Speaker Identification using Row Mean Vector of Spectrogram

H B Kekre
Computer Engineering Department,
Senior Professor,
MPSTME, SVKM's NMIMS University
Mumbai, 400-056, India.
+91-9323557897
hbk@yahoo.com

A Athawale
Computer Engineering Department,
Asst. Professor,
Thadomal Shahani Engg. College, Bandra (W),
Mumbai, 400-050, India.
+91-9226977842
Athawalearchana@gmail.com

M Desai
Computer Engineering Department,
Lecturer,
K. J. Somaiya Institute of Engg. And Information Technology, Sion (W),
Mumbai, 400-022, India.
+91-9869918310
desaimrunali@rediffmail.com

ABSTRACT
In this paper a simple approach to text dependent speaker identification using spectrograms and row mean is presented. This, mainly, revolves around trapping the complex patterns of variation in frequency and amplitude with time while an individual utters a given word through equalized spectrogram. These equalized spectrograms are used as a database to successfully identify the unknown individual from his/her voice. The features used for identifying, rely on optimal spectrogram segmentation and the Euclidean distance of the distributional features of the spectrograms of the unknown voice with that of a given known speaker in the database. Performance of this novel approach on a sample collected as two separate databases from 12 speakers and 28 speakers show that this methodology can be effectively used to produce a desirable success rate.

Categories and Subject Descriptors
I.5.4 [Pattern Recognition] : Applications – Signal Processing, Waveform analysis.

General Terms
Experimentation, Performance, Verification.

Keywords
Speaker Identification, Speaker Recognition, Spectrograms, Row Mean.

1. INTRODUCTION
Nowadays Information Technology based applications are growing rapidly and with the increase in these applications there is need for secure transaction. There are many areas where we need secure transaction as banking by phone, remote access of terminal, biometric identification etc. Speaker identification is very chief and robust technology as compare to other biometric techniques to secure access of many applications.

ICWET'11, February 25–26, 2011, Mumbai, Maharashtra, India.

As human beings, we are able to recognize someone just by hearing him or her talk. Usually, a few seconds of speech are sufficient to identify a familiar voice. The idea to teach computers how to recognize humans by the sound of their voices is quite evident, as there are several fruitful applications of this task. Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices.
The problem of speaker recognition can be divided into two major sub problems: *speaker identification* and *speaker verification*. Speaker identification can be thought of, as the task of determining who is talking from a set of known voices of speakers. It is the process of determining who has provided a given utterance based on the information contained in speech waves. The unknown voice comes from a fixed set of known speakers, thus the task is referred to as closed set identification. Speaker Verification on the other hand is the process of accepting or rejecting the speaker claiming to be the actual one. Since it is assumed that imposters (those who fake as valid users) are not known to the system, this is referred to as the open set task. Adding anone of the above option to the closed set identification task would enable merging of the two tasks, and it is called open set identification. Error that can occur in speaker identification is the false identification of speaker and the errors in speaker verification can be classified into the following two categories: (1) false rejections: a true speaker is rejected as an imposter, and (2) False acceptances: a false speaker is accepted as a true one [3].

2. RELATED WORK
All speaker recognition systems at the highest level contain two modules, feature extraction and feature matching. Feature extraction is the process of extracting unique information from voice data that can later be used to identify the speaker. Feature matching is the actual procedures of identifying the speaker by comparing the extracted voice data with a database of known speakers and based on this a suitable decision is made.

Mel-frequency Cepstral Coefficients (MFCC), based on short-time spectral analysis, are commonly used feature vectors for speaker identification [2]. VQ (Vector Quantization) technique is widely used in text-dependent and text-independent speaker recognition systems. In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ algorithm based on a training sequence to generate codebook [7]. The Gaussian mixture model (GMM) is a density estimator and is one of the most commonly used types of classifier[1].

3. PROPOSED APPROACH

In the proposed method, speaker identification is carried out by means of speech spectrograms [6]. Speaker Identification task includes the basic components: (I) feature extraction (II) speaker modelling (III) speaker matching and (IV) decision logic. The feature extraction module converts the raw speech waveform in the given sample to a spectrogram. Distributional features of the spectrograms are then used to make representative codebooks of speaker's voice patterns and use them to create a database.

Spectrograms were created using Short Time Fourier Transform method as discussed below:

In the approach using STFT, digitally sampled data are divided into chunks of specific size say 128, 256, 1024 etc. Fourier transform is then applied to calculate the magnitude of the frequency spectrum for each chunk. Each chunk then corresponds to a vertical line in the image, which is a measurement of magnitude versus frequency for a specific moment in time. Thus the speech database is converted into image database [5]. The spectrogram is cut out from the middle. Thus, the half spectrogram is obtained and then equalized image of half spectrogram is obtained.

For feature extraction the row mean is calculated from the half equalized spectrogram. The row mean vector is the set of averages of the intensity values of the respective rows [4]. If fig.1 is representing the sample image with 4 rows and 4 columns, the row and column mean vectors for this image will be as given below.

Row Mean Vector = [Avg(Row 1), Avg(Row 2), …., Avg(Row n)] (1)

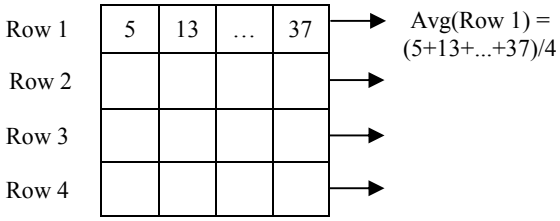

Figure 1. Row Mean Calculation.

Average of row mean vector is calculated and feature vector (x) of size 8, 16 and 32 is obtained. The same process is applied for test image and feature vector (y) of size 8, 16 and 32 is extracted. The Euclidean distance (D) between the database image and test image is calculated using the following formula 2.

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

The database sample with smallest Euclidean distance is declared as the identified speaker from the set of n speakers.

4. EXPERIMENTS AND RESULTS

Implementation for the proposed approach was done on Intel Core i3 Processor, 2.26 GHz, and 3 GB of RAM. Windows 7 Operating System and MATLAB 7.0 are used. For recording purpose Windows Sound Recorder is used. The implementation process is started by collecting the speech samples of various speakers. That will become a database for the project. For recording three sentences are selected as follows.

S1 The central processing unit is the brain of the computer.
S2 FFT is used for implementation of DFT.
S3 An operating system is an interface between user and hardware.

Two separate databases are created. In one database audio format is kept same for all the speakers that is PCM 22.050 kHz, 16 bit Mono. There are 12 speakers in first database and each will speak each sentence 6 times. So there are total 72 occurrences for one sentence, from these 24 samples are used for testing and remaining 48 will be treated as database samples. In second database different PCM rates are chosen. There are 28 speakers and each sentence is repeated for 3 times only. Thus, there are 84 samples for each sentence. From 3 occurrences 1 occurrence is used for testing and other 2 are used as database samples. Recording was done at varying times.

From these speech samples spectrograms were created with window size 1024 and no overlap. As spectrograms are mirrored image it is cut down from the middle and half spectrogram is used for analysis. The histogram equalized image of this half spectrogram is stored as database sample. This forms the closed set for experiment. The Half spectrogram is then used to calculate row mean vector. From this row mean vector, feature vectors of size 8, 16 and 32 are obtained by sectionalizing row vector. Then Euclidean distance is used for similarity calculation. The same process is applied on the test sample and obtained the test image. Euclidean distance between feature vectors of test images and database images was used as a measure of similarity between images. In this approach no pre-processing or normalization is done.

Figure 4 and Figure 5 shows the graph of original speech sample of sentence S2 spoken by speaker 1 and speaker 2 respectively. In these graphs, x-axis indicates number of samples and y axis indicates the value at the time instant.

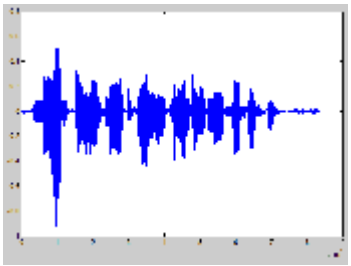

Figure 2. Original speech sample of sentence S2 spoken by speaker 1.

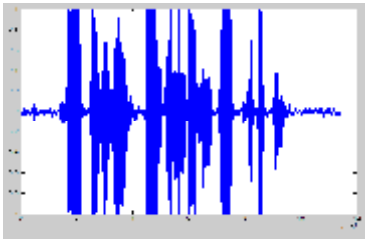

Figure 3. Original speech sample of sentence S2 spoken by speaker 2.

Figure 4 shows that the image obtained for the sentence S2 spoken by two different speakers is different. It can be seen that the spectrogram of the same sentence spoken by two different speakers are also different. Since the size of each spectrogram is different they can not compared directly. Hence the row mean vector as the representative of each spectrogram has been taken. The size of feature vector is further reduced by sectionalizing this row mean vector.

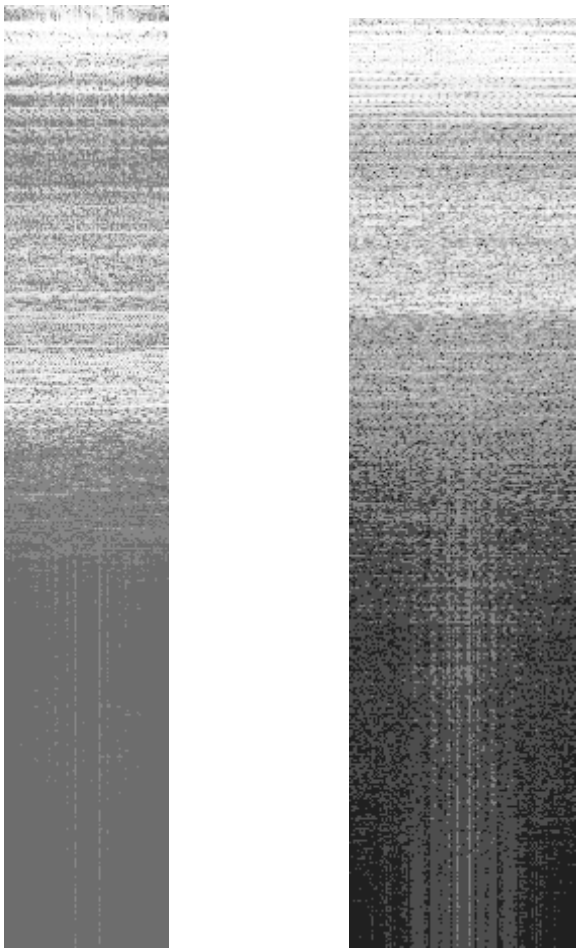

Figure 4. Spectrogram of Sentence S2 uttered by Speaker 1 and speaker 2 respectively.

(A) Results

Table 1 and Table 2 shows the identification rate of sentences S1, S2 and S3 when feature vector size is 8, 16, and 32 for database 1 and database 2 respectively. It can be seen from the table that the results with database 1 are better than results with database 2. And also with feature vector of size 32.

Table 1. Identification rate of three sentences S1, S2, and S3 when feature vector size is 8, 16 and 32 for database 1.

Sentence	Feature Vector Size		
	8	16	32
S1	75%	87.5%	**91.67%**
S2	87%	91.30%	**95.65%**
S3	91.30%	91.30%	**100%**

Table 2. Identification rate of three sentences S1, S2, and S3 when feature vector size is 8, 16 and 32 for database 2.

Sentence	Feature Vector		
	8	16	32
S1	67.85 %	**71.42 %**	**71.42 %**
S2	50 %	53.57 %	**57.14 %**
S3	60.71 %	**67.86 %**	**67.86 %**

Figure 5 and figure 6 shows the accuracy rate versus No of features for database 1 and database 2 respectively.

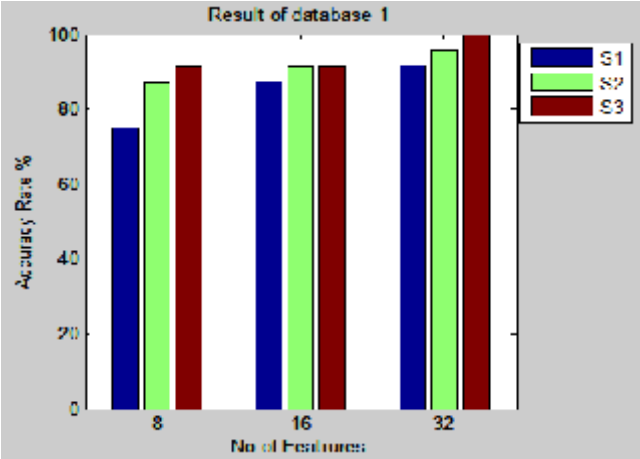

Figure 5. Accuracy rate versus nunber of features for S1, S2 and S3 Sentences database 1.

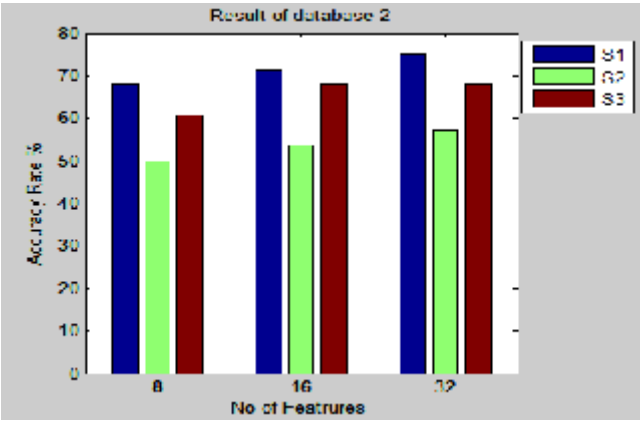

Figure 6. Accuracy rate versus no of features for S1, S2 and S3 Sentences database 2.

5. CONCLUSION

In this paper, text-dependent speaker identification using spectrogram and row mean is proposed. Spectrograms are used for obtaining feature vectors. Row mean is applied on spectrograms to obtain feature vector of size 8, 16 and 32. Then, Euclidean distance is calculated to identify the speaker.

The experiments performed with two databases and it can be seen from the results that results using database 1 are better than results using database 2 mainly because second data base contains samples with different sampling rates with less number of samples per speaker. Results using feature vectors of size 32 are better than results using feature vectors of size 8 and16. As accuracy close to 100% is achieved for database 1 for 32 sections higher no of sections will not be cost effective.

6. REFERENCES

[1] Abdul Manan Ahmad, Loh Mun Yee "Vector Quantization Decision Function for Gaussian Mixture Model Based Speaker Identification", 2008 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008) Swissôtel Le Concorde,Bangkok,Thailand

[2] Ali Zulfiqar, A. Muhammad, A. Enriquez, A.M., "A Speaker Identification System using MFCC Features with VQ Technique", 2009 Third International Symposium on Intelligent Information Technology Application, Vol 3, pp 115-118.

[3] Bojan Imperl, "Speaker recognition techniques", Laboratory for Digital Signal Processing, Faculty of Electrical Engineering and Comp. Sci., Smetanova 17, 2000 Maribor, Slovenia.

[4] Dr. H. B. Kekre, S D Thepade, A Athawale, A Shah, P Verlekar, S Shirke, "Image Retrieval using DCT on Row Mean, Column Mean and Both with Image Fragmentation", International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) – TCET, Mumbai, India, February 26-27, 2010.

[5] Dr. H. B. Kekre, Dr. Tanuja K. Sarode, Shachi J. Natu, Prachi J. Natu, "Speaker Identification Using 2-D DCT, Walsh And Haar On Full And Block Spectrogram", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1733-1740.

[6] [Tridibesh Dutta, "Text dependent speaker identification based on spectrograms", Proceedings of Image and vision computing, pp. 238-243, New Zealand 2007.

[7] Y. Linde, A. Buzo, R. M. Gray, "An algorithm for Vector Quantizer Design", IEEE Transaction on Communications, 28: 1980, pp 84-95.