Engineering and Technology Journal
Journal homepage: engtechjournal.org

A Proposed Speaker Recognition Method Based on Long-Term Voice Features and Fuzzy Logic

Iman H. Hadi [a] *, **Alia K. Abdul-Hassan** [b]

[a] GSCOM, Baghdad, Iraq. iman.h.1439@gmail.com.

[b] Computer Sciences Department, University of Technology, Baghdad, Iraq. 110018@uotechnology.edu.iq.

*Corresponding author.

Submitted: 10/06/2019 Accepted: 15/09/2019 Published: 25/03/2021

K E Y W O R D S

Speaker identity, Voice, frequency features, Fuzzy Vector.

A B S T R A C T

Speaker recognition depends on specific predefined steps. The most important steps are feature extraction and features matching. In addition, the category of the speaker voice features has an impact on the recognition process. The proposed speaker recognition makes use of biometric (voice) attributes to recognize the identity of the speaker. The long-term features were used such that maximum frequency, pitch and zero crossing rate (ZCR). In features matching step, the fuzzy inner product was used between feature vectors to compute the matching value between a claimed speaker voice utterance and test voice utterances. The experiments implemented using (ELSDSR) data set. These experiments showed that the recognition accuracy is 100% when using text dependent speaker recognition.

How to cite this article: I. H. Hadi and A. K. Abdul-Hassan, "A Proposed Speaker Recognition Method Based on Long-Term Voice Features and Fuzzy logic," Engineering and Technology Journal, Vol. 39, Part B, No. 01, pp. 1-10, 2021.
DOI: https://doi.org/10.30684/etj.v39i1B.343

1. INTRODUCTION

One of the most critical tasks of information security is developing a method to recognize the identity of the user based on the digitization of that identity which is required to access these systems with an acceptable level of trust. The identity refers to a group of well-defined properties that make an entity recognized compared to other entities [1]. While the digital identity is a set of features owned by an entity used by information systems to represent an identity of individual. Voice attributes rich with information that could be digitized to recognize between users. The voice attributes have many advantages , they are unique and easy to capture.

I. Related Work

The powerful recognition method depends on unique biometric attributes, like voice attributes, so many researches proposed methods that depend on speaker voice attributes. A. U. Khan et al. (2012)

proposed a text-dependent speaker identification method which used the feature of the extraction takes place by Zero Crossing Rate (ZCR) and similarity distance for recognition [2]. G. Luqman (2013) Utilizes discrete Fourier transformation (DFT) to compare the frequency spectrum of two voice utterances, in addition to Euclidean Norm with template matching with accuracy 70% [3]. R. A. Sadewa et al. (2015) proposed a method that uses the characteristic of the voice by extracting features called Mel Frequency Cepstral Coefficients (MFCC) and matching these features using Vector Quantization (VQ), this method modified with a predefined threshold and the true acceptance using trained data is around 86% [4]. A. Banerjee et al. (2018) used short term spectral features learned from the Deep Belief Networks (DBN) augmented with MFCC features to perform the task of speaker recognition, they achieved a recognition accuracy of 95% as compared to 90% when using standalone MFCC features on the ELSDSR dataset [5].

II. Voice Biometric Features

Voice biometric feature has a uniqueness attribute that can be used efficiently to discriminate between identities in the recognition process [6]. Every person has different vocal features. These features can be divided into two main types: long-term (Prosodic) features comprise of the pitch, (ZCR). Short-term features are extracted within transforming the voice wave to the frequency domain. This is done by applying the (DFT) techniques. Then the voice signal is broken down into short portions having intervals of 25–30 milliseconds [7]. It is important to consider the variability aspect of voice sample that may be due to illnesses or other reasons, but these samples are quite reiterate and several of samples would be sufficient for acceptable recognition [8]. To overcome the variability in voice, it is possible to create a monitoring technique to keep track of the changes in the voice during the use of the authentication system.

2. SPEAKER RECOGNITION METHODOLOGY

Speaker recognition (SR) is a method used to verify user personality in security application. SR system has two approaches: text dependent or text independent approaches. Text-dependent speaker recognition enforces users to speak an exact sentence; this sentence is predefined in the enrollment phase, to be recognized from other speakers. While in text independent approach, the user could use any phrase to be recognized by the system. In the speaker recognition, a pretended speaker presents his identity through his voice sample. The SR system verifies his identity if it is right otherwise he is rejected. This is done by matching the voice utterance with a set of known speaker voice utterances for proofing the actual identity [9]. The voice prints data used in the SR system are usually break into two main divisions: first part for training and the second for test. Train voice samples are labeled with the user identity. Test voice samples are also labeled for performance evaluation purpose.

In Figure 1, the speaker recognition general structure is presented [9]. The voice attributes are derived from the voice utterances that are designed to symbolize the speaker's voice. For the pretended user voice utterance, the attributes are derived and matched against the attributes of known speakers. The matching score indicates whether these features refer to proofed speaker. If the score is above a specific threshold, the SR system accepts the test data that belong to a specific speaker.

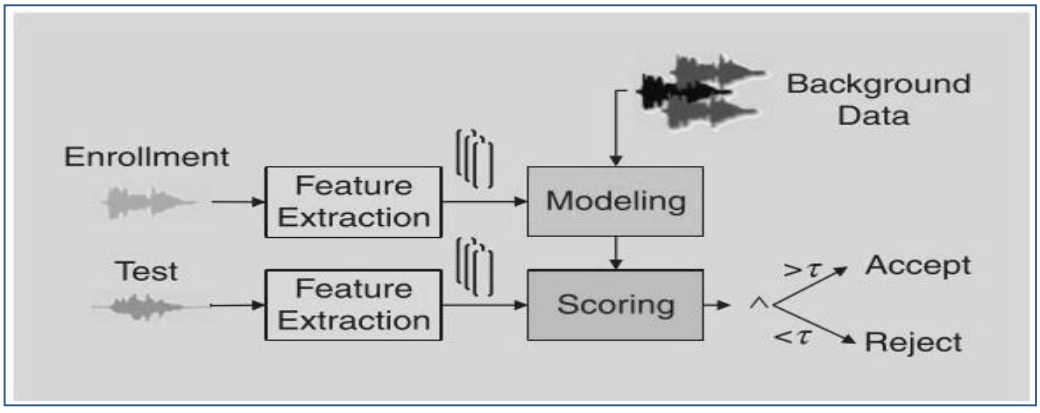

Figure 1: the Speaker Recognition Architecture

3. LONG TERM VOICE FEATURES EXTRACTION

Long term feature extraction is related to the characteristics of the voice signal as a whole. That mean they are computed without framing these data into small portions as short-term features (e.g.) Mel frequency cepstral coefficients (MFCC) [9]. In this paper, the long term features are used to characterize the identity of the speaker:

I. The Pitch

Pitch is a sensed quantity which is related to the fundamental frequency of oscillation of the vocal's cords during some duration [8] , [10]. It is the rise and fall of the voice during speaking, sometimes called "highness" or "lowness" [11]. In this paper, the pitch is estimated using quadratic-ally interpolated FFT (QIFFT) method which is implemented using Librosa python library.

II. Maximum *Peak Feature:*

Maximum peak in frequency of voice signal is one feature that can be used in speaker recognition. This feature can be extracted by converting this signal from the time domain into the frequency domain using the (DFT) [12]. The following Algorithm presented the computation of frequency maximum peak (FMP) as long term feature of the wave signal. This algorithm depends on DFT which could be explained as following: Let S(n) represent a speech signal. S(n) is transformed to frequency domain by an N point using equation 1:

$$\text{DFT(S)} = \sum_{m=0}^{N-1} s_m \exp\{-2\pi i \frac{mk}{N}\} \qquad (1)$$

This equation determines the frequency content of the signal by computing DFT(S) coefficients. These coefficients represent how much similarity between the original signal S and complex exponential function $\exp\{-2\pi i\frac{mk}{N}\}$.

III. Zero Crossing Count feature :

The ZCC is a measure of how many the signal values crosses the zero line. The idea is that the ZCC gives information about the spectral data of the waveform [11]. Algorithm ZCR shows the calculation of the zero crossing count in the voice signal.

Algorithm FMP

Input: voice audio file S, Fs signal sampling rate

Output: Maximum peak in S

```
Begin
Step 1: compute DFT(S)←eq. 1
Step2: max =maximum (DFT(s)).
Step3 :for i=0 to length(DFT(s))do
Temp=i
 freq=temp*(Fs/length(DFT(s)))
step4:for j=0 to length(DFT(s)) do
if freq[j] < Fs/2 then
useful_freq[i] = freq[i]
step5:ind=round(lenghth(DFT(s))/2)
for k=1 to ind do
useful_DFT = abs(DFT[k])                    # Take the absolute magnitude.
Step6: return the index of  max (useful_DFT )
END
```

Algorithm ZCR
Input: voice audio file S
Output: zero crossing rate(ZCR) in S
Step1 :Repeat
Step 2:If sign(s(i-1)) not equal to sign(s(i))
Step3: Zcc=zcc+1
Step4 : i=i+1
Step5 :ZCR= Zcc/i
Step6:Until EOF(s),
Step7: Return ZCR

4. FEATURE MATCHING USING FUZZY VECTORS

Fuzzy logic is stand on the approximate rather than crisp logic. This theory depends on "truth represents the degree of approximation in sets, which is different from the likelihood of a condition, since these sets are based on vague definition, not randomness" [13]. The utterance samples of the speaker's voice (training sample and test sample), sometimes have very similar values. The fuzzification of such attributes vectors can improve the recognition process.

There are certain features and operations implemented using fuzzy sets which could be used as fuzzy pattern recognition methods which depend on the" *maximum approaching degree*" [14]. To clarify this concept, let us define *a* and *b* as feature vectors, their length is represented by *n*, then the fuzzy inner product is as in equation (2):

$$\Lambda_{i=1}^{n}(ai \wedge bi) \qquad (2)$$

After fuzzifying these two vectors, and if they are similar *a=b*, the inner product reaches a maximum value compared with other samples. This norm, the inner product, can be used in any pattern recognition method (like speaker recognition) because it recognizes the closeness or the similarity between features of different patterns.

Let X= [-∞,∞], a 1D dimension universe on the real domain, A and B are two vectors with normal Gaussian membership, defined mathematically by the equations:

$$\mu A(x) = \exp[-(x-a)\hat{}2/\sigma_a^2] \qquad (3)$$
$$\mu B(x) = \exp[-(x-b)\hat{}2/\sigma_b^2] \qquad (4)$$

Where σ is the standard deviation, and *a*, *b* are the mean of A and B.

This operation is efficient when used as a measure of similarity between two feature patterns. The inner product between them shown in figure 2, are computed using Gaussian membership function as in the following equations (9) [14] :

$$inner\ product\ (A,B) = \exp[-(a-b)\hat{}2/(\sigma a+\sigma b)\hat{}2\ ] \qquad (5)$$

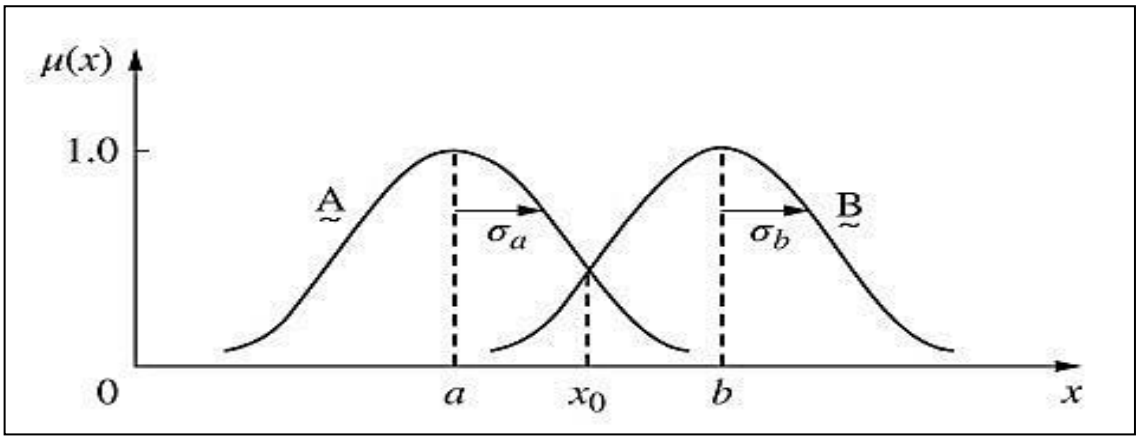

Figure 2: The inner product between A and B

The following example shows patterns which can all be represented by Gaussian membership functions, *Ai*, where i =1,2,…,6 and parameters a_i and σ_{ai} define the shape of each membership function. Table I provides information for the six regions. The unknown pattern, represented by a fuzzy set *B*, with the following characteristics [14]: $b=41$, $\sigma_b = 10$. In order determine the maximum approaching degree , calculation has been done using the equation (5) which represents the inner product between B and Ai Where i=1,..., 6.

TABLE I: Parameters for Gaussian membership function

	A1	***A2***	***A3***	***A4***	***A5***	***A6***
a_i	5	20	35	49	71	92
σ_{ai}	3	10	13	26	18	4

The inner product results : (B,A1)=0.5 ,(B,A2)=0.67 ,(B,A3) =0.97 ,(B,A4)= 0.98 ,(B,A5)=0.65(B,A6)=0.5

As a result, **B is similar to A4**, because the inner product value between B and A4 has the maximum value 0.98.

The implementation of this method in speaker recognition could be done by comparing the data sample to each of the known voice features vectors in pairwise order, to find the approaching value for each pair, and select the pair with the maximum value. The recognized vector is the pattern with the maximum approaching value [14].

5. THE PROPOSED SPEAKER RECOGNITION METHOD

The proposed method for speaker recognition is consists from two main processes. The first process is the long term voice features extraction. The second process is the features matching method using fuzzy inner product which presented in section 4. These two processes are employed in the proposed algorithm (fuzzy matching).

The input of the algorithm is the voice utterances of the claimed user and test voice utterances. In addition, the sampling rate of voice signals is the last input of the algorithm. The features extraction is implemented by employing the algorithms presented in section 3. The output of the algorithm is the recognized identity of the claimed speaker voice utterance. This recognized identity is produced using Eq.5 that is presented in section 4. The main steps of the proposed method could be summarized by the following algorithm:

Algorithm Fuzzy Matching

Input: speaker voice file of the claimed speaker (S1) and testing voice utterances S2 , S_Rate the voice signal sampling rate.

Output: speaker recognized identity SP_ID

Begin

Step1 : load the voice signal S1.

Spep2:extract Max_freq← Algorithm FMP(S1, S_Rate)

Sp_Feat(S1) ← Max_freq #store Max_freq into feature vector Sp_Feat

Step3: extract ZCR ← Algorithm ZCR(S1)

Sp_Feat(S1) ← ZCR

Step 4:extract Pitch← QIFFT(S1)

Sp1_Feat(S) ← Pitch

Step5: for each speaker voice signal (S2) in test utterances do steps 6 and 7:

Step6: extract features of S2 :

Sp2_Feat_Test ← repeating steps 3,4,5 #Sp2_Feat_Test is the feature vector of test utterance

Step 7:Compute fuzzy inner product (FIP)

FIP(Sp1_Feat(S1) , Sp2_Feat(S2)) ← employing eq.5

Rec_list(sp_Test) =FIP

step 8:rec_id=Max(Rec_list)

step 9: return rec_id

End

6. Experiments and Result

The main implementation was done using English Language Speech Database for Speaker Recognition (ELSDSR) which consists of 7 audio file samples for each speaker; the total number of speakers is 22 volunteers. The text language is English [15]. The voice samples are recorded into the file type (.wav).

We choose two different recorded voice file for each speaker from this dataset. Then each file is loaded into an array (as plotted in figure 3) using "LibRosa" python library.

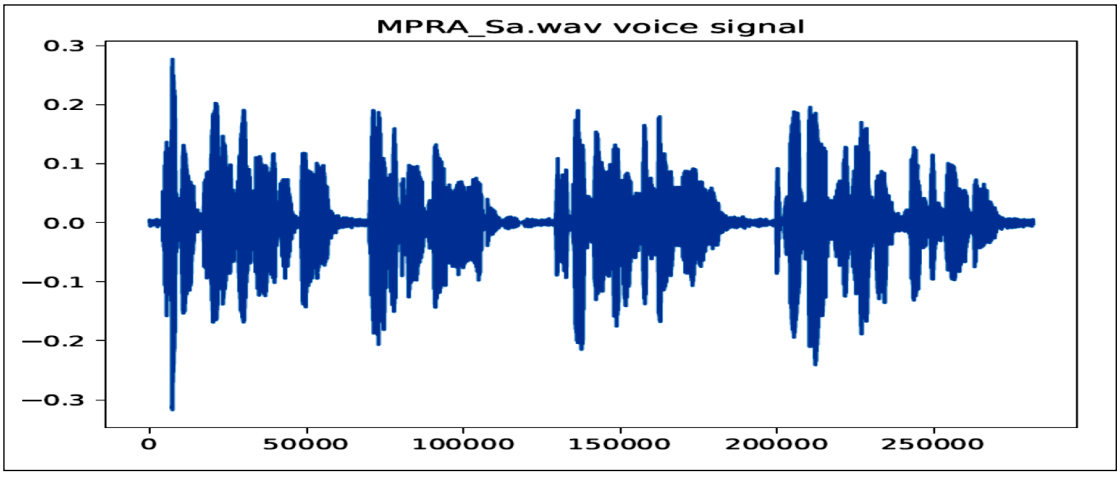

Figure 3: Voice wave signal

The extraction of long-term feature (as shown in figures 4,5,6,7) has been done in order to create a feature vector for each speaker that contains 3 values represent (maximum peak, ZCR, and Pitch) to use it in the recognition process.

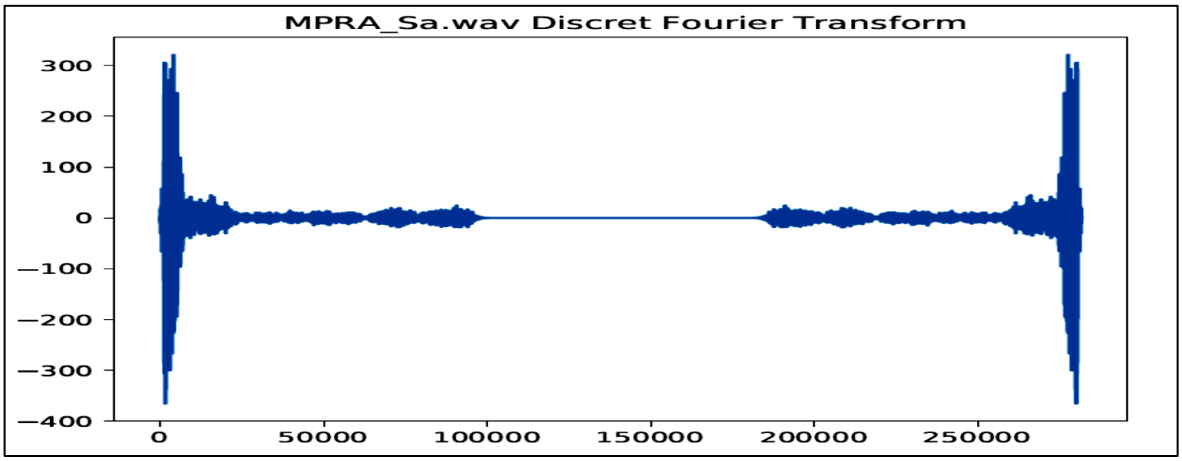

Figure 4: DFF transform

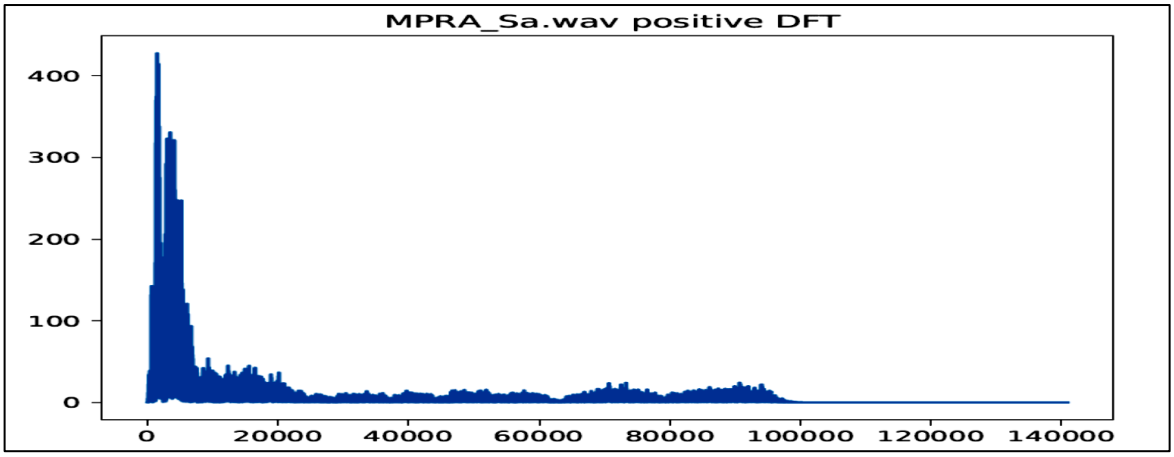

Figure 5: the positive frequency

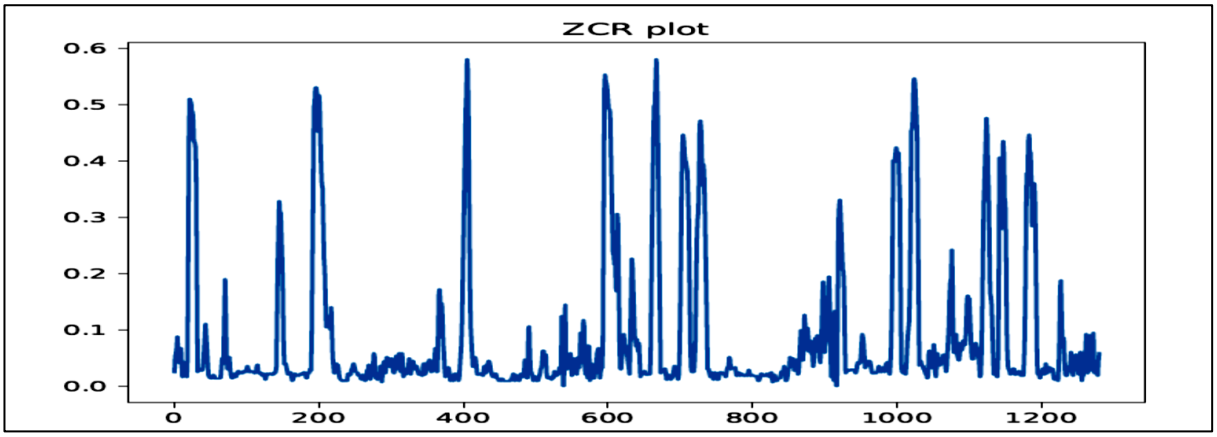

Figure 6: ZCR values

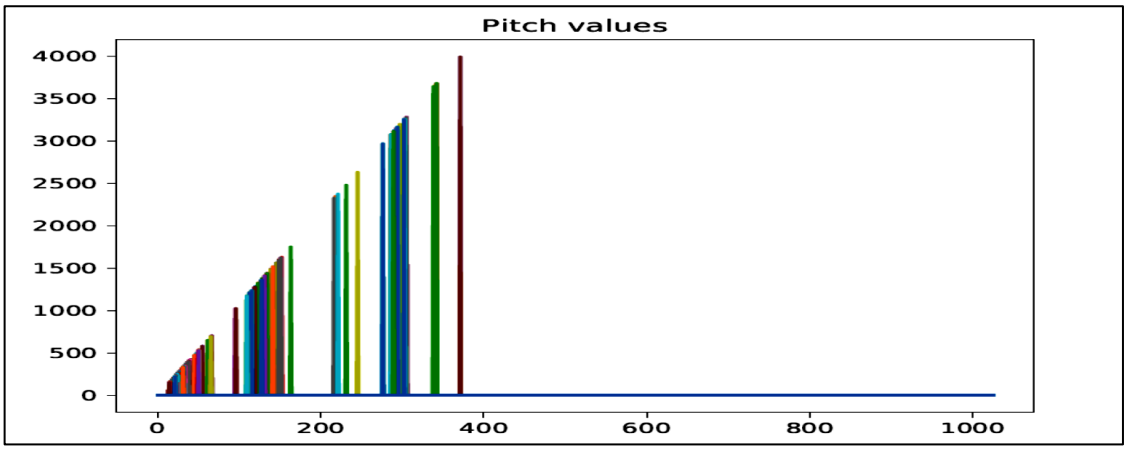

Figure 7: Pitch values

These features are computed from the voice signal as one entity (as shown in Table II), each feature type did not have discriminated property as a single feature for speaker identity recognition. But if we combine these three features in one vector it could be discriminated enough.

TABLE II: long term features

	MaxFreq	ZCR	Pitch
FAML	186.09	124.7	149.2
FDHH	189.82	109.4	145.4
FEAB	193.03	134.4	150.1
FHRO	243.88	122.5	182.6
FJAZ	199.5	120.5	165.6
FMEL	223.62	108.9	159.6
FMEV	236.79	104.9	175.6
FSLJ	177.47	103.0	147.2
FTEJ	208.69	93.86	149.4
FUAN	325.98	85.39	145.7
MASM	315.35	63.62	145.5
MCBR	234.56	135.5	145.5
MFKC	235.2	76.65	146.1
MKBP	344.16	80.81	145.8
MLKH	327.12	153.4	145.3
MMLP	331.48	88.77	147.7
MMNA	95.89	73.57	146.7
MNHP	105.14	109.1	145.4
MOEW	226.96	127.3	146.8
MPRA	114.08	110.8	145.5
MREM	245.33	110.0	146.3
MTLS	96.29	61.45	146.7

Table III shows the fuzzy matching of short-term features vectors using fuzzy inner product operation, to recognize the identity of speaker MTLS using text independent speaker recognition. The recognized identity was not accurate using text independent recognition. The speaker actual identity (MTLS) was recognized as false identities FAML,FDHH,FUAN and MKBP the maximum matching values (which is equal to 1) referred to other identities other than the actual identity (MTLS). Table IV shows the results of fuzzy matching values for speaker MTLS using text dependent recognition as an example. The maximum value of the fuzzy inner product between two voice utterances belong to speaker MTLS is equal to 1. That means it is important to use the same spoken words in speaker recognition when using long-term features.

TABLE III: Text-independent dependent recognition MTLS

Speaker	Fuzzy Matching
FAML	1
FDHH	1
FEAB	0.99
FHRO	0.93
FJAZ	0.93
FMEL	0.99
FMEV	0.99
FSLJ	0.97
FTEJ	0.99
FUAN	1
MASM	0.96
MCBR	0.98
MFKC	0.97
MKBP	1
MLKH	0.95
MMLP	0.95
MMNA	0.77
MNHP	0.85
MOEW	0.98
MPRA	0.87
MREM	0.98
MTLS	0.76

TABLE IV: Text dependent recognition of speaker MTLS

Speaker	Fuzzy Matching
FAML	0.84
FDHH	0.86
FEAB	0.83
FHRO	0.84
FJAZ	0.84
FMEL	0.86
FMEV	0.86
FSLJ	0.87
FTEJ	0.89
FUAN	0.9
MASM	0.94
MCBR	0.83
MFKC	0.92
MKBP	0.91
MLKH	0.89
MMLP	0.93
MMNA	0.86
MNHP	0.83
MOEW	0.86
MPRA	0.86
MREM	0.96
MTLS	1

Finally, the experiments show the results of text-dependent recognition of all speakers in the dataset are similar to speaker MTLS example that mean the accuracy is 100% when using inner product between identical voice utterances as shown for speaker MTLS in table IV.

7. CONCLUSION

Speaker recognition depends on the type of features extracted from the speaker voice signal. In addition, the accuracy of the recognition process relays on the matching technique between the train and test voice samples. In this paper, the conclusion resulted from the experiments shows that using the long-term one feature as alone decrease the recognition accuracy, while using a feature vector that contains more than one feature (maximum frequency, ZCR, and Pitch) is more accurate. Feature vector matching using the inner product of fuzzy vectors for matching two voice samples with the same spoken words increased the recognition rate more than text-independent approach.

8. FUTURE WORK

We intend to propose a fusion technique to combine more than one type of features, long-term features and short-term features like MFCC and linear predictive coefficients (LPC) to enhance recognition accuracy, especially with text-independent speaker recognition.

References

[1] G. B. Ayed, "Architecting User-Centric Privacy-as-a-Set-of-Services: Digital Identity-Related Privacy Framework", Springer, 2014.

[2] A. U. Khan, L.P. Bhaiya, S. K. Banchhor, "Hindi Speaking Person Identification Using Zero Crossing Rate," International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-3, July 2012.

[3] G. Luqman. "Voice Recognition System Using Template Matching.", International Journal of Research in Computer Science, 2013.

[4] R. A. Sadewa, T. A. B. Wirayuda, S. Sa'Adah, "Speaker recognition implementation for authentication using filtered MFCC - VQ and a thresholding method", 3rd International Conference on Information and Communication Technology, ICoICT, 2015.

[5] A. Banerjee, A. Dubey, A. Menon, S. Nanda, and G. Chand Nandi, "Speaker Recognition using Deep Belief Networks". Robotics and Artificial Intelligence Laboratory, Indian Institute of Information Technology, 2018.

[6] F. O. Karray and C. De Silva, "Soft Computing and Intelligent Systems Design: Theory, Tools, and Applications", 2004.

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

[8] H. Beigi, "Fundamentals of Speaker Recognition." 2011.

[9] J. H.L. Hansen, T. Hasan, "Speaker Recognition by Machines and Humans", IEEE Signal Processing Magazine, 2015

[10] https://pronuncian.com/pitch-lessons/

[11] D. Gerhard, "Pitch Extraction and Fundamental Frequency: History and Current Techniques," 2003.

[12] S. Memon ,"Automatic Speaker Recognition: Modelling, Feature Extraction and Effects of Clinical Environment", PHD thesis, RMIT University,2010

[13] K. R. Venugopal, K.G. Srinivasa and L.M. Patnaik, "Soft Computing for Data Mining Applications", Springer,2009.

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

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