Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Research Paper - A Hybrid Approach for Voice Recognition Using MFCC, VQ and K-Means Algorithm

Nisha Saini
M.Tech Scholar, Department of Computer Science & Engineering
Deenbandhu Chhotu Ram University of Science & Technology (DCRUST), Sonepat, India

Dr. Dinesh Singh
Asst. Professor, Department of Computer Science & Engineering
Deenbandhu Chhotu Ram University of Science & Technology (DCRUST), Sonepat, India

Abstract: The speech or voice of any human being is his/her unique personal characteristics. No two have almost identical voice there are some features which must be present in one voice and missing from another voice. In order to identify this uniqueness a robust and efficient technique is required so that we can accurately identify the genuine voice from the bunch of fake voices. The process of recognizing a voice of a given speech from the group of given speakers is called speaker identification. This paper proposes a new improved and efficient technique for voice recognition system by applying Mel Frequency Cepstral Coefficients (MFCC), Vector Quantization, K-means and Euclidean distance technique.

Keywords: Voice Recognition, MFCC, Vector Quantization, Euclidean Distance

I. INTRODUCTION

Biometrics refers to "automatic recognition of individuals based on their physiological and behavioral characteristics." Today there is need for simple and secure access control mechanism for authentication of a person. Biometrics technique for authentication is the solution for this. With biometric security we need not carry different types of ID cards or remembering various passwords or keys, our body parts are sufficient for authentication purpose. Even more secure & robustness can be achieved if we combine biometric security with other security modes such as key, password or any identification card.

There are various application areas where we can use biometric security such as banking application where signature recognition identify our authentication or our voice in case of phone banking, driving license and passport where our image (picture) is used for identification.

There are many advantages of using biometric security over normal security for authentication [1]. These advantages are listed below:

1. Reliability: The biometric cannot be lost and stolen, therefore it provides more reliability.
2. Uniqueness: Another important feature of biometric is that it is unique for everyone. For example the fingerprint of one person is different from another person.
3. Nothing to remember or carry: With biometric we need not carry different types of ID cards or remembering various passwords or keys, our body parts are sufficient for authentication purpose.

The biometric security works as follows:

First capture the biometric image, extract features & create database of these features. Now when any person comes for biometric verification then his/her biometric image is collected again features are extracted and then these features are matched with the features already stored in database and then matching is performed. One of important biometric system is human voice.

The main aim of this paper is voice or speaker identification. In this method we compare voice of any unknown person with database of known voices, It then returns the message whether voice matches or not. In this paper, we use hybrid approach of "Mel Frequency Cepstral Coefficients (MFCC) technique, Vector Quantization, K-means algorithm and Euclidean distance."

II. OVERVIEW OF WORK

Voice recognition identification [2] is the process of recognizing the speaker automatically. The speech waves of individual voice form the basis of identification of speaker. We can use voice identification in multiple application areas such as telephone banking, shopping through telephone, access to database information and voice mail. One of the powerful applications of voice recognition is for security purpose where a person can enter his/her voice for authentication.

The voice recognition system [3] is broadly divided into following two categories- "Speaker identification & Speaker verification". The process of recognizing a voice of a given speech from the group of given speakers is called speaker identification. The speaker whose maximum voice characteristics are matches with the stored voice is identified & the speaker whose voice characteristics are not matched is eligible for new entry in the database.

The main building blocks [4] of any speaker or voice recognition systems are "*feature extraction* and *feature matching*." As the name suggests the feature extraction block of voice recognition system is to extract the unique characteristics of any voice signal. These characteristics represent the identity of any person. On the other hand feature matching block is use to match the unknown voice signal with the database of given voice and identify the claim person.

A. Voice Feature Extraction [5, 6]

The voice feature extraction block of voice recognition system is to extract the unique characteristics of any voice signal. These characteristics represent the identity of any person. It converts voice into speech wave as shown in figure 1. It uses the concept of digital signal processing (DSP) tool for converting voice signal into speech wave.

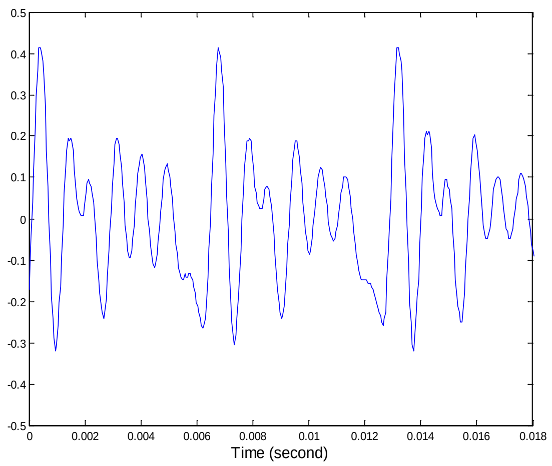

Figure 1: Example of speech wave

The voice signal can be converted into speech wave for extracting unique features by using one of the following methods:

1. Linear Prediction Coding (LPC)
2. Mel-Frequency Cepstrum Coefficients (MFCC)

In this work we use MFCC technique for feature extraction because it is well studied and popular method for voice feature extraction

MFCC [7] described by authors as "MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech." The MFCC is generraly expressed as *mel-frequency* scale.

B. Voice Feature Matching

Voice feature matching [8] block is use to match the unknown voice signal with the database of given voice and identify the claim person. Voice feature matching works as - When any person comes for voice verification then his/her voice signal is collected, then features are extracted and then these features are matched with the features already stored in database and then matching is performed.

The voice signal can be matched with voices already stored in database by using one of the following methods:

1. Dynamic Time Warping (DTW)
2. Hidden Markov Modeling (HMM)
3. Vector Quantization (VQ).

In this work, we use the VQ approach [9] because it gives high accuracy & its implementation is easy.

III. PROPOSED WORK

The main aim of this paper is voice or speaker identification. In this method we compare voice of any unknown person with database of known voices, It then returns the message whether voice matches or not. In this paper, we use hybrid approach of "Mel Frequency Cepstral Coefficients (MFCC) technique, Vector Quantization, K-means algorithm and Euclidean distance."

MFCC [10] is used to extract unique features from voice entered by the user. For estimation of probability distributions of the computed feature vectors, we use the concept Vector Quantization (VQ). VQ represents the centroid of extracted features. The k-means algorithm is used for dividing the features into n clusters; each cluster represents the unique characteristics of any feature of voice [11].

The steps performed in K-means algorithm are shown in figure 2 below:

1. Clusters the data into k groups where k is predefined.
2. Selects k points at random as cluster centers.
3. Assigns objects to their closest cluster center according to the Euclidean distance function.
4. Calculates the centroid or mean of all objects in each cluster.
5. Repeats steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

Figure 2: Steps in K-means Algorithm

The flowchart of above steps is shown in Figure 3 below:

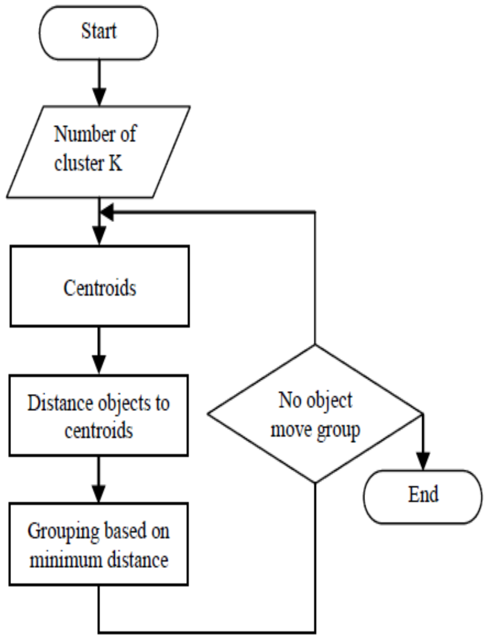

Figure 3: Flow chart of the K-means algorithm

Finally to identify the unknown speaker, we use Euclidean distance. The Euclidean distance measure the alteration distance of given two vector sets. It chooses one of the speakers with the smallest alteration distance for identification as the unknown person.

To achieve the desire result, first of all one has to divide the project in two parts: Training and Testing. In Training session the feature vectors are extracted from the training set and then the classifier is trained using these features. At testing session the classifier is tested with some unknown data. The following figure 4 illustrates the Speaker Recognition System concept.

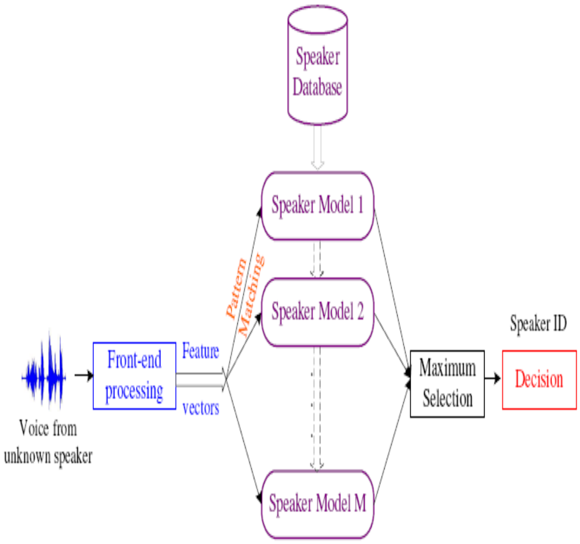

Figure 4: Block Diagram of the algorithm for Speaker recognition

In order to make a speaker identification system, main part is to find right feature.

IV. FRAMEWORK FOR VOICE RECOGNITION

The basic framework for voice recognition system consists of two phases – training phase and testing phase.
In the training phase, voice samples of different person are trained & stored in a database for creating reference model of voices. In the testing phase, actual decision for voice recognition is performed with the input voice sample. Decision is depend upon feature matching values of reference model of database and input voice.
The flowchart for training phase is shown in figure 5 & the flowchart for testing phase is shown in figure 6 below.

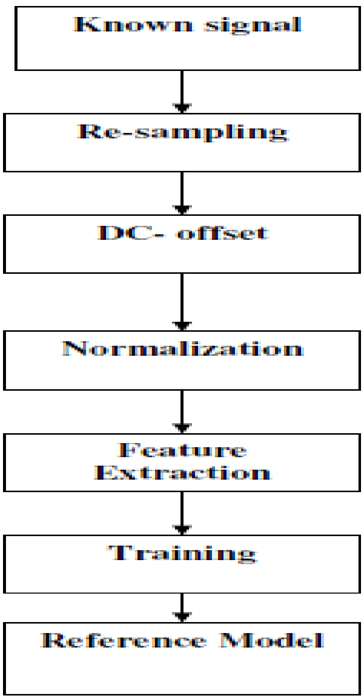

Figure 5: Testing Phase

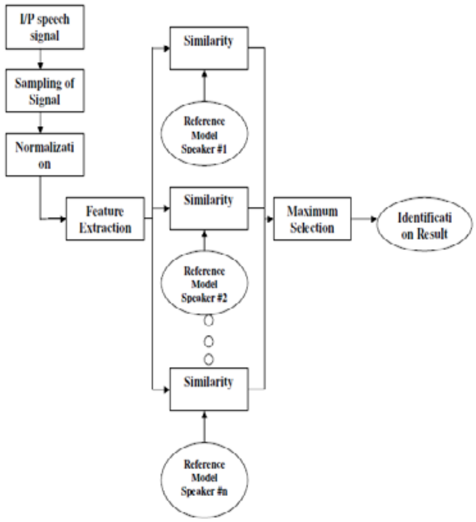

Figure 6: Testing Phase

First of all a set of speech signal is collected. Then they are re-sampled at 22050 samples per second with 16 bit per sample. DC offset of signal is corrected to 0. Then signal is normalized to 60%. The Mel-frequency cepstral coefficient of each signal is calculated and stored in database. This is the training phase. After that the unknown signal is taken for recognition. Then they are re-sampled at 22050 samples per second with 16 bit per sample. DC offset of signal is corrected to 0. Then signal is normalized to 60%. Then Mel frequency for each signal is calculated. Then Euclidean distance of the unknown signal is calculated. Minimum distance found is the result sound.
MATLAB [12] has been used to implement the above procedure. First training sample belonging to each class is passed to the function speaker_training.m, which reads each samples and calculates MFCC. The values of these MFCC's are then written into a database file called cepstrum.mat along with the name of training signal called name.mat. After that the unknown signal is passed through speaker_test.m which reads the signals and calculates MFCC. After that the cepstral coefficients of training signal and test signal are passed through distmeasure.m, which calculates the Euclidean distance between test signal and each of the training signals.

Figure 7 shows the general framework for voice recognition system.

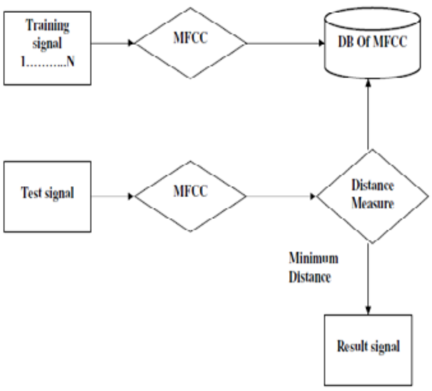

Figure 7: A framework for voice recognition system

V. ANALYSIS OF RESULT

In this work we compare voice of any unknown person with database of known voices, It then returns the message whether voice matches or not. In this paper, we use hybrid approach of "Mel Frequency Cepstral Coefficients (MFCC) technique, Vector Quantization, K-means algorithm and Euclidean distance."

MFCC is used to extract unique features from voice entered by the user. For estimation of probability distributions of the computed feature vectors, we use the concept Vector Quantization (VQ). VQ represents the centroid of extracted features. The k-means algorithm is used for dividing the features into n clusters; each cluster represents the unique characteristics of any feature of voice.

Finally to identify the unknown speaker, we use Euclidean distance. The Euclidean distance measure the alteration distance of given two vector sets. "It chooses one of the speakers with the smallest alteration distance for identification as the unknown person".

The figure 8 & 9 below shows the results obtained using the concept of feature extraction and feature mapping.

Command Window

Enter the directory for testing:test

Feature Matrix:

Columns 1 through 12

0	26.8854	26.9963	21.4765	16.7062	12.4984	26.3234	13.3399	16.2118	13.8117	18.3464	12.3106
18.4185	0	4.2721	22.4665	23.9453	14.9113	25.0625	19.5206	12.4781	25.8350	31.2962	23.5895
22.7681	9.8823	0	26.4881	25.0386	15.3603	23.6473	21.1390	13.8876	26.2100	29.2207	25.1517
10.1165	14.9667	19.4054	0	11.8591	8.3775	21.4304	10.3151	8.9860	21.4642	22.6087	11.5230
7.9922	14.9012	15.0420	11.7500	0	10.1251	22.6342	5.1456	9.9085	17.7243	22.5041	11.9257
9.6047	16.1677	13.9679	11.0052	10.0798	0	9.9649	6.7598	8.5171	9.7186	10.6351	9.7608
25.1049	30.0251	29.9015	27.4379	26.4522	16.2824	0	18.7583	15.2248	16.5175	13.2559	19.7756
10.1564	15.8545	13.2312	12.8584	7.9190	6.4191	11.8606	0	8.2300	10.8886	11.4654	10.7635
9.4231	14.4059	10.5331	9.0614	10.1524	5.8537	7.0111	6.9200	0	0.0934	7.9911	7.3692
36.7353	74.1376	45.6947	48.9109	44.1292	24.6580	26.8548	32.5735	19.4591	0	11.5633	23.0805
21.8060	58.3440	37.9056	36.4784	30.1825	19.7394	22.8006	21.8841	21.3969	8.7878	0	22.4543
12.9739	35.6208	24.0877	25.5952	17.5331	14.5675	19.9105	14.4860	10.2227	10.9644	14.9847	0
34.8265	71.2120	43.3976	49.3722	42.0313	25.9093	25.1780	31.9511	18.4316	6.5671	11.5271	20.9118
7.2266	35.9396	30.3324	25.4932	16.6892	14.7805	20.7728	12.1012	14.7256	9.7465	9.1356	12.6089
7.6003	21.0605	24.3346	15.5209	14.5728	14.0622	29.3056	14.5212	16.9753	19.3721	23.6665	11.1078
9.0582	25.3934	25.7145	20.5125	17.6420	12.9102	23.8953	13.0407	12.4803	11.3641	16.3998	7.4566
7.7618	36.8456	30.2027	25.4187	19.0078	15.9285	25.0555	14.4934	15.9272	11.6277	12.5679	15.9047
8.2592	39.0577	30.0444	25.9354	22.4712	14.8580	19.5338	14.6336	13.4155	9.0278	11.0664	11.0496
7.6669	30.1899	28.2718	18.5674	18.3417	14.7841	22.9204	14.1688	15.0734	10.9723	15.0789	10.8697
17.0606	10.7318	9.3687	25.5775	25.3127	13.6424	20.1224	20.7099	11.0075	22.3571	24.6903	22.3282

Figure 8: Result of feature comparisons

Command Window

Columns 13 through 20

17.6449	11.5052	7.9186	8.8238	7.7528	7.7957	7.5717	25.1806
31.5427	33.3587	22.3236	15.0649	21.8675	26.7587	19.6055	7.4460
29.8567	32.6589	23.9686	19.5910	25.1952	28.4966	23.9221	9.8823
23.9171	18.9674	7.0376	11.8295	11.7094	16.0774	11.7297	24.7770
19.4444	15.2081	10.1904	7.5244	11.2598	13.6034	8.8877	27.1709
10.0076	12.4314	9.5659	8.9121	9.6911	9.0425	9.2514	14.9503
14.7121	22.1671	21.9467	18.7863	21.6940	17.9955	19.3707	20.4589
12.6022	10.9383	9.6700	8.9486	10.2304	9.8090	8.5904	17.5894
9.9232	11.4892	11.2455	7.1471	7.6147	10.2022	6.9979	13.1814
6.3243	27.0662	38.4249	23.7905	33.8053	24.2978	29.5845	33.4289
8.1477	14.4057	25.5570	21.3310	18.5223	14.9490	19.5214	33.3441
11.4138	11.9980	14.7481	7.1316	12.9687	9.7332	10.8075	23.3473
0	24.2136	38.4883	23.4398	30.1316	21.1528	27.3862	31.7755
11.5921	0	10.0132	10.7478	6.8261	6.8784	9.1044	28.7623
25.0492	15.1063	0	7.6819	11.2575	13.2008	7.1864	27.2665
15.8443	11.4858	9.3230	0	8.1779	7.6545	6.2533	24.2274
13.0058	8.4639	12.7503	12.8442	0	6.1298	5.9202	28.4597
9.5273	9.6107	11.3059	8.7415	6.0195	0	5.2777	25.2325
15.3627	9.8609	7.5477	8.2840	5.6514	6.1630	0	26.1645
26.2960	30.2341	20.8675	17.1306	21.0601	23.3527	19.4485	0

Accuracy Percentage:

99.8502

Figure 9: Result of feature comparisons (cont…)

As we can clearly infer all the speakers 1 to 11 (in our example) were correctly identified. The final matrix contains the Euclidean distance from training to testing feature vectors. The value 0 in each row represents exact matching of voice. Hence accuracy of our algorithm achieves 99.85%.

VI. CONCLUSION

The process of recognizing a voice of a given speech from the group of given speakers is called speaker identification. This paper proposes a new improved and efficient technique for voice recognition system by applying Mel Frequency Cepstral Coefficients (MFCC), Vector Quantization, K-means and Euclidean distance technique. MFCC is used to extract unique features from voice entered by the user. For estimation of probability distributions of the computed feature vectors, we use the concept Vector Quantization (VQ). VQ represents the centroid of extracted features. The k-means algorithm is used for dividing the features into n clusters; each cluster represents the unique characteristics of any feature of voice. Finally to identify the unknown speaker, we use Euclidean distance.

REFERENCES

[1] Parwinder Pal Singh, Er. Bhupinder Singh, "Speech Recognition as Emerging Revolutionary Technology", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012.

[2] Campbell, J.P., "Speaker recognition: a tutorial", Proceedings of the IEEE Volume 85, Issue 9, Sept. 1997 Page(s):1437 – 1462.

[3] Surbhi Mathur, Choudhary S. K. and Vyas J. M., "Speaker Recognition System and its Forensic Implications", Open Access Scientific Reports, Volume 2, Issue 4, 2013.

[4] Douglas A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", ©2002 IEEE

[5] Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur, "The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review",IJARCSSE, Volume 2, Issue 2, February 2012.

[6] Genevieve I. Sapijaszko, Wasfy B. Mikhael, "An Overview of Recent Window Based Feature Extraction Algorithms for Speaker Recognition", IEEE, pp 880-883, 2012.

[7] Ufuk Ülüg, Tolga Esat Özkurt, Tayfun Akgül, "Bispectrum Mel-frequency Cepstrum Coefficients for Robust Speaker Identification", 2006 IEEE

[8] Maider Zamalloa, Germacn Bordel, Luis Javier Rodriguez, Mikel Penagarikano, "Feature Selection Based on Genetic Algorithms for Speaker Recognition", 2006, IEEE.

[9] HUO ChunBao SHAO Yan, "The improved VQ algorithm for speaker recognition",© 2009 IEEE

[10] Jorge Martinez, Hector Perez, Enrique Escamilla, Masahisa Mabo Suzuki, "Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques, IEEE, pp 248-251, 2012.

[11] Izuan Hafez Ninggal & Abdul Manan Ahmad, "The Fundamental Of Feature Extraction In Speaker Recognition: A Review", Proceedings of the Postgraduate Annual Research Seminar 2006.

[12] MATLAB Primer, MathWorks Inc, 2014.