Speaker Identification using Spectrograms of Varying Frame Sizes

H. B. Kekre
Phd,Senior Professor,
Computer Dept., MPSTME,
NMIMS University
Mumbai, 400-056, India.

Vaishali Kulkarni
PhD Scholar, Associate Professor,
EXTC Dept., MPSTME,
NMIMS University.
Mumbai, 400-056, India.

Prashant Gaikar, Nishant Gupta
Student (B.Tech)
EXTC Dept., MPSTME,
NMIMS University.
Mumbai, 400-056, India.

ABSTRACT
In this paper, a text dependent speaker recognition algorithm based on spectrogram is proposed. The spectrograms have been generated using Discrete Fourier Transform for varying frame sizes with 25% and 50% overlap between speech frames. Feature vector extraction has been done by using the row mean vector of the spectrograms. For feature matching, two distance measures, namely Euclidean distance and Manhattan distance have been used. The results have been computed using two databases: a locally created database and CSLU speaker recognition database. The maximum accuracy is 92.52% for an overlap of 50% between speech frames with Manhattan distance as similarity measure.

General Terms
Speaker Identification, Spectrograms

Keywords
Discrete Fourier Transform (DFT), Row Mean, Euclidean Distance, Manhattan distance

1. INTRODUCTION
The goal of automatic speaker recognition systems is to extract, characterize and recognize the information in the speech signal for conveying speaker identity [1]. Speaker recognition is divided into two areas: speaker identification and speaker verification. Speaker identification is deciding if a speaker is a specific person or is among a group of persons [1]. Speaker verification is deciding if a speaker is who he/she claims to be [2]. Speaker verification is a 1:1 match where one speaker's voice is matched to one template whereas speaker identification is a 1: N match where the voice is compared against N templates. Algorithms developed for speaker recognition depend on whether the system being analyzed is based on text dependent or text independent speech samples. In text dependent recognition, the phrase is known to the system and can be fixed or prompted. In text independent recognition the system must be able to recognize the speaker from any text [1, 4, and 5].

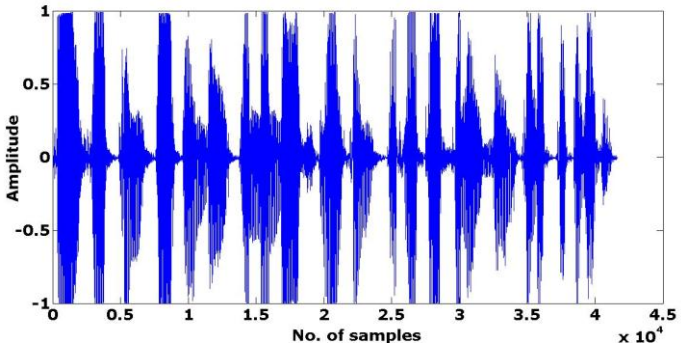
A spectrogram describes how the spectral density of a signal varies with time. The most commonly used form of a speech spectrogram is the frequency versus time plot with a third dimension indicating the amplitude of a particular frequency at a particular time represented by the intensity or color of each point in the image. The concept of using spectrograms for speaker identification has been around for decades. One of the first attempts for automatic speaker recognition were made in the 1960s [3]; by using filter banks and correlating two digital spectrograms for a similarity measure [6]. Results of experiments related to speaker identification by speech spectrograms have been compared and discussed as early as 1969[7]. Several techniques have been developed for pattern matching in speech signals using spectrograms. Technique in [8] describes the spectrogram band as a cluster and its mean pixel value, the centroid of cluster. Hence, given an unknown speaker's utterance of a known word, we would be looking for the database sample of that particular word with ordered cluster centroids having the closest Euclidean distance with those of the unknown speaker. Comparison of transformations such as the DCT, Haar and Walsh on the spectrograms has been discussed in [9]. Using row mean on Kekre's transform of spectrogram images of different frame sizes for speaker identification [10] and applying 2D DCT on full/block spectrogram and 1D DCT on row mean of spectrogram [11] have shown favorable results.

The paper is organized as follows; Section 2 describes the spectrogram generation technique, Section 3 explains the feature vector extraction process, in section 4 the feature matching has been explained, followed by the decision making in Section 5, in Section 6 a brief description of the database is given. Section 7 discusses the results and conclusion is given in Section 8.

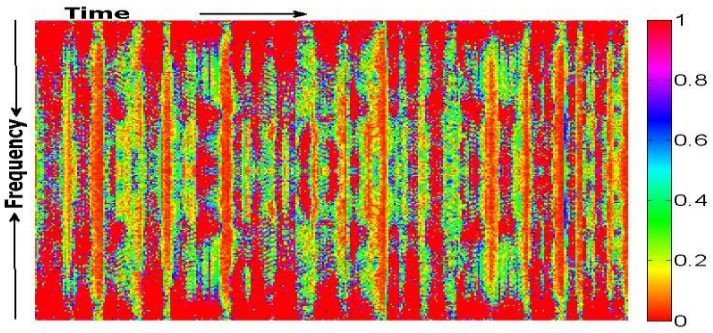
2. SPECTROGRAM GENERATION
For the present work on Speaker Identification, spectrograms have been generated using the following steps [12, 13]:

1. The speech signal has been first divided into frames, of sizes from 32 to 512 with step size of 32 with an overlap of 25% or 50%.
2. These frames have been arranged column wise to form a matrix. E.g. if the speech signal is a one dimensional signal of 44096×1. This is first divided into frames of 256 samples each with an overlap of 25% between consecutive frames i.e. overlap of 64. These 229 frames are then arranged column wise to form a matrix of dimension 256×229.
3. Discrete Fourier Transform (DFT) has then been applied to this matrix column wise.
4. The spectrogram has then been plotted as the absolute magnitude of this transform matrix.

In Fig. 1 (a) a speech signal of one of the speaker in the database is shown. Fig. 1 (b) shows the spectrogram generated for this speech signal using pseudo colors. There is a lot of information in the temporal-spectral dynamics contained in the complete speech signal that can help speaker-identity. Fig. 2 shows the spectrograms of two different speakers for two different iterations of the same sentence. Fig. 2 (a) shows the spectrogram of speaker 1 for iteration 1. Fig. 2 (b) shows the spectrogram of speaker 2 for iteration 1. As can be seen there is lot of difference between these two spectrograms. Fig. 2 (c) shows the spectrogram of speaker 1 for iteration 2. There is a similarity between Fig. 2 (a) and Fig. 2 (c). Fig. 2 (d) shows the spectrogram of speaker 2 for iteration 2. There is a

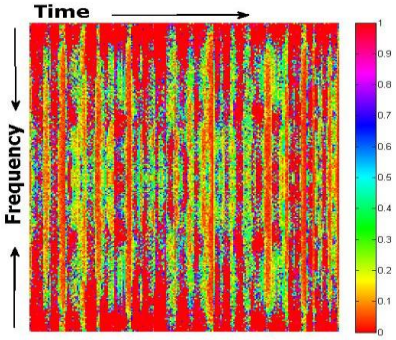


(a)

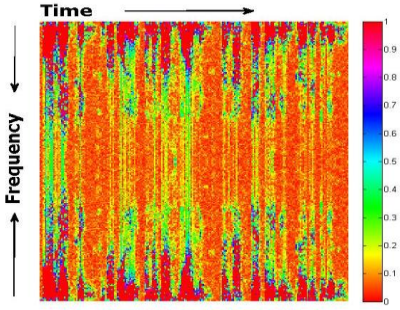


(b)

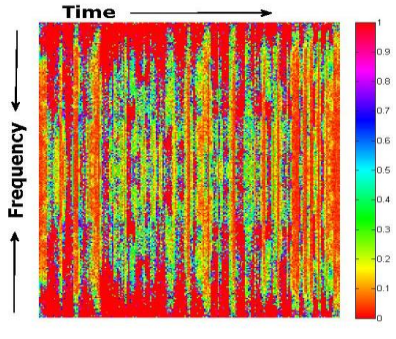
Fig. 1 Speech and its spectrogram



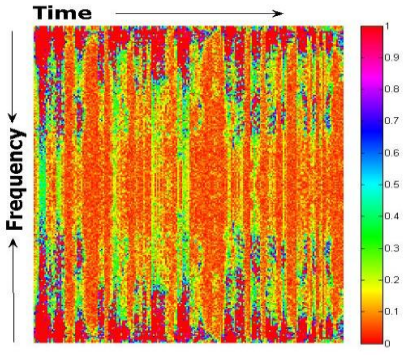
(a) Speaker 1, iteration 1



(b)Speaker2,iteration1



(c) Speaker 1, iteration 2



(d) Speaker 2, iteration 2

Fig. 2 Spectrograms of the same sentence for two different speakers for frame size of 256 with 25% overlap. a) Speaker 1, iteration 1 b) Speaker 2, iteration 1 c) Speaker 1, iteration 2 d) Speaker 2, iteration 2.

similarity between Fig. 2 (b) and Fig. 2 (d). Thus as can be seen visually, the spectrogram has the temporal-spectral information which can be used to identify a speaker.

3. FEATURE VECTOR EXTRACTION

The procedure for feature vector extraction is given below:

- The spectrograms of all the speech waveforms have been generated for the different frame sizes as described in section 2.
- The mean of the absolute values of the rows of the spectrogram matrix is then calculated.
- These row means form a column vector (M×1), where M is the number of rows in the spectrogram matrix.
- This column vector forms the feature vector for the speech sample.

The feature vectors for all the speech samples have been calculated for different values of n (frame size) and stored in the database.

4. FEATURE MATCHING

In the work proposed in this paper two distance measures, Manhattan Distance (MD) and Euclidean Distance (ED) have been explored and comparative performance of both has been given. Manhattan distance (MD) [15] is defined as the Minkowiski distance of the order 1 or 1-norm distance (where p=1). The 1-norm distance is called the taxicab norm or Manhattan distance, because it is the distance a car would drive in a city laid out in square blocks (if there are no one-way streets). In n dimensions, the MD between two points A and B is given by eq. (1), where xi (or yi) is the coordinate of A (or B) in dimension i.

$$d_{AB} = \sum_{i=1}^{n} | x_i - y_i | \quad \textbf{(1)}$$

Euclidean distance is defined as the Minkowiski distance of the order 2 or 2-norm distance (where p=2). Euclidean Distance (ED) [14, 15] is defined as the straight line distance between two points. It is what would be obtained if the distance between two points were measured with a ruler: the "intuitive" idea of distance. In n dimensions, the ED between two points A and B is given by eq. (2), where xi (or yi) is the coordinate of A (or B) in dimension i.

$$d_{AB} = \left(\sum_{i=1}^{n} | x_i - y_i |^2 \right)^{1/2} \quad (2)$$

5. DECISION MAKING

The final step in speaker recognition process is the decision making. The feature extraction and pattern matching are same for different speaker recognition tasks, but the decision depends on the task: closed set or open set. Let us denote generally a speaker model of speaker i by Si, and let S = {S1,…….,SN} be the speaker database of N known speakers. Without assuming a specific speaker model/classifier, let score(X, Si) be the match score between the unknown speaker's feature vectors X = {x1,…….,xT} and the speaker model Si. In the case of distance based classifiers, minimum distance corresponds to best match. In closed-set speaker identification task, the decision is simply the speaker index i that yields the minimum distance, where i is given by eq. (3).

$$\mathbf{i = min_i\ dist(X,\ Si)} \quad (3)$$

where the minimum is taken over the speaker database S. In the open set identification task, the decision is given as given by eq. (4).

$$dist(X, S_i) \begin{cases} < \Theta_i, accept \\ \geq \Theta_i, reject \end{cases} \quad (4)$$

Where Θi is the threshold. The threshold can be set the same for all speakers, or it can be speaker-dependent. The threshold is determined so that a desired balance between the two types of errors False Acceptance Rate (FAR) and False Rejection Rate (FRR) is achieved [5, 16].

FRR and FAR are given by eq. (5) and eq. (6) respectively.

$$\textbf{FRR = (true claims rejected/total true claims) × 100} \quad \textbf{(5)}$$

$$\textbf{FAR = (imposter claims accepted/total imposter claims) ×100} \quad \textbf{(6)}$$

$$\textbf{GAR = 100 − FRR} \quad \textbf{(7)}$$

GAR given by eq. (7) is defined as the **Genuine Acceptance Rate (GAR)**, in percentage.

Thus FAR is the error with which an imposter is accepted and FRR is the error with which a genuine or true speaker is rejected. There is a trade-off between the two errors. When the decision threshold Θ_i is increased: FAR increases but FRR decreases, and vice versa. The balance between these two depends on the application. Since either of the two types of errors can be reduced at the expense of an increase in the other, a measure of overall system performance must specify the levels of both types of errors. The trade-off between FAR and FRR is a function of the decision threshold. FAR and FRR are plotted against the decision threshold. The point of intersection of these two curves is defined as the Equal Error Rate (EER). The EER is the value for which the FAR and FRR are equal. The system performance can be given by Performance index (PI), which is defined as given by eq. (8).

$$\textbf{PI (\%) = 100-EER (\%)} \quad \textbf{(8)}$$

5. DATABASE DESCRIPTION

5.1 Locally Created Database

The speech samples used in this work are recorded using Sound Forge 4.5. The sampling frequency is 8000 Hz (8 bit, mono PCM samples). Table I shows the database description. The samples are collected from different speakers. Five iterations of four different sentences (E1, E2, E3 and E4) of varying lengths are recorded from each of the speakers. Twenty samples per speaker are taken. For text dependent identification, four iterations of a particular sentence are kept in the database and the remaining one iteration is used for testing. These speech signals have a amplitude range of '-1' to '+1'.

5.2 CSLU Voices for Speaker Recognition Version 1.1

The CSLU Speaker Recognition Database consists of telephonically recorded speech spanning twelve collected over a two year period. These speech signals have very low amplitude range. These signals are scaled up to the level '-1' to '+1'. Also preprocessing was done to remove the long silent parts in between the words.

Table1. Description of Local Database

Parameter	Sample characteristics
Language	English
No. of Speakers	107
Speech type	Read speech, microphone recorded
Recording conditions	Normal
Sampling frequency	8000 Hz
Resolution	8 bps

Table2. Description of CSLU Database

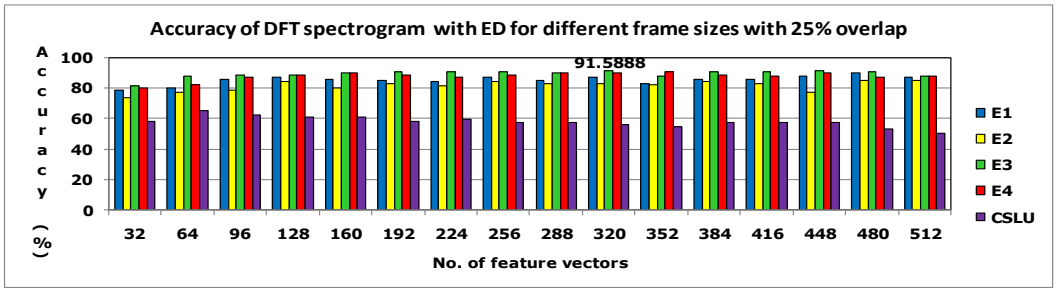
Parameter	Sample characteristics
Language	English
No. of Speakers	77
Speech type	Read speech, telephonically recorded
Recording conditions	Normal
Sampling frequency	8000 Hz
Resolution	16 bps

6. RESULTS AND DISCUSSION

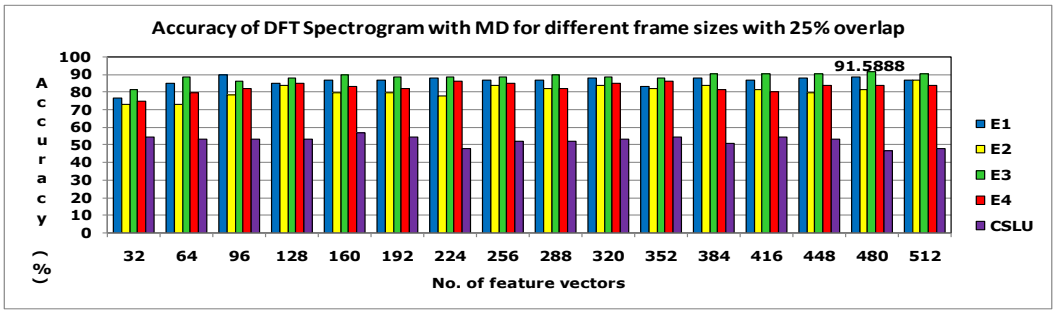
The experiments have been performed on spectrograms of different frame sizes with 25% and 50% overlap between consecutive frames. In the first set of experiment, spectrograms generated with 25% overlap have been considered. The results have been computed on the four sentences (E1, E2, E3 and E4) of the local database and one phrase of the CSLU database. The row mean which forms a column vector as described in section 3 forms the feature vector. For closed set identification, the feature vectors have been calculated for the reference speech samples and stored in the database. For testing, the test speech sample has been similarly processed and feature vector has been computed. The similarity measure Euclidean distance (ED) or Manhattan Distance (MD) between the database feature vectors and test feature vector has been calculated. The speaker whose reference feature vector gives the minimum distance with the test feature vector has been declared as the speaker recognized.

Fig. 3 shows the performance of DFT spectrogram for varying frames sizes with 25% overlap for the different sentences. Fig. 3 (a) shows the results with ED as similarity measure. It can be seen from the results that, accuracy increases as the feature vector size is increased up to a certain value for all the sentences. After that the accuracy decreases or remains at almost the same level. The maximum accuracy for E1 is 89.71% for a feature vector of size 480, for E2 it is 85.04% for a feature vector of size 480. E3 gives a maximum accuracy of 91.58% for a feature vector of size 320. E4 gives a maximum accuracy of 90.65% which is obtained for a feature vector of size 352. The CSLU database gives comparatively lower results with a maximum accuracy of 64.93% for a feature vector of size 64. Fig. 3 (b) shows the results with MD as similarity measure. The maximum accuracy for E1 is 89.71% for a feature vector of size 96, for E2 it is 86.91% for a feature vector of size 512. E3 gives a maximum accuracy of 91.58% for a feature vector of size 480. E4 gives a maximum accuracy of 85.98% which is obtained for a feature vector of size 224. The CSLU database gives comparatively lower results with a maximum accuracy of 57.14% for a feature vector of size 160. The comparison of the best performance of DFT spectrogram with 25% overlap for ED and MD are shown in Table 3.

In the work proposed in this paper, open set identification has been done on one sentence E4 from the local database, for which the speech samples from the imposter speakers have been collected. There are 31 imposter speakers. For the open set identification, False Rejection Rate (FRR) and False Acceptance Rate (FAR) have been calculated for E4 for the frame size of 352 and 224 with 25% overlap for ED and MD respectively by varying the threshold. Fig. 4 (a) shows the % rate for FAR and FRR with ED for varying threshold. The EER is 13.7% and the PI is 86.3%. Fig. 4 (b) shows the % rate for FAR and FRR with MD for varying threshold. The EER is 16.9% and the PI is 83.1%. In the problem of Speaker Identification, the parameter EER does not play any important role. However, the threshold value, which gives the margin of operation for 0% FAR, is important. Hence for comparing performance of DFT for threshold parameter, the ratio of maximum permissible threshold at 0% FAR and at crossover point (EER) of FAR and FRR has been considered. The conflict for same ratio is resolved by considering GAR at a point where FAR is 0%. Table 4 gives the comparative threshold performance of DFT spectrogram with 25% overlap for FAR and GAR with respect to both similarity measures ED and MD.

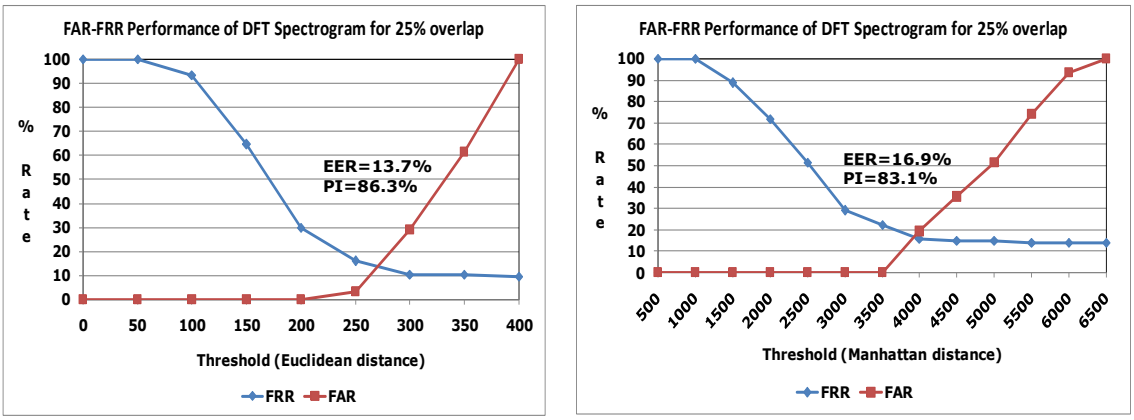


(a)



(b)

Fig. 3: Accuracy of DFT spectrogram with 25% overlap between frames. a) with ED and b) with MD



(a) **(b)**

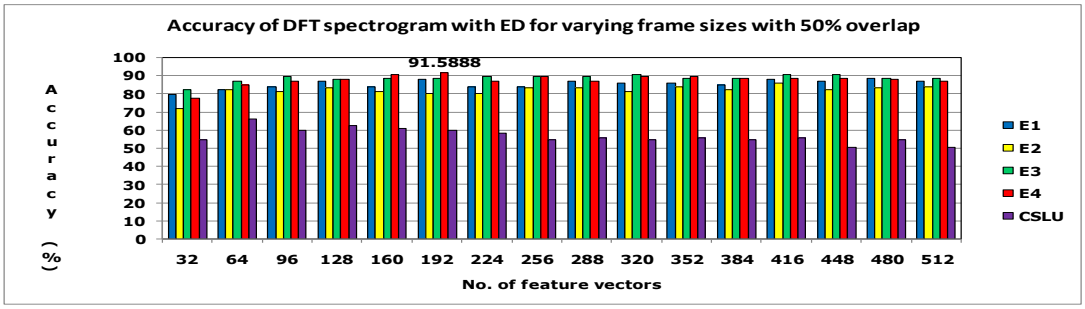
Fig. 4. FAR-FRR for DFT spectrogram for frame size of 352 and 224 with 25% overlap for ED and MD respectively for varying threshold.

Table 3 Best results for DFT spectrogram with 25% overlap

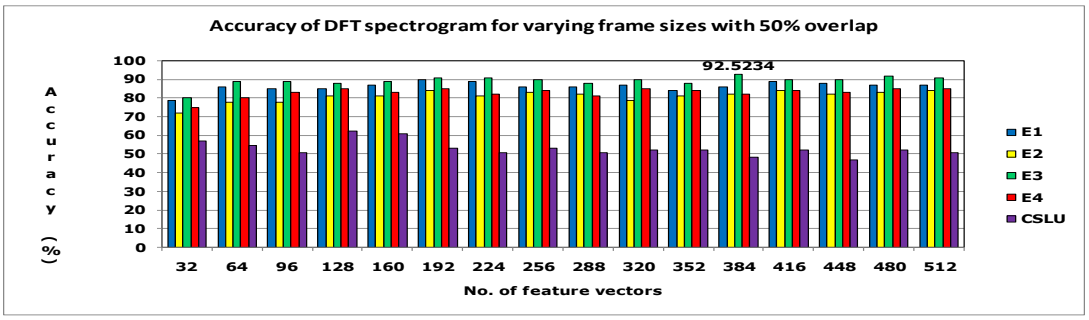
Sentence	Total samples tested	ED as similarity measure		MD as similarity measure	
		Feature vector	Accuracy (%)	Feature vector	Accuracy (%)
E1	107	480	89.71	96	89.71
E2	107	480	85.04	512	86.91
E3	107	320	91.58	480	91.58
E4	107	352	90.65	224	85.98
CSLU	77	64	64.93	160	57.14

Table 4 Comparison of Threshold of DFT spectrogram with 25% overlap for FAR and GAR with ED and MD.

Similarity measure	GAR for 0% FAR	Threshold		
		At EER	At 0% FAR	%
ED	70.09	270	200	74.07
MD	77.57	4000	3500	87.5

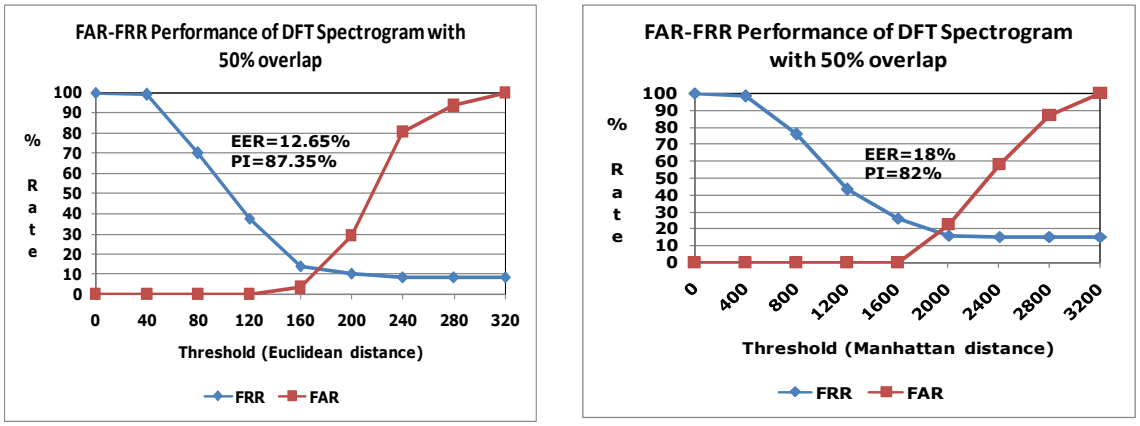


(a)



(b)

Fig. 5: Accuracy of DFT spectrogram with 50% overlap between frames. a) with ED and b) with MD



(a) **(b)**

Fig. 6: FAR-FRR for DFT spectrogram for frame size of 192 with 50% overlap for ED and MD for varying threshold.

Table 5 Best results for DFT spectrogram with 50% overlap

Sentence	Total samples tested	ED as similarity measure		MD as similarity measure	
		Feature vector	Accuracy (%)	Feature vector	Accuracy (%)
E1	107	480	88.78	192	89.71
E2	107	416	85.98	192	84.11
E3	107	416	90.65	384	92.52
E4	107	192	91.58	192	85.04
CSLU	77	64	66.23	128	62.33

Table 6 Comparison of Threshold of DFT spectrogram with 50% overlap for FAR and GAR with ED and MD.

Similarity measure	**GAR for 0% FAR**	**Threshold**		
		At EER	**At 0% FAR**	**%**
ED	62.61	175	120	68.57
MD	73.83	1900	1600	84.21

In the second set of experiment, spectrograms generated with 50% overlap have been considered. Fig. 5 shows the results obtained for the different frame sizes with 50% overlap. Fig. 5 (a) shows the results with ED as similarity measure. The best result is 91.58% which has been obtained for E4 for a feature vector of size 192. Fig. 5 (b) shows the results with MD as similarity measure. The best result is 92.52% which is obtained for E3 for a feature vector of size 384. The comparison of the best performance of DFT spectrogram for 50% overlap with ED and MD is shown in Table 5.

For the open set identification, FRR and FAR was calculated for E4 for the frame size of 192 with 50% overlap by varying the threshold. Fig. 6 (a) shows the % rate for FAR and FRR with ED as the threshold. The EER is 12.65% and the PI is 87.35%. Fig. 6 (b) shows the % rate for FAR and FRR with ED for varying threshold. The EER is 18% and the PI is 82%. Table 6 gives the comparative threshold performance of DFT spectrogram with 50% overlap for FAR and GAR with respect to both similarity measures ED and MD.

From these experiments it can be observed that:

- The accuracy increases as the feature vector size (frame size) is increased.
- It can be observed that for ED, 25% overlap gives maximum accuracy of 91.58% for a frame size of 320, whereas 50% overlap gives same accuracy for a frame size of 192 reducing the computational complexity by a factor of 1.66. For MD, the maximum accuracy with 25% overlap is 91.58% and it increases to 92.52% with 50% overlap.
- As far as FAR/FRR results are concerned, the performance is much better with 50% overlap. It can be seen that for 50% overlap, MD gives better performance with GAR of 73.83% for a threshold level of 83.33% of that at EER.

- Accuracy also depends on the nature and length of the sentences in the database. The results obtained for E3 and E4 which are longer than E1 and E2 are better.
- Instrumentation used for recording voice also plays an important role in deciding the accuracy of speaker identification. The results of CSLU database are poorer as compared to the local database because CSLU is telephonic recording whereas for the database microphone, which has larger bandwidth, has been used.

7. CONCLUSION

In this paper, we have proposed a technique for text-dependent speaker identification for a closed set as well as open set using row mean of spectrograms it can be observed that accuracy increases with the size of feature vector. Also 50% overlap gives comparable results with lesser computational complexity. Manhattan distance (MD) has edge over Euclidean distance (ED). The maximum accuracy is 92.52% with 50% overlap with MD as similarity. The study is ongoing and different techniques to extract the features from the spectrogram are being explored.

8. REFERENCES

[1] D.A. Reynolds, "An overview of automatic Speaker Recognition Technology", ICASSP 2002, pp 4072-4075.

[2] J. M. Naik, "Speaker Verification: A Tutorial", IEEE Communications Magazine, January 1990, pp.42-48.

[3] S. Furui, "Fifty years of progress in speech and speaker recognition," Proc. 148th ASA Meeting, 2004.

[4] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D.Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," EURASIP J. Appl. Signal Process., no. 1, pp. 430–451, 2004.

[5] Joseph P. Campbell, Jr., Senior Member, IEEE, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.

[6] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J.A.S.A.*, 35, pp. 354-358, 1963.

[7] R. H. Bolt, F. S. Cooper, E. E. David, Jr., P. B. Denes, J. M. Pickett and K. N. Stevens "Identification of a Speaker by Speech Spectrograms", Science Volume 166, pp. 338-343 (1969)

[8] T. Dutta, 'Text Dependent Speaker Identification Based on Spectrograms', Proceedings of Image and Vision Computing New Zealand 2007, pp. 238–243, Hamilton, New Zealand, December 2007.

[9] 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.

[10] Dr. H. B. Kekre, Vaishali Kulkarni, "Speaker Identification using row Mean of Haar and Kekre's Transform on Spectrograms of Different Frame Sizes", International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.

[11] H. B. Kekre, Tanuja Sarode, Shachi Natu, Prachi Natu, "Performance Comparison Of 2-D DCT On Full/Block Spectrogram And 1-D DCT On Row Mean Of Spectrogram For Speaker Identification", (Selected) CSC-International Journal of Biometrics and Bioinformatics (IJBB), Volume (4): Issue (3).

[12] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," IEEE Spectrum, vol. 7, pp. 57–62, August 1970.

[13] W. Koenig, H. K. Dunn, and L. Y. Lacey, "The sound spectrograph," Journal of the Acoustical Society of America, vol. 18, pp. 19–49, 1946.

[14] Paul E. Black, "Euclidean distance", Dictionary of Algorithms and Data Structures [online].

[15] Paul E. Black, ed., U.S. National Institute of Standards and Technology.17 December 2004. Available from: http://www.nist.gov/dads/HTML/euclidndstnc.html

[16] Micki Krause and Harold F. Tipton, "Handbook of Information Security Management", Auerbach Publications, CRC Press, ISBN: 0849399475.