculated.

A. FEATURE EXTRACTION USING MFCC's

The input audio signal of 10sec is chosen and segmented into frames of 40~60ms with an optional overlap of 35~55ms. Such a small frame size is chosen because the audio signal is nonstationary signal. By framing we are trying to map long nonstationary signal into small stationary signals. This means the frame length for 16kHz signal is 0.060*16000=960 samples. Now the MFCC is to be calculated as per the following steps.

Step 1: The each frame is now multiplied by hamming window [1] by using the below equation:

$$X(n) = S(n) \times H(n) \tag{1}$$

$$Where\ S(n) is\ the\ Framed\ signal$$

$$H(n) = Hamming\ window\ Function = \alpha - \beta \cos\left(\frac{2\Pi n}{N-1}\right)$$

Step 2: On each windowed signal, the Descrite Fourier Transform (DFT) is performed to obtain the magnitude frequency response. DFT is performed to analyze the different timbers in terms of frequency. The equation for DFT is:

$$X(k) = \sum_{n=0}^{N-1} X(n) e^{-\frac{j2\Pi kn}{N}}, \quad 0 \le k \le N-1 \tag{2}$$

Step 3: In the next step the magnitude frequency response is multiplied by 24 triangular band pass filters. The triangular filters [4] are used to reduce the size of the feature points and it smoothes the spectrum also. These filters are equally spaced along the Mel frequency. The liner frequency to Mel conversion takes place as written in Eq. 3:

$$Mel(f) = 1125 \times \ln\left(1 + \frac{f}{700}\right) \tag{3}$$

Step 4: In the next step the output filtered frequencies are again converted into a time-like domain called quefrency by applying discrete cosine transform (DCT) on it. DCT is chosen because it has excellent energy compaction as maximum of the energies are concentrated on low frequency components. The output of DCT [5], [2] is called as cepstrum, and as they are on Mel scale these feature points are called Mel frequency cepstral coefficients (MFCCs). The Eq. 4 describing the DCT is:

$$Y(k) = \sum_{n=0}^{N-1} \cos[\frac{\Pi}{n}(n + \frac{1}{2})k] \quad 0 \le k \le N-1 \tag{4}$$

Step 5: For efficient performance of the system, we add the logarithmic energy and compute the first difference (Δ). Taking logarithmic of energy compresses the dynamic range and makes the feature less variable to acoustic coupling changes. Moreover the human auditory system response is logarithmic in nature. Generally the log energy value of the feature is added as the 25th component of MFCC. Sometimes the differential of MFCC [1], [4] are

also of interest and hence can be added as 26th component. Several other components can also be added. Often, for speaker reorganization systems the MFCC of 39 dimensions is used.

B. DYNAMIC TIME WARPING OF TIME SERIES

The DTW [6] is used to find the similarity or difference between two multi-variate time dependent signals. Say we have two time series A of length n and B having length m. These two sequences are arranged on two sides of a grid of size n*m, one with on the top and other on the left side of the grid as shown in Fig. 2.

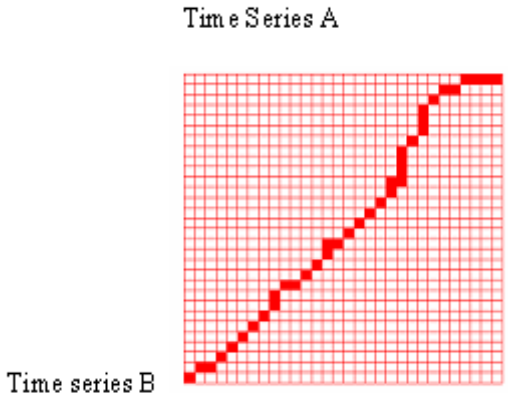


Fig. 2: Orientation of time series signal in DTW

The red dots in the grid shows the optimal distance path between two time series signal A and B. The optimal path between two sequences is found by minimising the total distance between them. For this all the possible routes [9] between A and B are computed and for each path we calculate the overall distance. Finally the path with minimum distance is chosen (as shown by red dots in Fig. 3). To normalize the length the weighing function is used. The distance between two points can be one to one mapped or may be one-to-many mapped. Generally one-to -many mapping is done. The one to one and one-to-many mapping is shown in Fig. 3:

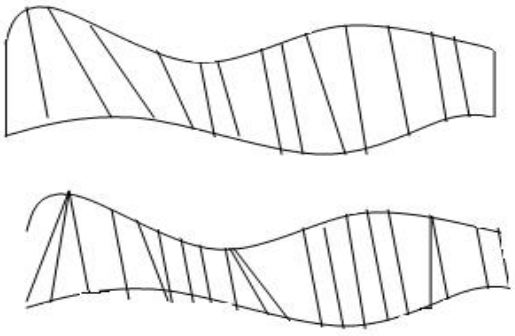


Fig. 3: One to one and one–to-many mapping in DTW

Some of the characteristics of the desired path are:

1. Monotonic & Continuity Condition: The path must be monotonic in nature i.e. the path will not turn back on itself. It should be a continuous monotonically increasing function.
2. Boundary Condition: The path must starts from bottom left corner and must end at top right corner. Otherwise the path is not said to be complete.
3. Warping Window Condition: An acceptable path is unlikely to wander very far from the diagonal. The maximum distance that is allowed to vary from diagonal is equal to the window size.
4. Scope Constrained Condition: The desired path should not be too steep and should not be too shallow.

The flow chart for the implementation of DTW is shown in Fig. 4:

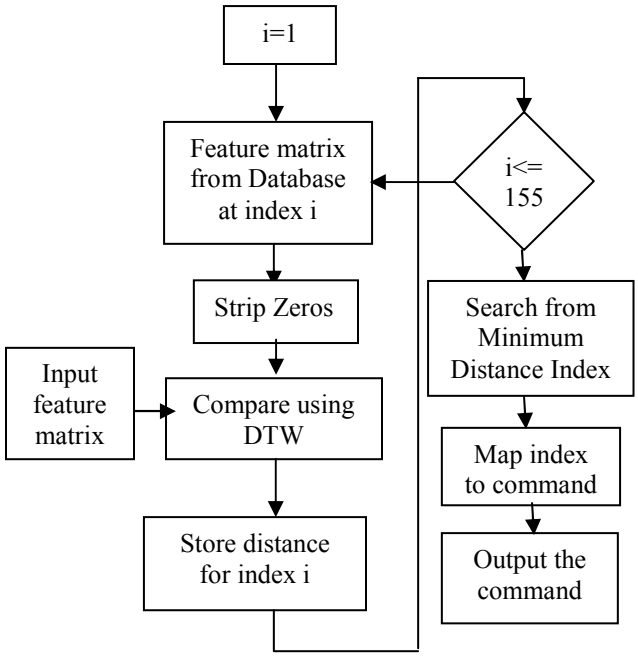
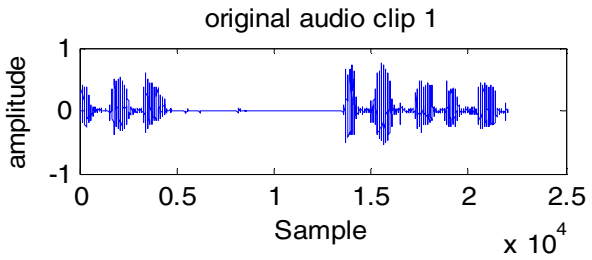


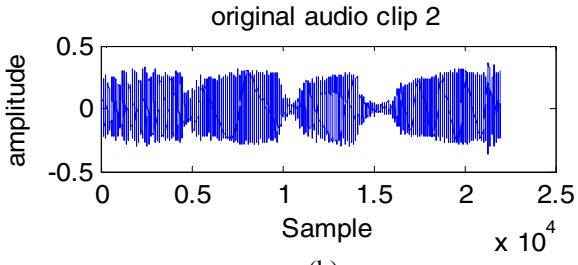
Fig. 4: Flow Chart for the Implementation of DTW Algorithm

III. EXPERIMNETAL RESULTS & DISCUSSION

Experiments have been performed in many songs of English and Hindi (Indian Language) for testing and validation of results. For example different segments of Hindi song of duration 2 minutes was considered and examined. The repeating segment in the song was taken as the query segment and was matched for calculating the number of repeating segments in rest of the song. The original audio clip which acted as a reference is shown in Fig. 5. The length of the reference is 10sec. Frames of 60 ms having an overlap of 55msec is implemented to extract MFCCs.



(a)



(b)

Fig. 5: (a) Plot of reference audio clip 1
(b) Plot of reference audio clip 2

After windowing and using FFT, the frequency components were passed through triangular filter bank shown in Fig. 6.

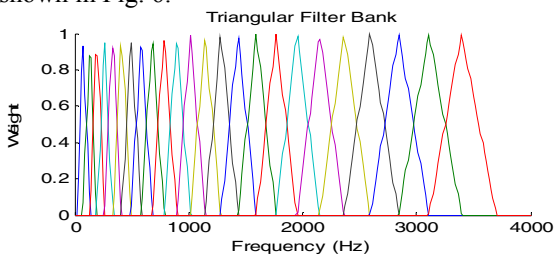
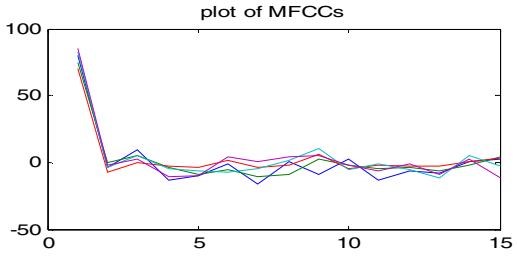


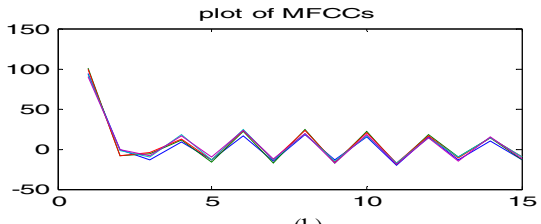
Fig. 6: Triangular band pass filter

After passing the filtered frequencies from DCT, log energy and delta cepstrum state we get the desired MFCCs. The plot of MFCCs is shown in Fig. 7 and their spectrogram is shown in Fig. 8.

After performing the matching part by DTW, the similarity function between query and reference was generated. This similarity function is plotted in Fig. 9. The lowest peaks show the minimum difference between two. In this clip the smallest distance is 0.0012 and maximum distance is 330.9149. These results indicate that that the query is most similar to the frame from which the distance is 0.0012.Or in other words we have got a similar chorus as that of our query.

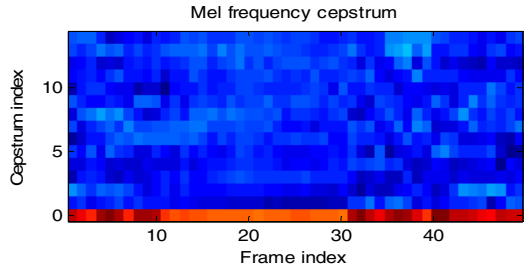


(a)



(b)

Fig. 7: (a) Plot of MFCCs of audio clip 1
(b) Plot of MFCCs of audio clip 2



(a)

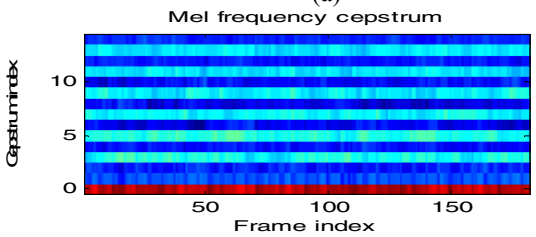
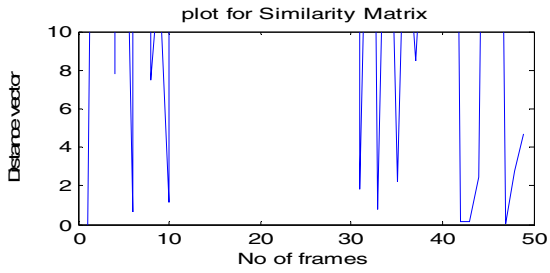
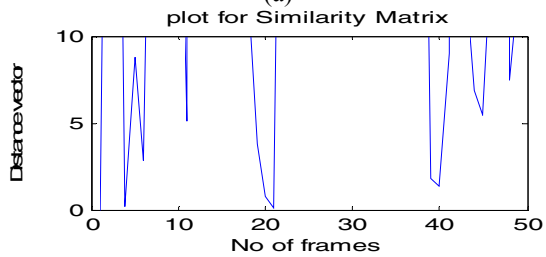


Fig. 8: (a) Cepstrogram for audio clip 1
(b) Cepstrogram for audio clip 2



(a)



(b)

Fig. 9: (a) Plot of similarity Matrix for audio clip 1
(b) Plot of similarity Matrix for audio clip 2

The distance matrix for audio clip 1 and audio clip 2 is represented in table 1 and table 2 respectively. The first element shows the matching of reference with itself. The minimum distance for audio clip 1 is 0.0012. Other approximate distances are 0.1360 and 0.1510. For audio

clip 2 the minimum distance is 0.1016 and approximate distance is 0.1996. These results reveals that there are three repeating segments present in the audio clip 1 and two repeating segments are present in audio clip 2. Similar experiments were preformed on different audio signals and the results are found to be encouraging.

The same algorithm is tested on different audio clips (English and Hindi). It was found that the proposed method is working efficiently on songs as well as on audio clips with accuracy from 98% to 92 %. For the accuracy calculation, the Euclidean distance between the query segment and other segments was calculated. The minimum distance between the segments help us to detect the repeating segment and the best match is found with the minimum distance parameter.

Table I : The Distance Matrix or Similarity Matrix when a chorus is Searched in an audio clip 1 having duration of 1min.

0	27.8737	103.7912	7.7665	30.0703	0.6412	88.7115
7.4823	12.5491	1.1667	133.7495	214.3206	164.7893	229.5122
248.8795	280.2300	287.6093	298.6330	299.7166	300.8247	330.9149
299.2211	308.3478	268.5738	242.0395	226.0998	256.8373	286.4003
308.7034	327.0083	1.8397	37.9737	0.8017	37.0026	2.2348
18.8174	8.4868	15.3857	140.6811	49.7645	37.9613	0.1510
0.1360	2.4701	35.9747	47.6288	0.0012	2.7803	4.6896

Table II: The Distance Matrix or Similarity Matrix when a chorus is Searched in an audio clip 2 having duration of 2mins.

0	40.4192	26.2113	0.1996	8.8011	2.8044	47.2941
78.4355	68.6495	37.5031	5.1010	182.1883	130.2512	73.6016
145.8980	129.6125	77.5715	14.1841	3.7503	0.7404	0.1016
73.7267	50.4832	44.8588	42.2283	33.4968	65.0921	45.1446
49.4923	55.6628	115.7910	186.8406	74.0829	37.0509	19.3256
55.9639	103.3541	46.9379	1.8049	1.3794	8.9855	58.7653
13.8607	6.8390	5.4578	15.8685	37.6368	7.4410	12.3564

Table III : Simulation Results for different audio clips

S.No.	Query Segment	Length of query	Length of reference	Min. Distance	No. of chorus	Accuracy (%)
1	Audio clip 1	1min	6sec	0.0012	3	98%
2	Audio clip 2	2mins	12 sec	0.1016	2	97%
3	Audio clip 3	2mins	10sec	0.0085	4	98%
4	Audio clip 4	3mins	10sec	0.6842	2	90%
5	Audio clip 5	3mins	10sec	0.4862	3	92%

IV. CONCLUSION

The Proposed scheme of detection of chorus from an audio clip using dynamic time warping algorithm is presented in this paper and experimental results indicate that this scheme gives results which has an accuracy up to 98% for detection of repeating segments. The features used for this method is Mel frequency cepstral coefficients based and the matching was done using dynamic time warping. The proposed method has been tested on a database of audio signals and the results are encouraging. Future work may be to test this algorithm in local languages and improve the accuracy of the results. Another possible futuristic work may be the use of frequency warping on audio clips for similar work and investigate if competitive or better results are achieved.

REFERENCES

[1] Yun Yun Chu, Wei Hua Xiong, Wei Wei Shi, Yu Liu, "The Extraction of Differential MFCC Based on EMD" Applied mechanism and materials, volume 313-314, 2013, page(s): 1167-1170.

[2] Osama Alhamdani, Ali Chikma, Jamal Dargham and Sh Hussain Salleh, “Efficient Speaker Verification System Based on Heart Sound and Speech” International Conference on Latest Computational Technologies, 2012, Bangkok, page(s): 65-68.

[3] Regunathan Radha Krishnan and Wenyu Jiang, “Repeating Segment Detection in Songs using Audio Fingerprint Matching’’ IEEE Transactions on Multimedia, Volume: 7, Issue: 1, Publication Year: 2012, page(s): 144-149.

[4] J. Alam , T. Kinnunen, P. Kenny, P. Ouellet, D. O'Shaughnessy, “Multi-taper MFCC features for speaker verification using I-vectors” IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)”, 2011, Wailoba, HI, page(s): 547-552.

[5] M.A. Hossan, S. Memon and M.A. Gregory, “A novel approach for MFCC feature extraction” IEEE conference on signal processing and communication systems, 2010, Gold Coast, QLD, page(s): 1-5.

[6] Fu. Wenjie, Yang Xinghai, Yutai Wang, “Heart Sound Diagnosis based on DTW and MFCC” IEEE 3rd International congress on Image and signal processing, Volume 6, 2010, Yantai, page(s): 2920-2923.

[7] Yuan Yujin, Zhao Peihua , Zhou Qun, “Research of speaker recognition based on combination of LPCC and MFCC” IEEE international conference on intelligent Computing and intelligent systems, Volume 3, 2010, Xiamen, page(s):765-767.

[8] Shih-Hao Chen, Shi-Huang Chen, R.C.Guido, “Music genre classification algorithm based on dynamic frame analysis and SVM” IEEE international symposium on multimedia, 2010, Taichung, page(s): 357-361.

[9] A.Zulfiqar, A. Mohammad, A.M.M. Enriquez, “ A speaker identification system using MFCC feature with VQ technique” IEEE Third international Symposium on Intelligent information technology application, Volume 3, 2009, Nanchang, page(s): 115-118.

[10] M. Goto, “A chorus section detection method for musical audio signals and its application to a music listening station,” IEEE Transactions on Audio, Speech, and Language Processing Volume: 14, Issue: 5, Publication Year: 2006, Page(s): 1783 -1794.

[11] M.A. Bartsch, G. H. Wakefield, “Audio thumb nailing of popular music using chroma-based representation,” IEEE Transactions on Multimedia, Volume: 7, Issue: 1, Publication Year: 2005, Page(s): 96- 104.