Audio Retrieval based on Cepstral Feature

R.Christopher Praveen Kumar
Assistant Professor
V.S.B College of Engineering
Coimbatore

S.Suguna
Assistant Professor
V.S.B College of Engineering
Coimbatore

J.Becky Elfreda
Assistant Professor
V.S.B College of Engineering
Coimbatore

ABSTRACT

The interest towards music is rapidly growing in our day to day life. It is necessary to have efficient system to retrieve relevant music for the user. The audio retrieval system mainly depends on the feature extraction process because only the meaningful feature will provide better retrieval task. In this work, audio information retrieval has been performed on GTZAN datasets using weighted Mel-Frequency Cepstral Coefficients (WMFCC) feature which is a kind of cepstral feature. The results obtained for the various stages of feature extraction WMFCC and retrieval performance plot has been presented. The mean precision values obtained for the audio files from the GTZAN database are 96.40% respectively.

General Terms

Segmentation, query, Audio, Filters.

Keywords

Audio Retrieval, Cepstral Feature, WMFCC, Feature Extraction, Mel Filter Bank.

1. INTRODUCTION

Audio information plays a major role in many multimedia applications. Nowadays, there is a lot of research going on towards audio retrieval and its application such as music similarity retrieval, artist identification, musical genre, or instrument recognition [1]. Audio information retrieval involves retrieval of similar pieces of music, instruments, artists, musical genres, and the analysis of musical structures [2]. The extraction of pitch, attack, and duration and signal source of each sound in a music piece are all related to music transcription. Retrieval of natural sounds other than speech and music are nothing but Environmental sound retrieval. New devices have been developed to hold the large collections of music piece. As a result an efficient search engine is required to pick the music of user’s choice [3]. In order to accomplish the task of searching the relevant music in efficient way, the entire system depends on the ability to retrieve audio files based on their content. An efficient method especially computerized method is required in order to access the large collection of music and moreover to retrieve relevant audio data from the large collection of music, an efficient and automated content-based retrieval system is needed. The major task in audio retrieval system is feature extraction process. The feature extraction process involves extracting audio features from audio data which will give meaningful information about the audio data. The feature which was selected for audio retrieval process will able to discriminate among various sounds. In this paper, Weighted Mel Frequency cepstral Co-efficient (WMFCC) feature has been used for audio retrieval process. The different stage of feature extraction process of WMFCC and its corresponding outputs has been explained in the next sections.

2. RELATED WORK

In this section, the different feature extraction methods used for audio retrieval process have been discussed. In [4], audio tag annotation task which has been performed on CAL500 and CAL10k corpora has based on Dirichlet mixture model (DMM) approach and this method uses Mel frequency cepstral coefficient (MFCC) for the retrieval task. In [5], with the help of Distance-from-Boundary (DFB) and Support vector machine (SVM), audio retrieval and classification task which use Mel cepstral feature had been performed on a database which consists of 409 sounds of 16 classes. A new approach for automatically annotate and retrieve audio files with the help of MFCC features has been proposed in [6]. The first step in this process involves segmenting audio clips into frames and for each frames, ensemble classifier is used to train each tag. The above mentioned task has been performed on audio database which contains 2,473 clips and the duration of each clip is 10 seconds or less. The efficient modeling of timbre feature for each individual instrument for music analysis and retrieval purpose is one of challenging task of most of the researchers. The modeling of timbre feature which has been done in [7] uses the Gaussian distributions over a space of Cepstral coefficients. The timbre model will give a lot of information which can be used to discriminate among various songs. An effective algorithm proposed in [8] for automatic classification of audio files based on MFCC, linear predictive coefficient (LPC) features uses support vector machine along with radial basis neural network in order to classify the audio file. This method provides better result for classification. The audio classification based on timbral feature which has been proposed in [9] makes use of Gaussian mixture model and clustering approach. A General Audio Data (GAD) classification based on Mel Frequency Cepstral Coefficient (MFCC) and linear prediction coefficient (LPC) has been performed in [10]. With the addition of segmentation-pooling scheme noise occur at boundary of audio segments during classification has been reduced. The emotion recognition based on the formant frequencies obtained from linear predictive filters has been proposed in [11]. This process was performed on corpus FAU abio and it makes use of MFCC feature and it shows significant result for emotion recognition. From the above discussion, it was clear that Mel Frequency Cepstral Coefficient (MFCC) provides better retrieval task. In this paper Weighted MFCC feature which is an extended form of MFCC has been used for audio retrieval tasks. Its extraction procedure has explained in the further sections.

3. METHODOLOGY

3.1 Content based Audio Retrieval

In the Content-based audio retrieval (CBAR) system the query audio feature is compared with the audio database feature in order retrieve the specific audio data. The first stage in this process involves extracting feature from the audio files. The features extracted from each audio file from the database are stored separately as feature database. When the query audio file has given, the feature has extracted from the query audio file and compares this feature with the features which was already extracted from the audio database.

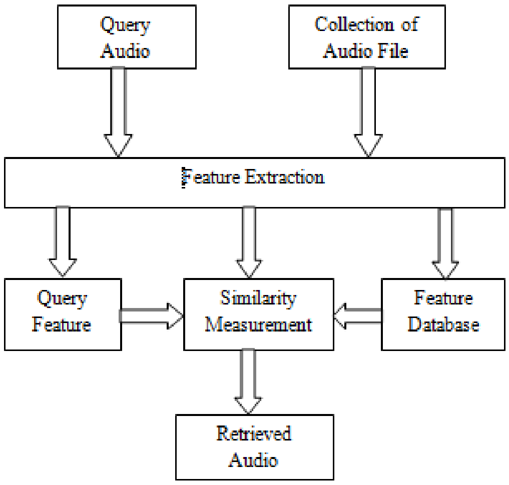

Fig. 1. Block diagram of content-based audio retrieval

When any audio features matches with the query feature, the audio file corresponding to that feature has been extracted. In this paper the feature which is extracted from audio file is weighted MFCC feature. This feature has been used for audio retrieval task. The similarity measurement used for comparing the audio feature is Euclidean distance which is the simplest and efficient way to compare the audio features.

3.2 Weighted Mel Frequency Cepstral Coefficient

Mel Frequency Cepstral Coefficients (MFCCs) which is traditional feature used for audio retrieval process constitutes Mel Frequency Cepstrum (MFC) which is used for the representation of audio signal [12]. The frequency is equally spaced on Mel scale compared with frequency bands in the normal cepstrum in MFC. In WMFCC feature the critical band energies are mapped onto the spectral weights to obtain weighted MFCC feature.

3.3 WMFCC Feature Extraction Process

The feature extraction process of WMFCC feature involves the mapping of spectral weights produced from LSF feature extraction process to the critical band energies which is produced from the early stages of MFCC feature extraction. In MFCC feature extraction process the first stage involves passing the audio signal x (n) shown in Fig.2 to the pre emphasis stage where the high frequency components can be preserved with the help of high pass filter.

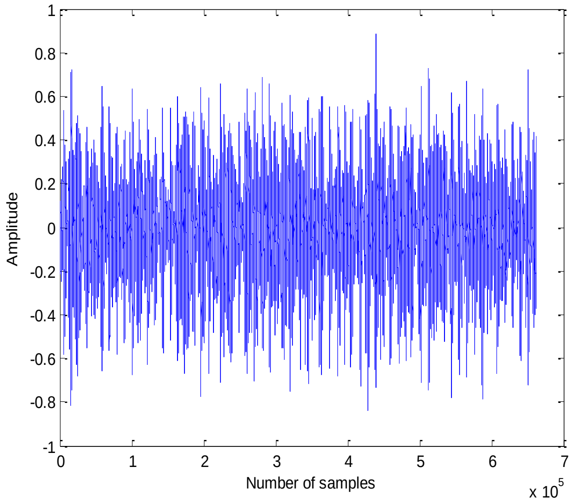

Fig. 2. Original signal

The high frequency components mostly get suppressed during the sound production mechanism of humans and the output of this stage will be pre-emphasized signal shown in Fig. 4 has been obtained [14].

$$x_2(n) = x(n) - a * x(n-1) \qquad (1)$$

Where x_2 (n) is the output signal and the value of *a* is usually between 0.9 and 1.0. The z- transform of the filter is given by [15],

$$H(z) = 1 - a * z^{-1} \qquad (2)$$

Next step involves breaking of audio signals into frames of 20~30 ms with an optional overlap of 1/3~1/2 of the frame size [15]. Then windowing stage involves product of each frame with the hamming window so that discontinuity near the boundary of audio segments can be reduced.

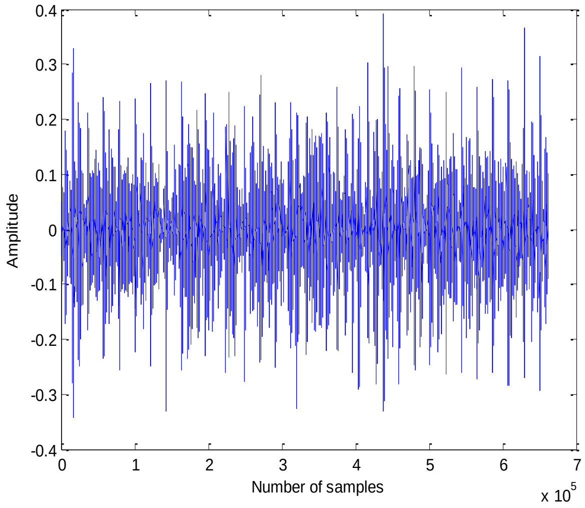

Fig. 3. Pre-emphasis signal

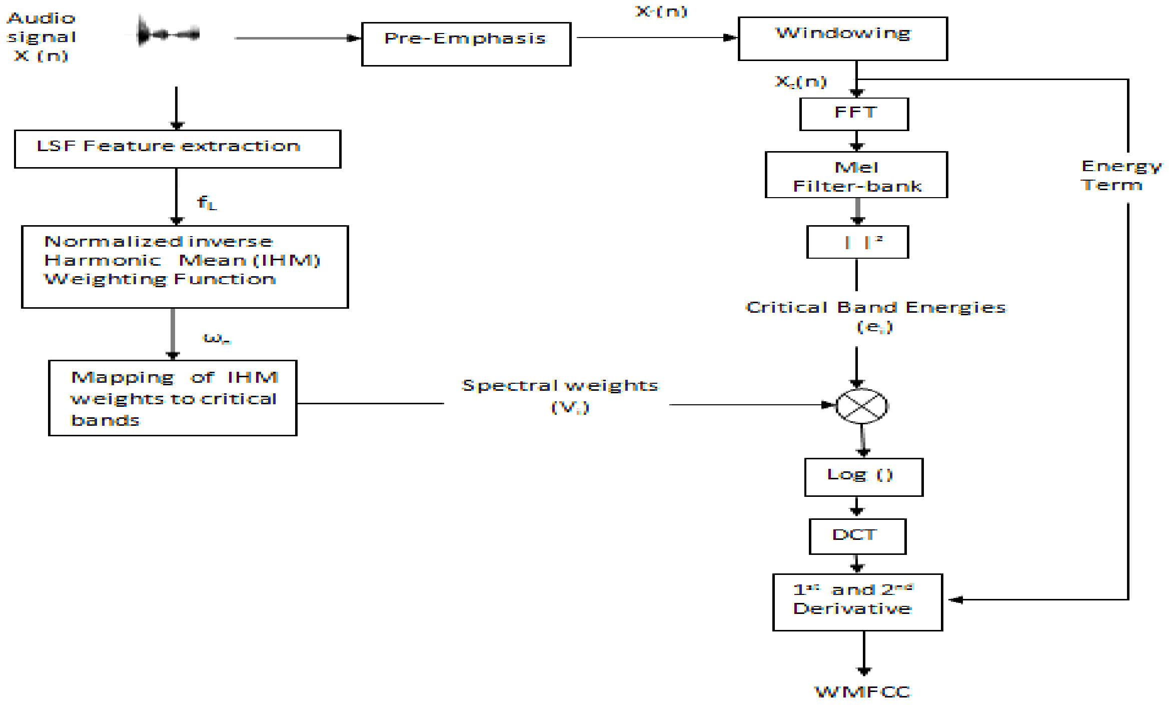

Fig. 4. Steps involved in MFCC feature extraction Process

If x (n) and w (n) are the input signal and windowing function respectively. Then the resulting signal is given by x (n)*w (n), and the hamming window function is given by [16],

$$w(n) = 0.54 - 0.46 * \cos\left(2\pi n / L-1\right), \quad 0 \le n \le L-1 \qquad (3)$$

Since it is difficult to identify the variation in the audio segments in spatial domain, the Fourier transform has been used to find spectral components of the audio segments which will provide better variation among the audio segments. The spectral component for each frame is shown in Fig. 5.

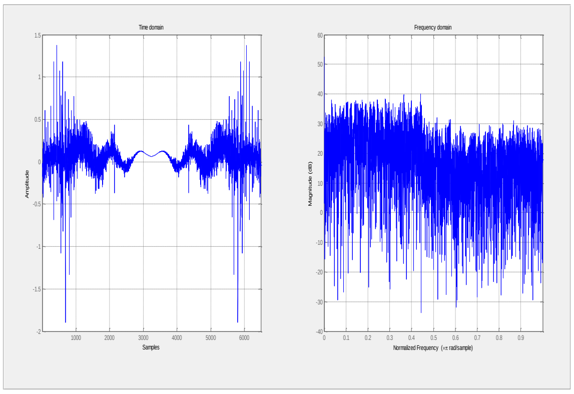

Fig. 5. Transformed signal

Mel filter bank, which make use of triangular band pass filter has been used in order to generate critical band energies.

These critical band energies can be obtained by multiplying the magnitude frequency response of each frame with the band pass filter on Mel scale. The Mel scale filter and the critical band energies are shown in Fig. 6 and Fig. 7. The main advantage of using the triangular band pass filter is that it can reduce feature size and smoothen the magnitude spectrum so that harmonics gets flattened.

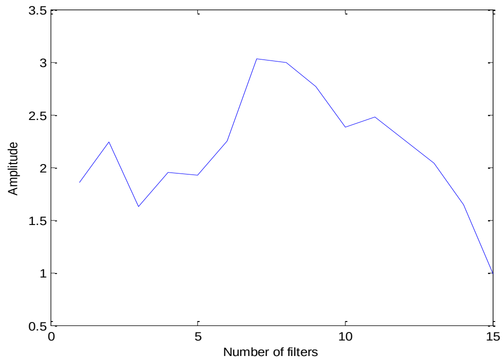

Fig. 6. Mel frequency curve

The relation between the Mel frequency scale and linear frequency scale is given by the following equation [17],

$$mel(f) = 1125 * \ln(1 + f/700) \qquad (4)$$

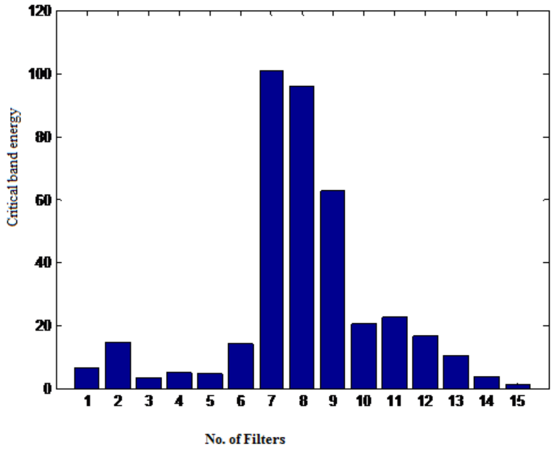

Fig. 7. Critical band Energy

Next step is to map the critical band energies with the spectral weights. The spectral weights can be obtained using line spectral frequency (LSF) extraction process. Line spectral frequency can be represented with the help of linear prediction (LP) filter which involves linear time invariant all pole filter of the form H (z).The LP filter can be represented as

$$H(z) = \frac{1}{A(z)} = \frac{2}{P(z)+Q(z)} \qquad (5)$$

P (z) and Q (z) are polynomials which have p/2 zeros on the unit circle. The p zeros of the polynomials represents LSF feature. As the neighboring LSF features are close to each other, it can be reduced to their mean value using inverse harmonic mean.

The Inverse Harmonic Mean function (IHM) w_i is defined by

$$w_i = \begin{cases} \dfrac{1}{f_L^{i+1} - f_L^{i}} & i=1 \\ \dfrac{1}{f_L^{i} - f_L^{i-1}} + \dfrac{1}{f_L^{i+1} - f_L^{i}} & i=2,3,\ldots,p-1 \\ \dfrac{1}{f_L^{i} - f_L^{i-1}} & i=p \end{cases} \qquad (6)$$

Where f_L^i represents line spectral frequency for p^th order filter. The linear interpolation of IHM weights can be defined by

$$v_j = \frac{(w_n - f_L^{n-1}) + w_{n-1}(f_L^{n} - m_i)}{f_L^{n} - f_L^{n-1}}, i = 1,2,\ldots,N_B \qquad (7)$$

Where N_B is the number of critical bands. The normalization of V_i value will give the spectral weights which is shown in Fig.8.

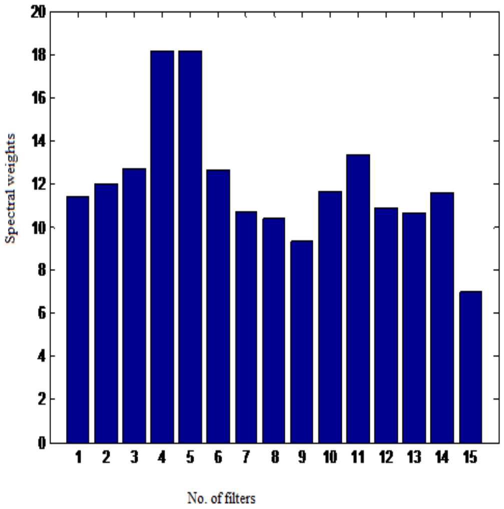

Fig. 8. Spectral weights

After the spectral weights has found out, it has to be mapped with the critical band energies. Then the mapped feature has to be compressed using logarithmic function and then discrete cosine transform has been applied in order to de-correlate the spectral feature. The formula for calculating WMFCC is given by [11],

$$f_w^j = \frac{1}{N_B}\sum_{i=1}^{N_B} \log(v_i e_i)\cos(i-0.5)\frac{j\pi}{N_B}, j = 1,2,\ldots,N \qquad (8)$$

Where N is the number of WMFCC feature extracted. The obtained WMFCC can be further extended to first and second order derivatives. More detailed information about audio segments can be obtained using the derivative of the WMFCC.

3.4 Similarity Measurement

After the feature has been extracted from the audio files, the next step involves comparison of the query audio file with the database audio files. This comparison is used to find the relevant audio file. The most commonly used comparison method is Euclidean distance method which is used to compare the two data and produce the distance. If the distance is of larger value then the query audio file is not similar to the database audio file. If the distance is of smaller value then the query audio file is closer to the database audio file. Suppose the distance value is zero then the query audio file is exactly matching with the database audio file. The Euclidean distance is defined as the root of the sum of the squared difference between the pairs of feature vector elements. Let x and y be the two vectors. Then the Euclidean distance between two vectors is given by the formula [17],

$$d(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \qquad (9)$$

3.5 Performance measure

The performance of content based audio retrieval system has to be evaluated. This will provide the consistency and efficiency of the system. Most common evaluation method used for audio retrieval task is precision and recall [18]. The precision value can be defined as the ratio of number of relevant audio to number of retrieved audio and the recall is defined as the number of relevant audio selected to the total number of relevant audio. Precision and recall graphs are mainly used to measure audio retrieval ability.

$$\text{Precision} = \frac{\text{Number of relevant audio}}{\text{Number of retrieved audio}} \qquad (10)$$

$$\text{Recall} = \frac{\text{Number of relevant Audio}}{\text{Total number of relevant Audio}} \qquad (11)$$

4. Results and Discussion

The audio retrieval task has been performed with the help of MATLAB software and the database used for the audio retrieval task is GTZAN dataset. It consists of 1000 songs which can be classified into ten genres. The ten genres are blues (blu), Classical (clas), country (con), disco (dis), Hip-hop (hip), jazz (jaz), metal (met), pop, reggae (reg), and rock (rok). The datasets have 100 songs per genre all of which are single-channel and sampled at 22.05 kHz [18]. The entire database has been used and WMFCC feature has been extracted from each audio file from the GTZAN datasets and stored it separately as feature database. From query audio file WMFCC feature has been extracted and stored it as feature vector. Both the feature vector and the feature database files are compared and the similar features are found. Then the corresponding audio file has been retrieved. Results are obtained for different stages of WMFCC feature extraction process and the retrieval performance has been found out by considering query audio and thousand audio file from GTZAN database.

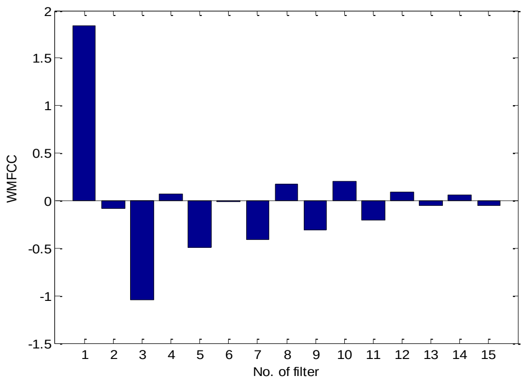

Fig. 9. WMFCC feature

In Fig. 9, the WMFCC extracted for a query audio has been shown. Top k-audio has selected (k=3, 5, 10) which indicates the number of relevant audio file being retrieved. For each category the number of relevant audio which has been retrieved has calculated with the help of the mean precision value. The Queries taken for this evaluation consist of 100 audio clips and also for each audio category, number of retrieved files has calculated using average recall value.

The mean precision value for the entire audio category for top-k (k=3, 5, 10) audio files has been calculated and the results are shown in Fig.10.

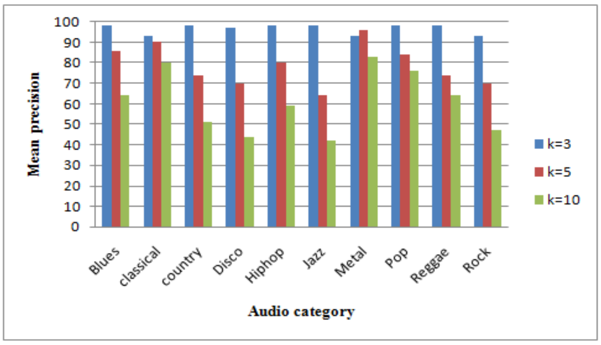

Fig. 10. Mean precision for top-K audio

The average recall rate for all the given query audio files has been calculated and the results are shown in Fig.11 which indicate the recall rate gets better after certain number of relevant audio files.

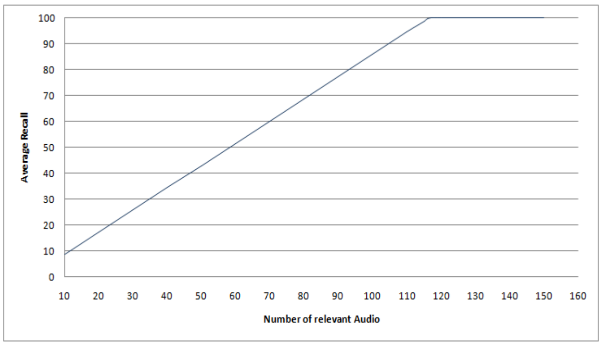

Fig. 11. Average recall rate

From each audio category, one audio data file has selected as a query and for that query the precision rate has calculated by considering the entire database For the sample queries, the mean precision rate has obtained for top-3 relevant audio data retrieved is shown in Table I. The average precision rate is 96.40%, which is an acceptable one.

Table I Precision rate of Various Queries for Top-3 Audio Files

Query	Retrieved Audio Files for K = 3			Precision (in %)
blu.00006	blu.00006	blu.00010	blu.00099	98
clas.00011	clas.00001	clas.00093	blu.00068	93
con.00001	con.00001	con.00005	con.00008	98
dis.00005	dis.00005	dis.00017	dis.00073	97
hip.00021	hip.00021	hip.00098	hip.00095	98
jaz.00005	jaz.00005	jaz.00016	jaz.00026	98
met.00004	met.00004	met.00007	met.00006	93
pop.00001	pop.00001	pop.00052	pop.00074	98
reg.00004	reg.00004	reg.00017	reg.00056	98
rok.00006	rok.00006	rok.00048	rok.00034	93
Average				**96.4**

5. CONCLUSION AND FUTURE WORK

In this work, the importance of audio retrieval system has been described. The feature extraction which is one of the important tasks in audio retrieval system has also been reported. Based on the literature survey the significance of feature extraction process has been addressed and it has been found out that only the sufficient feature will make audio retrieval task more effective and it has been concluded that MFCC feature provides relevant information for audio retrieval. Among the variants of MFCC feature, WMFCC feature which is a cepstral feature provides meaningful information about audio file. The results for the various stages of WMFCC feature extractions have been obtained. With the help of WMFCC feature, audio retrieval task has been performed on the GTZAN database. The mean precision rate of about 96.40% and better recall performance has been obtained. In the future, the other audio features which will give meaningful information about audio signal are being taken and their feature extraction techniques will be analyzed.

6. REFERENCES

[1] Tomi Kinnunen, Rahim Saeidi, "Low-Variance Multitaper MFCC Features: A Case Study In Robust Speaker Verification", IEEE Transactions On Speech, Audio And Language Processing,2012

[2] Masayuki Suzuki, Takuya Yoshioka, "MFCC Enhancement Using Joint Corrupted And Noise Feature Space For Highly Non-Stationary Noise Environments", ICASSP 2012

[3] Dalibor Mitrovi´C, Matthias Zeppelzauer, and Christian Breiteneder, "Features For Content-Based Audio Retrieval", Advances in Computers Vol. 78, pp. 71-150,2010

[4] Guodong Guo and Stan Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector

Machines", IEEE Transactions on Neural Networks, Vol. 14, No. 1, January 2003.

[5] Riccardo Miotto and Gert Lanckriet, "A Generative Context Model for Semantic Music Annotation and Retrieval", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 4, May 2012.

[6] Hung-Yi Lo, Ju-Chiang Wang, Hsin-Min Wang, "Homogeneous Segmentation and Classifier Ensemble for Audio Tag Annotation and Retrieval", National Science Council of Taiwan, 2008.

[7] Jean-Julien Aucouturier, François Pachet, And Mark Sandler ,"The Way It Sounds": Timbre Models For Analysis And Retrieval Of Music Signals, IEEE Transactions On Multimedia, Vol. 7, No. 6, December 2005.

[8] P. Dhanalakshmi, S. Palanivel, V. Ramalingam, "Classification of audio signals using SVM and RBFNN", Expert Systems with Applications Elsevier, 2008.

[9] Thibault Langlois, Gonc¸Alo Marques, "A Music Classification Method Based On Timbral Features", International Society for Music Information Retrieval Conference (ISMIR) 2009.

[10] Dongge Li, Ishwar K.Sethi, "Classification of General Audio Data for Content Based Retrieval", Pattern Recognition Letter, Elsevier 2001.

[11] Elif Bozkurt, Engin Erzin, "Formant position based weighted spectral features for emotion recognition", Elsevier 6 May 2011.

[12] G. Salton. The SMART Retrieval System. Prentice Hall,Englewood Cliffs, NJ, 1971.

[13] D. Sturim, D. Reynolds, E. Singer, J. Campbell. "Speakerindexing in large audio databases using anchor models." Proc. Of ICASSP, vol. I, pp. 429–433, 2001

[14] George Tzanetakis, "Musical Genre Classification of Audio Signals", IEEE Transactions On Speech And Audio Processing, Vol. 10, No. 5, July 2002.

[15] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet, "Semantic Annotation and Retrieval of Music and Sound Effects", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, February 2008.

[16] Tao Li, Mitsunori Ogihara, Qi Li, "A Comparative Study on Content-Based Music Genre Classification", SIGIR'03, 2003.

[17] Chandika Mohan Babu, Manish Puri and Anamika Das, "Effective principle analysis of speech recognition systems using MFCC and time domain approach for isolated word for training phase spectrum", World Journal of Science and Technology, April 2012.

[18] Atanas Ouzounov, "Cepstral Features and Text-Dependent Speaker Identification –A Comparative Study", Cybernetics and Information Technologies Volume 10, No 1, 2010.

[19] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora, "Audio Classification Based on MPEG-7 Spectral Basis Representations", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 5, May 2004.

[20] Vibha Tiwari, "MFCC and its applications in speaker recognition", International Journal on Emerging Technologies, November 2009.

[21] A. Ghias, J. Logan, D. Chamberlin: Query By Humming - Musical Information Retrieval in An Audio Database", Proc. ACM Multimedia Conference, pp.231-235, Anaheim, CA, 1995.

[22] J. Foote: \Content-Based Retrieval of Music and Audio", Proc. SPIE'97, Dallas, 1997.

[23] Elias Pampalk, Arthur Flexer, and Gerhard Widmer, "Improvements of audio-based music similarity and genre classification," ISMIR, 2005.

[24] M. S. Lewicki, "Efficient coding of natural sounds", in Nature Neuroscience, Vol. 5,No. 4, pp 356-363, 2002.

[25] S. Sundaram and S. Narayanan ,"Analysis of Audio Clustering using Words". Presented at ICASSP, Hawaii, USA. 2007.