Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Speaker Classification using DTW and A Proposed Fuzzy Classifier

Alia Karim Abdul-Hassan [a] , ***Iman Hassoon Hadi*** [b]

[a] *University of Technology/Iraq- Computer Sciences Dept., e-mail: hassanalia2000@yahoo.com*

[b] *University of Technology/Iraq- Computer Sciences Dept., e-mail :iman.h.1439@gmail.com*

ARTICLE INFO

Article history:
Received: 06 /08/2019
Rrevised form: 28 /08/2019
Accepted : 01/ 09 /2019
Available online: 29 /09/2019

Keywords:

MFCC, DTW, Fuzzy inner product, speaker classification

ABSTRACT

The speaker classification is considered with how the identity of the speaker is represented as a unique class label. This identity is characterized by the voice features belong to the speaker. The speaker classification has many application related to the security and forensic systems. There are many classification methods that could be used in speaker classification but the such classifier must has the ability to discriminate the between voice feature vectors which overthought there are a small differences. In this work, a proposed fuzzy classifier has been used for speaker classification using fuzzy inner product (FIP) and Mel frequency Cepstral coefficients (MFCC) features. This proposed classifier is evaluated by a comparison with Dynamic Time Warping (DTW) as traditional method. The proposed classifier was more accrued than DTW, since it classify speakers in ELSDSR data set with 90.91% while the accuracy of DTW classifier was 77.27%.

DOI : 10.29304/jqcm.2019.11.4.625

1. INTRODUCTION

One reason for designing and implementation of speaker classification is the process of identifing or verifing the identity of users for grant them a secure access to the information system[1]. Speaker classification can be thought of as speaker identification in which each class is a speaker. The speaker voice unique features can be utilized by applications to recognize and classify the authenticated and unauthenticated users [2].

In general, each speaker recognition system should contain speaker classification stage. So the speaker classifier is the process to create a model for each speaker in the training stage and use that model to authenticate the identity of each speaker in the test stage. There two major types of speaker classifiers : supervised and unsupervised.

In this paper , two types of speaker classifiers are used for comparison purpose , the first one is a proposed fuzzy classifier which is based on fuzzy set theory which is one of the soft computing as part of the artificial intelligent. The second classifier is based on Dynamic Time warping (DTW) method as crisp method. These two classifiers are reprsent the classification process in the speaker recognition system that include beside the classifier, the feature extraction of MFCC voice feature and the training and testing voice data. In this paper , the ELSDSR dataset is used to evaluate the recognition accuracy [3].

2. LITERITURE REVIEW

Togneri et al. (2011) Studied Speaker Classification with GMM-UBM and GMM-SVM, the implementation and evaluation were done under different experiments. they conclude that GMM-UBM and GMM-SVM

Corresponding author Iman Hassoon Hadi

Email addresses: iman.h.1439@gmail.com

Communicated by Qusuay Hatim Egaar

classification have same similar performance [4]. Arora (2016) In this paper, various strategies are discussed in speech recognition along conclude that CMN and GMM model used in last stages are more accurate even in noisy environment as compared with other techniques [5]. Gan et al. (2016) propose an evaluation of many classification algorithms for a speaker identity selection process. They presented a fusion engine for combining the scores from a number of classifiers, which uses the GMM-UBM approach to match speaker identity [6]. Swathy et al. (2017) they implemented a survey on classification methods related to speaker recognition; they concluded that GMM, ANN and a proposed fuzzy classifier are the important classification techniques [7]. Nayana et al. (2017) they use Gaussian Mixture Models (GMM) and i-Vector methods with two features PNCC and RASTA PLP coefficients. They concluded that accuracy is better with pitch and formants are added to basic features in speaker classification. In addition, the accuracy of i-vector with PLDA classifier is better than CDS classifier [8].

3. SPEAKER CLASSIFICATION METHODOLGY

Speaker recognition is a powerful tool for verifying identity in many applications. Speaker recognition may work on the user voice sample that is text dependent or text independent. This is more suitable in authentication systems—where a claimed user says specific phrase, such as a password or personal identification number, to be authenticated to access to the information system . In the proposed intelligent authentication, a claimed speaker claims an identity, and the main task is to verify if this identity is true. This done by classifying his voice sample with a set of models of authenticated speaker samples and deciding if the claimed speaker (class) is authenticated . This is main task of the classifier which is the essential of the recognition process. The data used in the proposed speaker classification system are subdivided into two portions: training data and test data. Train data samples are labeled with (the speaker class) as identification label to which this sample belongs. Test data are samples of voice belong to authenticated speakers which are labeled to testing the overall performance of the classification process [8].

Figure 1 show the general speaker classification system. MFCC features are extracted from the voice samples, these features represent the voice characteristics of the speaker. For the claimed speaker voice sample, the same features are extracted, and are compared against the features of other speakers. The comparison is implemented with contrast to a threshold. This threshold comparison indicates whether the two voice samples refer to the same speaker. If this comparison result is higher than a predefined threshold, then the system authenticated the speaker [2].

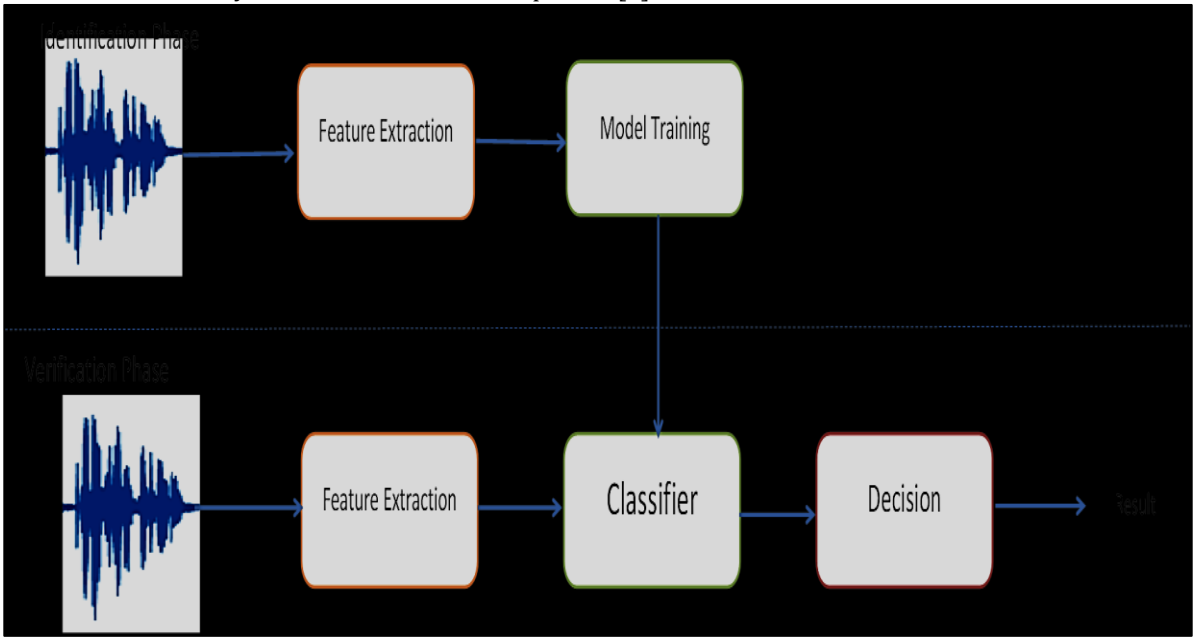


Figure 1 The General Speaker Classification System

3.1 MFCC Feature extraction

Short-term voice features are regarded for short duration because the voice signal is unceasingly changing as an effect of articulation. The voice signal is divided into short frames with durations of 20–30 ms. As a consequences of these small duration, the features are regarded to be stationary and these frames are chosen for spectral features obtaining. Mel is considered as a unit of sensed fundamental frequency [9]. MFCC is the most popular short-term acoustic features; these features are better from prosodic. The latter features suffer from many disadvantages, such as the difficulty of identifying the part of the signal that contains important information and determining an appropriate model of calculation as well as what is the amount of robust and efficiency when combined with the other characteristics [10]. These features are extracted from short voice frames of duration within 20–25 milliseconds [8]. This extraction process mimics the human hearing system. The following steps are used to compute MFCCs coefficients [11]

The MFCC computation could be summarized as follows:

1. Segment the signal into frames of 20 ms.
2. compute the periodogram estimate of the power spectrum of each frame.
3. compute the Mel filterbank of the power spectra,then sum the energy in each bank.
4. for all filter banks energies, compute logarithm.
5. for all filter banks energies, Compute the DCT.
6. preserve DCT coefficients 1-13.

3.2 Speaker Classification using DTW

DTW is a method "for measuring similarity between two time series which may vary (i.e. warp) in timing" [12] . This method can be used to discover the optimal alignment line between times sequences if one-time sequence may be "warped" by "stretching or shrinking it along time axis". This twisting between two time signals can be one choice to discover corresponding regions between the two time signals or to determine the similarity between them [13] .

The detailed computation of this method can be presented in Figure 2, each line connects a point in the sequence X to its similar point in the sequence Y. The lines have identical values on the y-axis, in the same time they have been dislocated so the vertical lines can be viewed. There are many metrices to find a distance between two sequences X and Y. In this work, the distance is the Euclidean distance defined by ||x−y||.

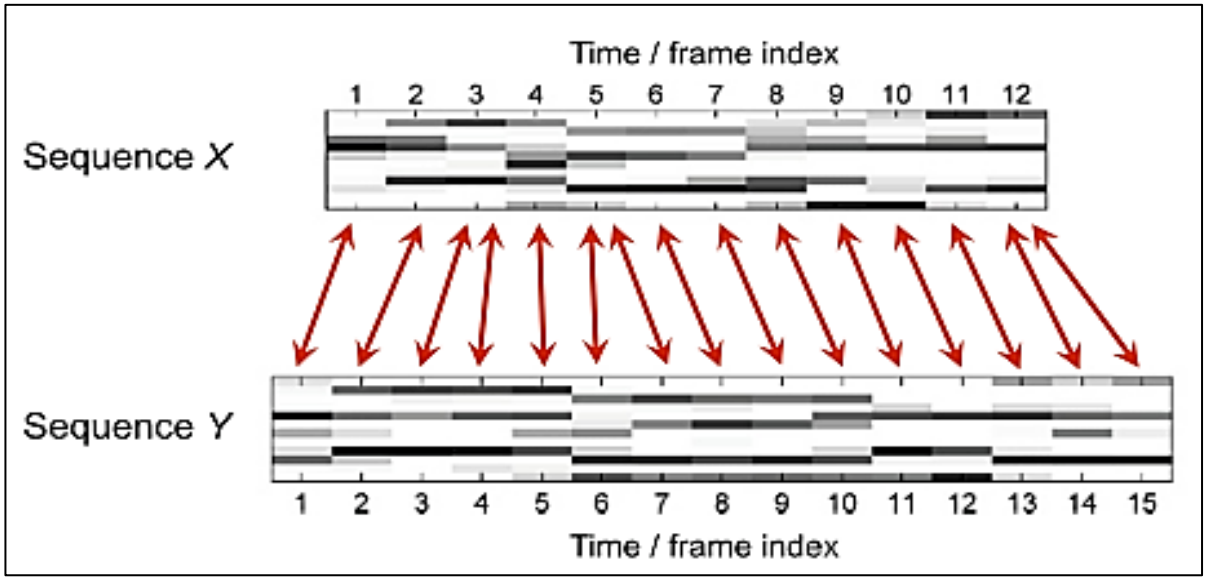


Figure 2 DTW Wrapping Process

The utilization of the DTW method in speaker classification is could be explained in the algorithm *DTW-Classification* :

Algorithm DTW-Classification *Input :MFCC1 features of test data set as unknown speaker and the MFCC2 features of the train data from dataset.* *Output: unknown speaker identity label*
Step1: load the MFCC feature base *Step2: assign the speakers identities as class labels, to have each feature vector is associated with a class label that represent speaker identity.* *Step3:for each MFCC of voice sample in test data do steps 4 and 5* *Step4: compute DTW distance (DIST) between MFCC1 and all MFCC2 features of the train data.* *Step 5: add DIST to rec_list and the identity (label) of speaker voice sample associated MFCC2.* *Step 6: select the minimum distance (Min_Dist) from rec_list* *Step 7:return the identity associated with Min_Dist.*

3.3 Speaker Classification using FIP:

Fuzzy set theory is based the approximate rather than crisp logic. Fuzzy truth represents the degree of approximation in sets, which is different from likelihood of a condition, since these sets are depend on vague concept, not randomness [15]. The two voice samples data (training and test) sometimes too close values, so fuzzification of these feature vectors can enhance recognition performance.

The recognition process goal is to find which element in feature vector Ai and the feature vector B most matches. To solve this problem, the inner product of fuzzy vectors is used.

One of the most important operation on fuzzy vectors that used in pattern recognition, is the fuzzy inner product. Assume *a* and *b* are fuzzy vectors of length *n*, then the FIP as follows [16]:

$\bigwedge_{i=1}^{n}(ai \wedge bi)$ …. (1)

If two fuzzy vectors are similar, a=b, the inner product has a maximum. These operations very useful when used in as a metric of sameness between two fuzzy vectors. The inner product of two fuzzy vectors could compute using Gaussian membership function as follows:

Let X= [-∞ ,∞], a 1D univers, A and B are two fuzzy sets having "normal, Gaussian membership "which are defined as:

$$\mu A(x) = \exp[-(x-a)\text{^}2/\sigma_a^2]$$

$\mu B(x) = \exp[-(x-b)\text{^}2/\sigma_b^2]$ …(2)

Where σ is standard deviation.

As shown in figure 3, and FIP of A, and B could be computed as follows:

$= \exp[-(a-b)\text{^}2/(\sigma a+\sigma b)\text{^}2\,] \;= \mu A(x0) = \mu B(x0)$ … … … … (3)

The utilization of such approach in speaker classification could be done by comparing the unknown data, to each of the known data pattern in pairwise order, to find the similarity value. The final decision is produced by selecting the data pattern which has the maximum approaching degree value. This recognized pattern is the pattern most like the unknown pattern. This concept defined as the maximum approaching degree [14].

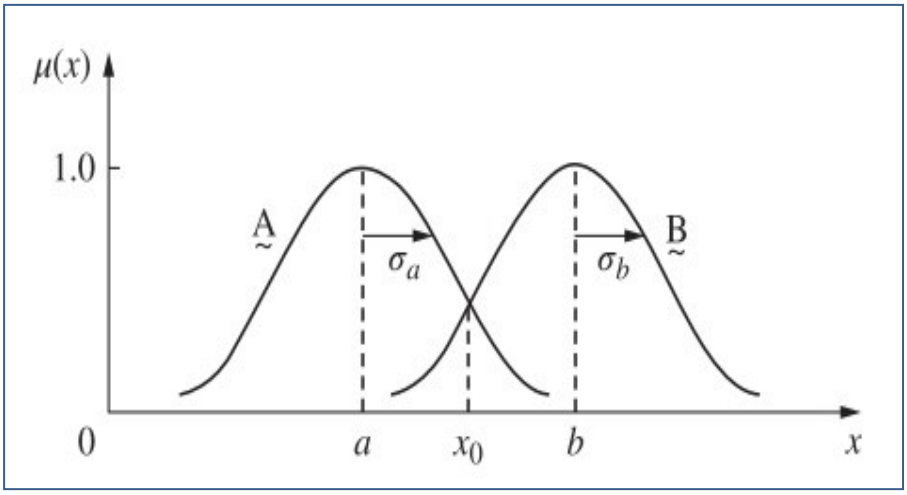
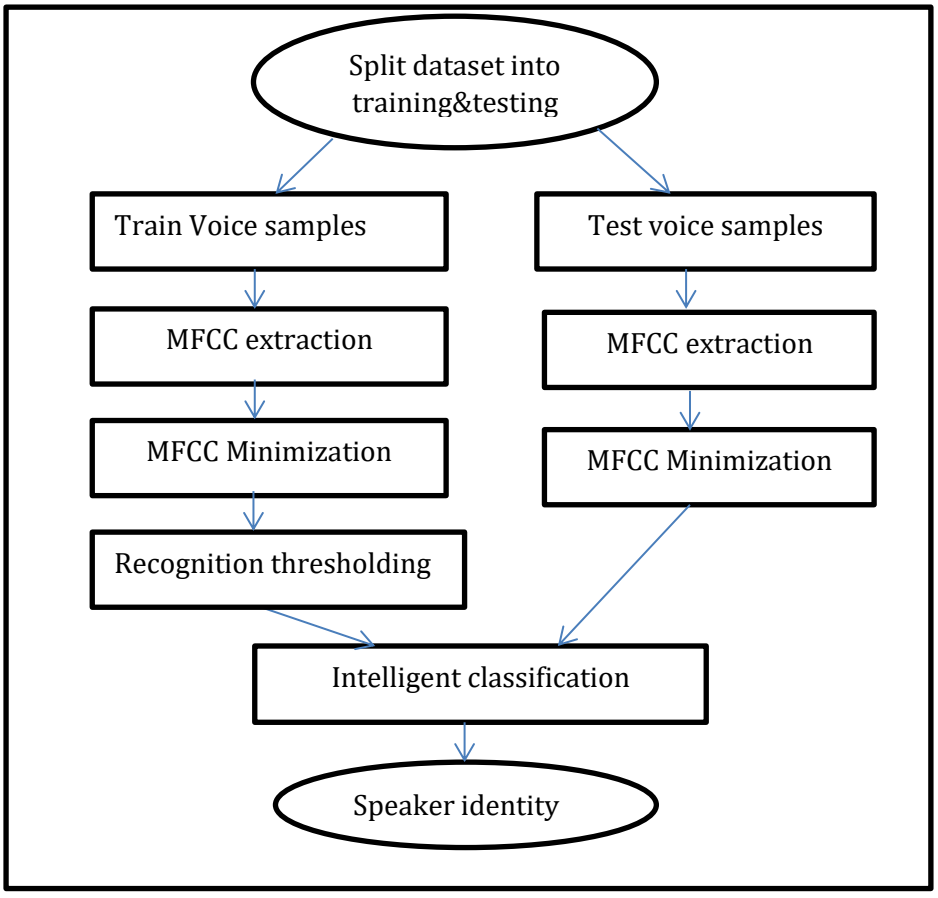


Figure 3 The Fip Of A, And B

The general workflow diagram of the proposed speaker classification using MFCC features and thresholding can be seen in figure 4:



While the detailed steps are presented in algorithm Intelligent Classification Method (ICM). The input of are the MFCC feature vector of unknown speaker which include 13 coefficients, and the database that store the MFCC features of speakers. Each speaker has a set of MFCC feature vectors each vector hold 13 coefficients. The last input is decision threshold (TH) that obtained from training .The output of the algorithm is the intelligent classification decision of the identity of unknown speaker.

Algorithm (ICM) *Input :the unknown speaker signal, the voice signals data set, and decision threshold (TH).* *Output: the speaker identity*
Step1: compute the MFCC array of the unknown speaker, MFCC1(T1,13) // T1 represent length of voice signal and 13 number of MFCC coefficients *Step2: Repeat for each voice signal (speaker2) in dataset* *Step3: Retrive MFCC features vectors MFCC_m(13) and MFCC_SD(13) from dataset.* *Step4: minimize MFCC1, by computing mean value of each MFCC coefficient to get user1_mean(13) and standard deviation of MFCC coefficient user1_ SD(13),* *Step5: compute fuzzy inner product between user1_mean, user2_SD and MFCC_m(13) and MFCC_SD(13)* *Step6: append the fuzzy inner product value of corresponding speaker2 to recognition_test_list* *step7: Until the last MFCC feature Vector in dataset.* *step8: Select the identity number (id) corresponding to the maximum value in recognition_test_list.* *Step 9: compare maximum inner product value with Threshold (TH),* *Step 10: Return the recognized user identity (id)*

4. Experiments and Results

In section 3.1 the MFCC feature extraction is presented and in section 3.2 and 3.3 the speaker classification methods and the related algorithms are explained. In this section the implementations and their results of the classification algorithms.

Each speaker in ELSDSR has 7 voice utterances for training and 2 utterances for testing. The implementation was done using python 2.7. Each voice utterance is stored in (wav) file format figure 5 show an example of speaker (MPRA_Sa) voice utterance signal.

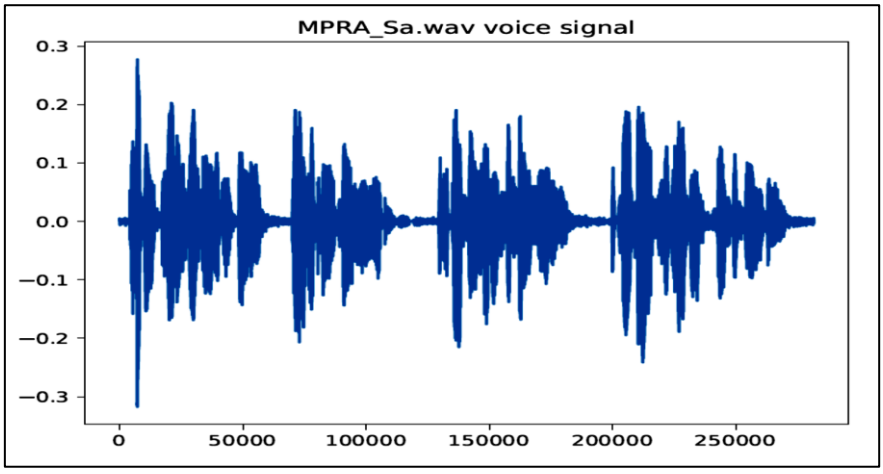


Figure 5 Example Of Voice Signal

Then these voice utterances transformed into MFCC features as explained in section 3.1. The output of MFCC extraction of each utterance is a matrix (13,T) where T is the length of the voice signal. Figure 6 shows an example of MFCC feature matrix where y-axis represent the power and x-axis is length of voice signal.

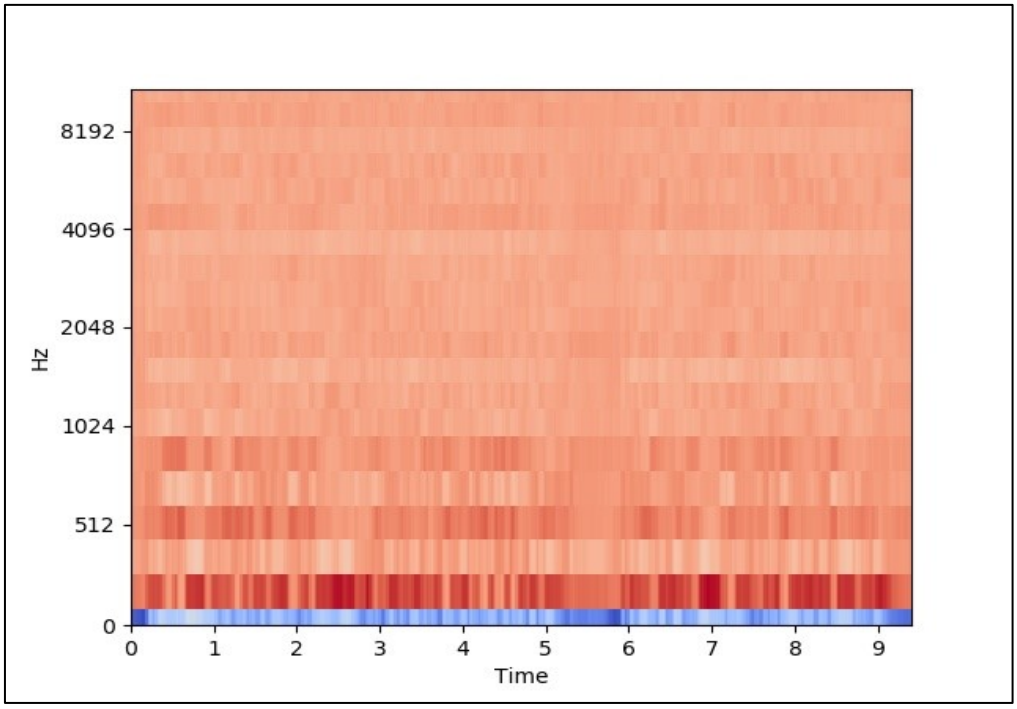


Figure 6 MFCC Features

The result of applying DTW algorithm is showed in table 1 using MFCC features and figure 7 and figure 8 shows the DTW warping line mentioned in section 3.2. This method is not discriminated enough as shown in table below. The main disadvantage of DTW algorithm is it could not create a model for the authenticated speaker compared. The second disadvantage , it is slower than the proposed fuzzy classifier.

Table 1 DTW Classification Result

Speaker actual class	Recognized speaker class	DTW distance
FAML	FAML	4406.55
FDHH	FDHH	5194.17
FEAB	FEAB	4153.37
FHRO	FEAB	4393.16
FJAZ	FJAZ	4774.74
FMEL	FMEL	3997.64
FMEV	FMEV	5080.03
FSLJ	FSLJ	3779.45
FTEJ	FTEJ	4154.65
FUAN	FUAN	3911.96
MASM	MASM	4497.54
MCBR	MCBR	3825.13
MFKC	FEAB	5688.36
MKBP	MKBP	4915.65
MLKH	MLKH	4845.6
MMLP	MPRA	5268.59

MMNA	MOEW	4561.57
MNHP	FAML	5678.58
MOEW	MOEW	5308.07
MPRA	MPRA	3832.33
MREM	MREM	4505.72
MTLS	MTLS	4369.58

The performance of DTW method in speaker classification is :

Total true accuracy =17/22*100= 77.27

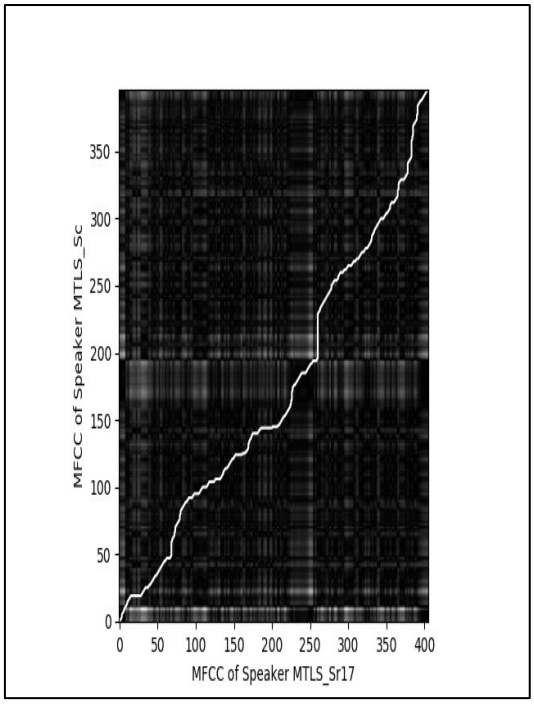


Figure 7 DTW Warping Line Of Two Different Voice Utterances Of The Same Speaker

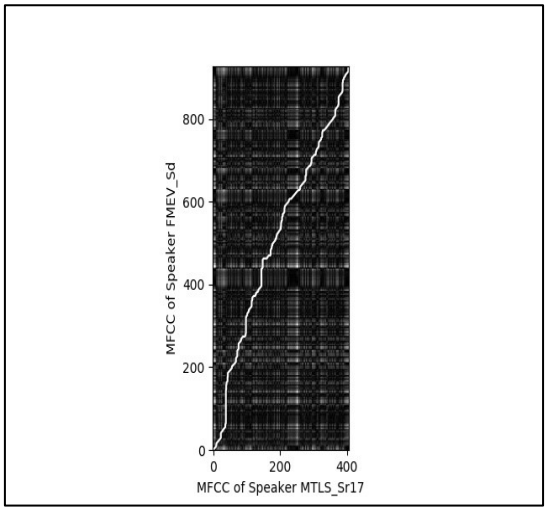


Figure 8 DTW Warping Line Of Two Different Voice Utterances Of Two Speakers

The implementation of the proposed fuzzy classifier showed high accuracy in speaker classification as shown in table 2.

Table 2 The Proposed Fuzzy Classifier Result

Actual speaker Class	recognized speaker class	Sum of Fuzzy inner product
FAML	FAML	12.7
FDHH	FDHH	12.91
FEAB	FEAB	12.86
FHRO	FHRO	12.78
FJAZ	FJAZ	12.75
FMEL	FMEL	12.79
FMEV	FMEV	12.82
FSLJ	FSLJ	12.88
FTEJ	FTEJ	12.95
FUAN	FUAN	12.83
MASM	MASM	12.76
MCBR	MCBR	12.83
MFKC	MFKC	12.81
MKBP	MKBP	12.7
MLKH	MLKH	12.83
MMLP	MCBR	12.71
MMNA	MMNA	12.71
MNHP	MNHP	12.37
MOEW	MASM	12.75
MPRA	MPRA	12.85
MREM	MREM	12.79
MTLS	MTLS	12.72

The total accuracy of the proposed classifier= 20/22*100= 90.91

5. CONCLUSION

The speaker classification methods is the core in many application related to the identity recognition of the user. Although there are many methods that applied in this type of classification like DTW, but the accuracy of this method is not good enough and this is important issue especially when the high recognition accuracy is required for specific security application like authentication. The main conclusion of this work, the proposed FIP classifier is more accurate than DTW method and it is faster since it implemented with less computation operation than DTW method.

6. FUTURE WORK

We intend to evaluate the proposed fuzzy classifier with other datasets and compare its results with other classifiers like artificial neural network and Gaussian Mixture Models.

7. REFERENCES

[1] J. G. Carbonell and J. Siekmann, "Speaker Classification I Fundamentals, Features, and Methods",Lecture Notes in Artificial Intelligence, 2007.

[2] P. T. Nguyen, "Automatic Speaker Classification Based on Voice Characteristics," MSC. Thesis, 2010.

[3] L. Feng, English Language Speech Database for Speaker Recognition (ELSDSR), department of Informatics and mathematical modelling, Technical University of Denmark (DTU),2004.

[4] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," IEEE Circuits Syst. Mag., vol. 11, no. 2, pp. 23–61, 2011.

[5] S. Arora, "Speech Recognition of Offline Attendance System : A Review Speech Recognition of Offline Attendance System : A Review," no. June, 2016.

[6] H. Gan, I. Mporas, S. Safavi, and R. Sotudeh, "Speaker Identification Using Data-Driven Score Classification," Image Process. Commun., vol. 21, no. 2, pp. 55–64,2016.

[7] Swathy M S1 , Mahesh K R, "Review on Feature Extraction and Classification Techniques in Speaker Recognition," Int. J. Eng. Res. Gen. Sci., vol. 5, no. 2.

[8] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods," in Procedia Computer Science, 7th International Conference on Advances in Computing & Communications, ICACC-2017, Cochin, India 2017.

[9] J. H. L. Hansen, "Speaker Recognition by Machines and Humans," no. November, 2015.

[10] S. Sremath Tirumala, S. Reza Shahamiri, A. Singh Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," Expert Systems With Applications, vol. 90, pp. 250–271, 2017.

[11] E. S. Al-Shamery And W. M. Al-Hameed, "Two Scopes Of Acoustic Signal And Fuzzy-Relief Algorithm For Improving Automatic Speake Recognition", Journal Of Theoretical And Applied Information Technology, Vol.97. No 2 ,2019.

[12] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ accessed on 11:40 PM on 04/05/2019.

[13] A. Mueen and E. Keogh, "Extracting Optimal Performance from Dynamic Time Warping," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016.

[14] A. Bala, A. Kumar, and N. Birla, "VOICE COMMAND RECOGNITION SYSTEM BASED ON MFCC AND DTW," International Journal of Engineering Science and Technology, vol. 2, no. 12, pp. 7335–7342, 2010.

[15] M. Müller, Fundamentals of Music Processing. 2015.

[16] K.R.Venugopal, K.G. Srinivasa and L.M. Patnaik, Soft Computing for Data Min-ing Applications,Springer,2009.

[17] T. J. Ross, Fuzzy Logic With Engineering Applications, Fourth Edition, 2017.