IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. II (July – Aug. 2015), PP 86-91
www.iosrjournals.org

Singer's Voice Identification and Authentication based on GFCC using K-means Clustering and DTW

Swati Atame[1], Prof. Shanthi Therese S[2]
[1](.Department of Computer Engineering, Mumbai University,India)
[2](Department of Information Technology, Mumbai University, India)

Abstract: *The use of computers has increased to a very large extent due to its immense use and growing technologies that gives rise to transfer of digital media over longer distances. As the audio data may travel over large networks it needs to be purified or cleaned during recording or while the processing the data so that it can be identified. Cleaning is related to removing the noise that may exist in the signal. There are 2 ways in which the audio can be cleaned. First, recording can be done in clean environment that includes no noise at all, even the fans may create disturbances, recording done in a closed room. But feature extraction in this case is more suitable using MFCC. Second, if recording is already done and needs to processed and if it contains lot of noise than some noise reduction technique such as wavelet or filter can used to eliminate the noise. In this paper, we will implement Gamma tone Frequency Cepstral Coefficient (GFCC) to eliminate the noise components. This is due to the fact that MFCC does not give required accuracy in the presence of noise. It works well in clean environment. On the other hand, GFCC gives the required performance and accuracy in clean as well as in the presence of noise. This paper compares auditory feature based methods namely GFCC and MFCC and concludes which method is more suitable in which environment. This resulting GFCC features are used by k-means clustering to group the similar voices in to clusters and thus identify the singer. Dynamic time warping is used to align or match two sequences that may vary in time or speed. It is a pattern matching technique that will authenticate the person a claimed identity.*

Keywords: *Feature extraction, Gammatone Frequency Cepstral Coefficient (GFCC), Audio clustering, MFCC, noisy environment, Dynamic time warping(DTW).*

I. Introduction

Singer can be identified from the his voice that is enrolled during the training session. The voice input can be restricted to text dependent or text independent. In this paper, we will deal with text dependent voice data. Usually, features are extracted using Mel-frequency cepstral coefficients (MFCCs) which contains bank of triangular filters. This is due to the fact that MFCC does not give required accuracy in the presence of noise. It works well in clean environment. On the other hand, GFCC gives the required performance and accuracy in clean as well as in the presence of noise. For that purpose, Gamma tone filter bank(GTFB) is used to model cochlear filter more accurately. Overlapped band pass filters are used to model GTFB. The resulting features are called gamma tone frequency cepstral coefficients (GFCC). This resulting GFCC features are used by k-means clustering to group the similar voices in to clusters and thus identify the singer.

Identification is done in phases:

1) Feature Extraction
2) Modeling
3) Decision making

Feature are extracted from the audio signal is done using GFCC. K-means for speaker modelingwhich is used to classify the feature in to groups, by petitioning the input data in to accurate distributions of individual singers. Decision making refers to recognition decisions depending upon the probability of observing feature frames given a speaker model.

II. System Model

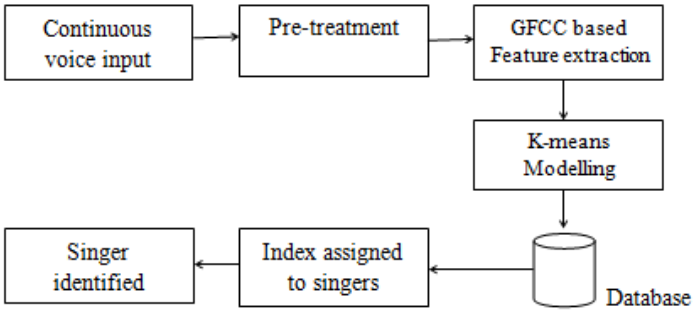

Fig 1. Singer identification architecture

In the above system, the continuous song is given as input to the system which is than given to the pre-treatment block. This input is given by many singers whose voice is identified using the system.

2.1 Pre- treatment[1][2]

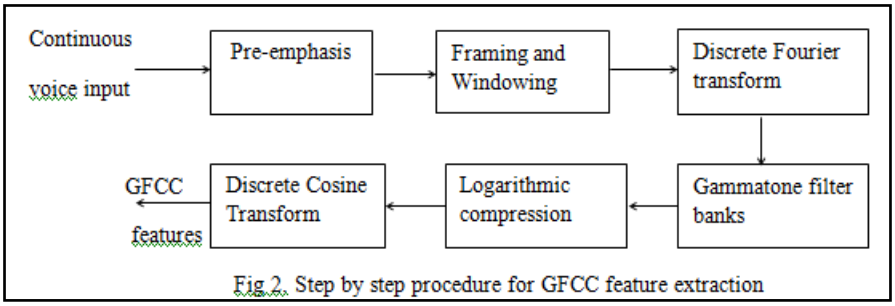

Fig 2. Step by step procedure for GFCC feature extraction

Pre-emphasis: The audio signal is converted to digital signal using pre-emphasis. It is used to emphasize high frequency components and suppress low frequency components present in the audio signal. It flattens the audio input before evaluating the spectral components. High pass filter is used to implement pre-emphasis stage and is shown in the following equation:

$$H_{preemp}(z) = 1-a_{preemp}\, z^{-1} \quad (1)$$

2.2 GFCC Feature Extractor

a. Framing: The output of pre-emphasis stage is high frequency components in digital form which is given as an input to the framing stage. This block divides the continuous signal into sequence of frames till the end of signal occurs. It is not possible to process entire signal at once for feature evaluation so framing of signal is essential. Framing helps to independently examine individual frames represented in the form of features. Since frames are considered to be static or stationary features can be evaluated easily as compared to that of dynamic or continuous signal. Short segments of time 20-30 ms are considered to be stationary.

b. Hamming window: It is used to avoid the discontinuities at edges that is at the beginning and ending of each frame, windowing is applied to each and every frame. This is widely used window which is described in following equation:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right) \quad (2)$$

c. Discrete Fourier Transform computation: Fast Fourier Transform algorithm of DFT is used for each frame to obtain its magnitude as phase does not contain any information required for feature evaluation.

d. Gamma tone filter banks stage [1][2] : Another technique used for audio signal identification based on FFT feature extraction is Gamma tone Cepstral Coefficients (GFCC). Instead of Mel bank filter we use Gamma tone filter banks that are implemented as a collection of band pass filters. Frequency selectivity property of basilar membrane is modeled by filter. Human auditory system is modeled as a collection of band pass filters is a technique based on Gamma tone filter bank (GTFB). Impulse response of each filter is given by the following equation:

$$g(t) = at^{n-1} e^{-2\pi bt}\cos(2\pi f_c t + \Psi) \quad (3)$$

where,
a= constant value and which is equivalent to 1,
n = filter order of the filter
Ψ= phase shift ,
fc =centre frequency, **b** = filter bandwidth in Hz.

The filters Equivalent Rectangular Bandwidth (ERB) is used to compute centre frequency and the bandwidth of each Gamma tone filter. ERB is related to centre frequency of the filter. The following equation represents the relation between ERB and centre frequency.

$$\text{ERB}(f_c) = 24.7\left(4.37\frac{fc}{1000} + 1\right) \quad (4)$$

The filters centre frequency and bandwidth of a Gammatone filter should be almost 1.019 times to that of Equivalent Rectangular Bandwidth (ERB) which is represented in the following equation:

$$b = 1.019\text{ERB} = 1.019\left(24.7\left(4.37\frac{fc}{1000} + 1\right)\right) \quad (5)$$

Gammatone filter (n=4) with 4th order auditory filter is shown below

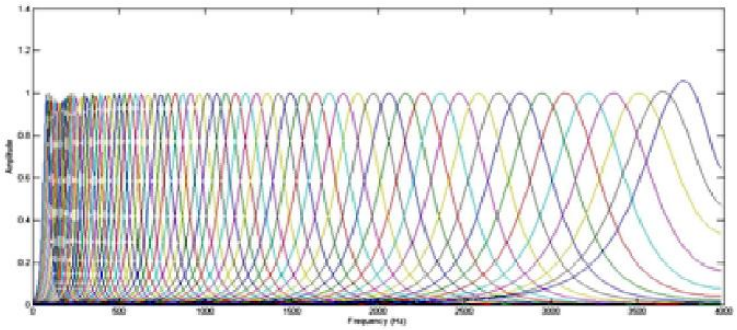

Fig 3. Gamma tone filter bank

e) Logarithmic compression stage

By applying the log compression stage we get log energy spectrum. To perform the simulation of human voice pitch at a particular intensity and to segregate the source by filter and vocal cords the is represented by vocal tract.

(f) Discrete cosine Transformation (DCT) stage

High uncorrelated feature vectors are generated by applying DCT on filters output. The truncated output is nothing but the featured vector.[1][2]

2.3 K-means clustering:

Grouping of identical elements with similar properties together is known as clustering. The output what we get are nothing but clusters. The main advantage of grouping the data into clusters is, it helps to arrange huge amount of data in to manageable pieces. It also helps to determine the pattern s of interest to the user. Also it is considered essential as it is one of the widely used techniques used to determine the groups of interest. Depending upon some predefined criteria, the input data set is divided in to partitions. This technique helps to distinguish between similarities and dissimilarities into as to compute useful inferences from them.

The k-means algorithm works on the concept of dividing the dataset of feature vectors into centroids. Initially, centroids of clusters among feature vectors are chosen. Assignment of each feature vectors to the closest centre point and new centre points are evaluated. The above steps are performed till the time the exit condition is fulfilled, it means there minimum distance between mean square error and cluster centre points or in other words there is no more changes in the cluster-assignment. This technique is same as that of expectation-maximization algorithm for Gaussian mixture distributions repetitive improvement approach that is implemented by both the algorithms. One more similarity is that both the methods models the data sets using cluster centres.[3][4].

Algorithm:

Step 1 :Begin by setting the value of k to the number of clusters.

Step 2 :Perform initial division and classification of the dataset in to k clusters. Assignment of dataset samples are assigned to the cluster as follows:

a. First training dataset sample value is taken to form single elements cluster.

b. The remainder (N-k) of training data set sample is assigned to the group or cluster with closest centre.
c. Re-computation of new centre of the gaining cluster is done after performing each assignment.

Step 3: Each dataset sample value is taken in series in order to evaluate the distance between the sample value and each cluster centre .If the currently selected sample is not nearest to the centroid of the cluster than switch the sample value to the cluster centre to which it is closest and update the cluster centroid.

Step 4: step 3 is iterated till the time there are no new changes in assignments to clusters and algorithm is deemed to have converged.

Distance measures for k-means are Euclidean distance and Manhattan distance is used to compute distance between the similar elements. The drawback of this algorithm is that the number of clusters needs to be evaluated beforehand so if cluster number is not selected properly it may result in poor outcomes. Applications of K-means algorithm is useful for undirected knowledge discovery and is relatively simple. K-means is widely used in most of the areas, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others. In Cartesian coordinates, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

III. Authentication Process

As we have already discussed identification methods we will mow process with the authentication by making use of DTW algorithm. Measures similarity between 2 signals. Authentication is done by calculating the following:
Local distance using Euclidean
Global distance using dynamic programming called as DTW[5][6]Suppose we have 2 time series Q and C , of length n and m for computing local distance where,

Q = q1, q2,…, qi,…,qn **(1)**
C = c1, c2,…, cj,…,cm **(2)**
d (qi,cj) = (qi - cj) **(3)**

Global distance is calculated using the following equation:

D(i, j) = min[D(i-1, j-1),D(i-1, j),D(i, j-1)] + d(i, j) **(4)**

DTW conditions:
Monotonic condition, Continuity condition, Boundary condition, Adjustment window condition, Slope constraint condition **.**

IV. Experiments

4.1 Database: For demonstrating the identification process a database of10 singers is considered. 5 song samples were recorded for each singer. So the database consists a total of 50 songs sung by 10 singers as a whole. Some voice recording is done in clean environment and some in noisy environment. The duration of continuous song input is for 10ms.Audacity software is used for recording and editing the song input. Same content song are sung by different singers. Format factory is used to extract audio signal from the videos available on the YouTube. The audio signal is than stored in .wav format in mat lab for further processing.

4.2 Methods
Algorithm

1. Continuous song is recorded using microphone.
2. The recorded continuous voice samples are transformed into .wav format.
3. The recorded sound files are loaded from the matlab database.
4. Singing parts and non singing parts are identified.
5. Vocal part is retained as it contains most of the essential information that helps in feature representation.
6. Features are extracted from the test file for identification and recognition.
7. Features are also extracted from all the training files already stored in the database.
8. The centre point of the test file is obtained using K-means clustering algorithm.
9. The centre point of the training file is obtained using the same algorithm used above.
10. Euclidean distance between test sample and each of the training samples is obtained.

11. The training sample that has the minimum distance with the test sample is found.
12. The sample corresponding to lesser distance is most likely the singer of the test song.

4.3 Results

1. **Comparison between MFCC and GFCC**: As we know MFCC works fine in clean environment but may degrade in performance in presence of noise in environment. In order to demonstrate this scenario a sample voice is recorded in noisy environment and is given to MFCC feature extraction algorithm and following results are observed:

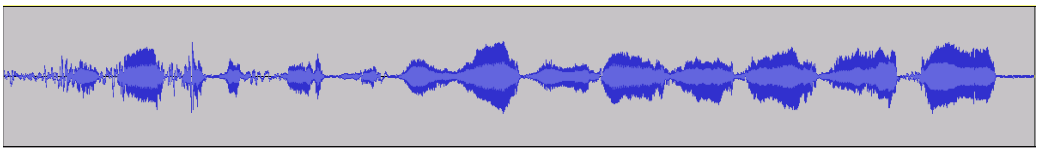

Fig 4. Continuous song input of singer 'vanshi9'

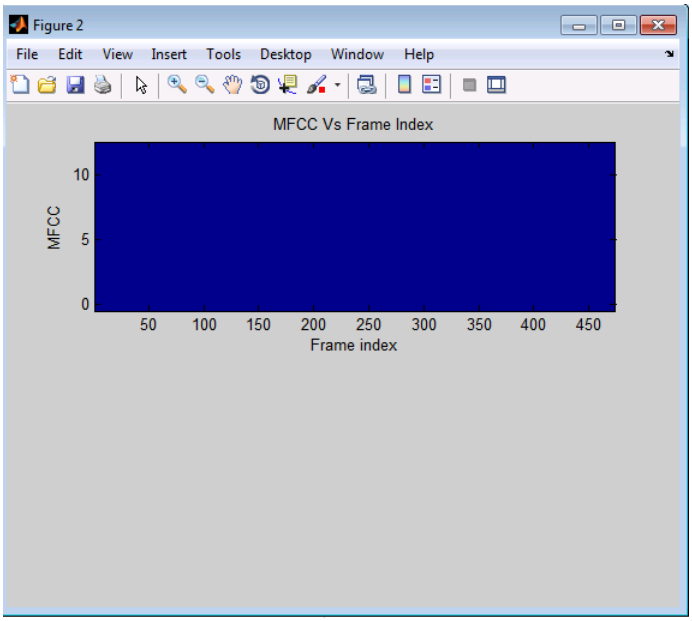

Fig 5. MFCC feature extraction

It has been seen that in some cases due the presence of noise the results are very much severe that the output is not displayed at all and all values are displayed as NaN. The features are not extracted due to the presence of noise in the input signal.

While on the other hand, when we use GFCC the following plot is obtained for the same continuous song input.

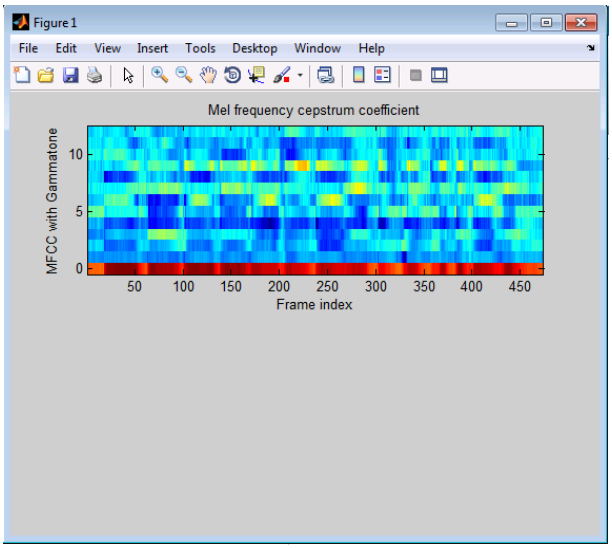

Fig 6. GFCC feature extraction.

By implementing the GFCC feature extraction, the features are extracted even if noise is present and at the output we get the extracted features as Gamma tone cepstral coefficients that is 13 co-efficients for input signal .

2. **K-means clustering**: A database is grouped in cluster of 4. The number of singers is 4 respectively. A sample of 3-4 songs for each singer is recorded in different environment. The song duration is for 10-15 ms.The input matrix is of size (n*1) and the output that is produced is classified into group of 4 clusters.

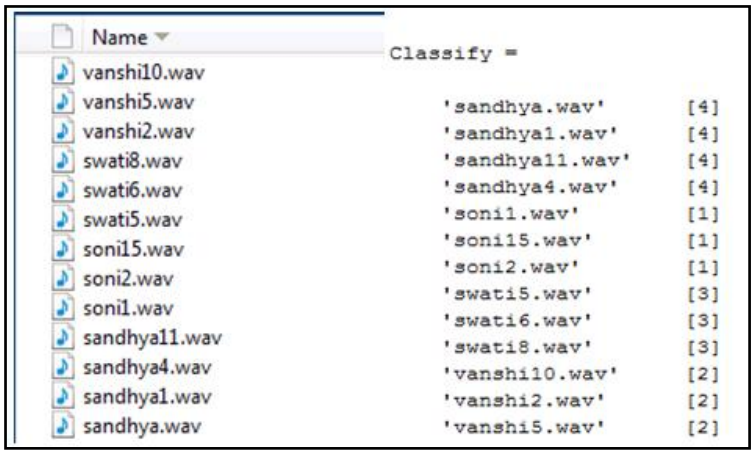

Fig 7a. Input dataset for clustering **Fig 7b.** Clustered Result

The above figure shows the data set for continuous voice input which needs to clustered and result of performing K-means clustering. It is observed that the clustering is performed accurately in to groups of 4 clusters as value of k was initially defined. Clustering thus helps in the identification of the singer. The singer is assigned an index with which identification is done.

3. **Authentication using DTW**: For simplicity we have taken 11 singers for authentication.

```
dtwDist =

  Columns 1 through 5

    [1.6479e+04]    [3.5085e+04]    [1.6215e+04]    [1.5030e+04]    [1.5466e+04]

  Columns 6 through 11

    [2.9364e+04]    [1.4474e+04]    [2.0361e+04]    [1.5800e+04]    [1.6100e+04]    [0]

AUTHENTICATED with vanshi4.wav
```

Fig 8 a. Distance between test and each training sample

The distance is calculated using DTW algorithm. The above snapshot shows distance between test and each training sample. It shows the differences where match is not found otherwise represents the minimum distance 0 in this case for match found.

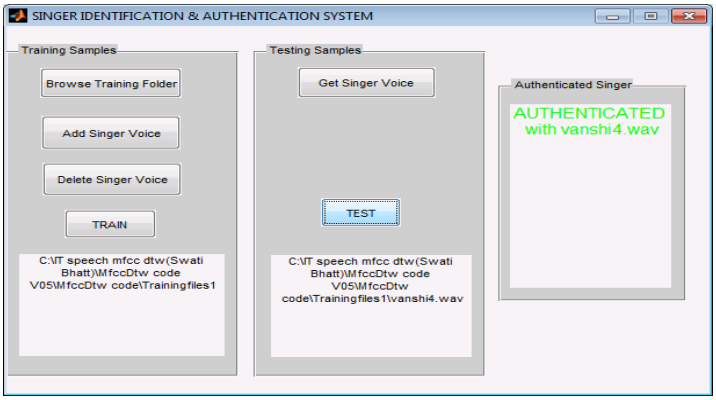

Fig 8 b. GUI for Singer authentication

The above GUI is used for authentication of the singer. In training phase, we need to browse training folder and then train the training folder. We can also add Singer voice to training or delete Singer voice from the training folder. In the testing phase, we get the test sample and click TEST to get the output of pattern matching between testing and training sample. If singer is identified, then message displaying "AUTHENTICATED with singer name.wav" will be displayed otherwise message displaying "File Not Found" will be displayed.This technique give accuracy of 85% in authentication the singer's voice.

V. Conclusion and future work

The conclusion of this paper is that the continuous voice (singer) can be identified using GFCC and compare with MFCC for feature extraction and DTW for pattern matching is used for authentication process. We observed that MFCC works fine with clean environment but issue arises when noise is present. Sometimes noise may affect more adversely. On the other hand, GFCC works very well in the noisy environment. The future scope of the proposed method can be extended to neural networks where the dataset size would be huge.

Acknowledgment

I express my gratitude to all my friends and faculty members of the Computer engineering Dept of Shree L.R .Tiwari of engg and technology, Maharashtra, India, for their support and enthusiasm . I am very grateful to Mrs. Madhuri Gedam(Assistant professor in Information Technology department) Shree L.R .Tiwari of engg and technology, Mumbai , for her guidance in my work. Last but not least I am very thankful to our HOD Prof. Vinayak Shinde for giving opportunity to work in the field of data hiding and retrieval.

References

[1]. Md. Moinuddin, Arunkumar N. Kanthi, **"Speaker Identification based on GFCC using GMM",**International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163,Volume 1 Issue 8 September 2014.

[2]. VenkateshDeshak, ArunkumarKanthi, "**Audio Clustering**", International Journal of Innovative Research in Advanced Engineering (IJIRAE) , ISSN: 2349-2163 Volume 1 Issue 8, September 2014

[3]. H. Zha, C. Ding, M. Gu, X. He and H.D. Simon,"**Spectral Relaxation for K-means Clustering**", Neural Information Processing Systems vol.14 (NIPS 2001).pp. 1057-1064, Vancouver, Canada. Dec. 2001.

[4]. J. A. Hartigan and M. A. Wong, **"A K-Means Clustering Algorithm"**, Applied Statistics, Vol. 28, No. 1, p100-108, 1979.

[5]. Abdul SyafiqAbdullSukor, **"Speaker identification system usinsg MFCC procedure and Noise reduction**",University Tun Hussein Onn Malaysia, January 2012.

[6]. ShivankerDevDhingra**,** GeetaNijhawan , PoonamPandit,"**Isolated Speech recognition using MFCC and DTW**", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 8, August 2013.