DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

Speaker Recognition System Based on Mel Frequency Cepstral Coefficient and Four Features

Ashraf Tahseen Ali[1], Hasanen S. Abdullah[2], Mohammad N. Fadhil[3]

[1, 2, 3] *Department of Computer Science, University of Technology- Iraq, Baghdad, Iraq*

[1]*cs.19.12@grad.uotechnology.edu.iq,* [2]*110014@uotechnology.edu.iq,*

[3]*mohammad.n.fadhil@uotechnology.edu.iq*

Abstract– *Biometrics signs are the most important factor in the human recognition field and considered an effective technique for person authentication systems. Voice recognition is a popular method to use due to its ease of implementation and acceptable effectiveness. This research paper will introduce a speaker recognition system that consists of preprocessing techniques to eliminate noise and make the sound smoother. For the feature extraction stage, the method Mel Frequency Cepstral Coefficient (MFCC) is used, and in the second step, the four features (FF) Mean, Standard Division, Zero-Cross and Amplitude, which added to (MFCC) to improve the results. For data representation, vector quantization has been used. The evaluation method (k-fold cross-validation) has been used. Supervised machine learning (SML) is proposed using Quadratic Discriminant Analysis (QDA) classification algorithms. And the results obtained by the algorithm (QDA) varied between 98 percent and 98.43 percent, depending on the way of features extraction that was used. These results are satisfactory and reliable.*

Index Terms— *SML, QDA, Voice Recognition, MFCC, FF.*

I. INTRODUCTION

Biometrics can be described as a science that is applied to know the personal identification by the employment of the physical and behavioral traits, such as fingerprints, finger veins, face, palm veins and voice [1]. In addition to recognizing handwritten fonts and knowing them, the person's font is one of the behavioral features through which a person can be identified [2]. Speech is the communication, or an illustration means that provides the knowledge and considering the special features of the speaker [3]. The voice is one of the means to identify a person. The speaker recognition system can check a person's condition by hearing his voice, for instance, the gender, speaker's identification, accent, and speech, emotion, and health conditions of the speaker. Along with technological improvements, it has also practical applications, for instance on voice-based security systems [4]. Speech Recognition (SR) methods can be classified toward a large amount of groups based on their capacity to identify various words. Classes of speech recognition are categorized into four techniques, the first is Isolated Speech (IS) which means Isolated words normally including a gap among two sentences; it didn’t indicate that it just allows one word, just requires one utterance at a time. The second is Connected Speech (CS) or Connected words (CWs) is alike to isolate words but allows disconnected utterances with the least pauses separating them. The third is Continuous Speech (CoS) which enables the user to speak relatively normally and is likewise named computer dictation. The last is Spontaneous Speech (SS) at a primary level; it can be imagined as speech that is natural-sounding and not related. An ASR system with a spontaneous

Received 6/7/2021; Accepted 26l8l2021

DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

speech facility should be ready to work a different of common speech traits like words being pronounced together and even little stutters [5]. Preprocessing is the first stage in the voice recognition system. The aim of preprocessing the audio is to decrease the signal to noise ratio. Various filters and techniques can be used to a voice signal to decrease the correlated noise. Framing, windowing, normalization, end-point detection and pre-emphasis are some of the commonly employed techniques to decrease noise in a signal [6]. The extraction for features is an essential and significant stage in the voice recognition system [7].

II. LITERATURE SURVEY

The fourth quarter of the twentieth century was considered the starting point for studies in automated speaking identification by machines, and has excited a lot of attention for many purposes varying of knowledge domain nosiness around the processes for the mechanical understanding of the persons' speech capabilities to automatize easy purposes that need human-machine interactions [8]. In this part, several of the preceding studies that used (SML) classifiers, which related to this research, are reviewed:

Athulya et al. [9] suggested a speaker verification method using the Gaussian Mixture Model with Universal Background Model (GMM-UBM) and SVM to recognize the speakers, applying MFCC and Power Normalized Cepstral Coefficients (PNCC) called (MPNCCs) as feature extraction techniques. The output obtained from the (GMM-UBM and SVM) classifiers showed a promising result reached 97.5 percent. Chauhan et al. [10] suggested a research to be done among several mixtures of features for speaker recognition method with Feed-Forward Artificial Neural Network (FFANN) and Support Vector Machine (SVM). Linear Predictive Codes (LPC), MFCC and zero-crossing rate (ZCR) are employed as the extraction of feature methods. Artificial Neural Network (ANN) was achieved an accuracy of 93.1 percent while SVM reached 79.1 percent. Herrera et al. [11] presented a voice corpus design, based on recordings from speakers distinct from any other corpus currently in practice. For its evaluation, they employed an MFCC with a GMM for features extraction and a Maximum Likelihood Estimation (MLE) method for classification. Results prove the accuracy of 93 percent. Devi et al. [12] suggested an Automatic Speech Recognition (ASR) using speech signals. Feature extraction employed by MFCC. The dimensions have been reduced by applying Self Organizing Neural Network (SOFM) so that the supervised classifier can be trained accurately. Lastly, applying the lessened input instance identification is done through Multilayer Perceptron (MLP) by Bayesian Regularization (BR). The training of the network has tested by a real speech dataset from Multivariability speaker recognition belonged to ten speakers. MLP-BR with SOFM presents enhanced performance and the accuracy reached 93.33 percent. Rao et al. [13] offer different methods of audio preprocessing such as trimming, split and merge, noise decrease and voiced improvements to improve the audios collected from real-world conditions. MFCC and Delta-Delta (DD) used for features extraction for each audio, along with them to evaluate k-Nearest Neighbor (KNN) methods. The algorithm has achieved an accuracy of 68.1 percent. Nawas et al. [14] implemented a voice recognition method using the Random Forest (RF) algorithm to recognize the different speakers, applying MFCC and Reconstructed Phase Space (RPS) as feature extraction methods on the data taken from the TIMIT corpus dataset. The result was achieved 97 percent.

III. THE SUGGESTED SYSTEM STAGES

The stages of the proposed method that uses voice recognition depending on (SML) are: Database Design consists of voice records, Preprocessing, Feature extraction, Evolution method and

Received 6/7/2021; Accepted 26l8l2021

DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

(QDA) classifier. *Fig.* 1 depicts the proposed method design.

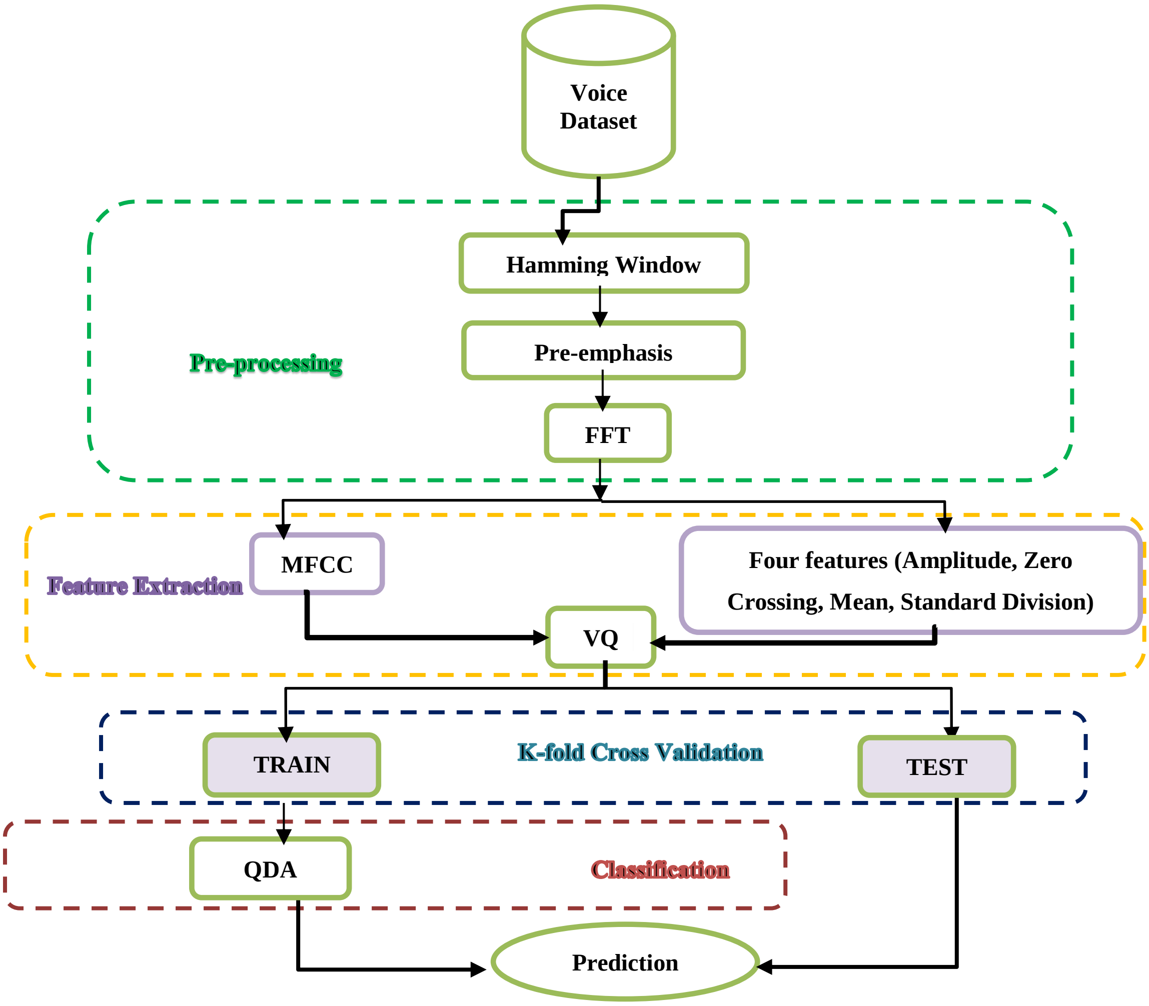


FIG.1. THE BLOCK DIAGRAM FOR PROPOSED SPEAK RECOGNITION MODEL

A. Database design

In this research, the voice dataset was used (Prominent leader’s speeches), including audio clips of 5 prominent politicians in their countries and the world, in the open air. It contains noise in a certain proportion, due to which the sound of the cheers of the masses can be heard. The dataset was obtained from the kaggle website. Table 1 shows the details of these dataset.

TABLE 1. DATASET MINUTIAE

Input File Data	Name of Dataset	File Format	Sample Rate	No. of Sample	No. of Persons
Voice Recognition	Prominent leadersspeeches	Wave	16khz	7500	5

Received 6/7/2021; Accepted 26l8l2021

DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

B. Feature Extraction

1. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is an audio extraction technique that extracts speaker-particular parameters from speech. (MFCC) is the most popular and common technique of extracting spectral characteristics for the speech [15] by applying the Fast Fourier (FFT) which converts the signal from the time domain to the frequency domain [16]. Dependent Mel spaced Filter Bank (MFB), Discrete Cosine Transform (DCT) is used on the log energy outputs. MFCC is strong, applied by the largest majority of the suggested methods, perfect for both Verification and Identification, and can be mixed with other feature extraction techniques to provide more reliable results [15]. Moreover, all MFCC steps will be indicated as shown in *Fig*. 2.

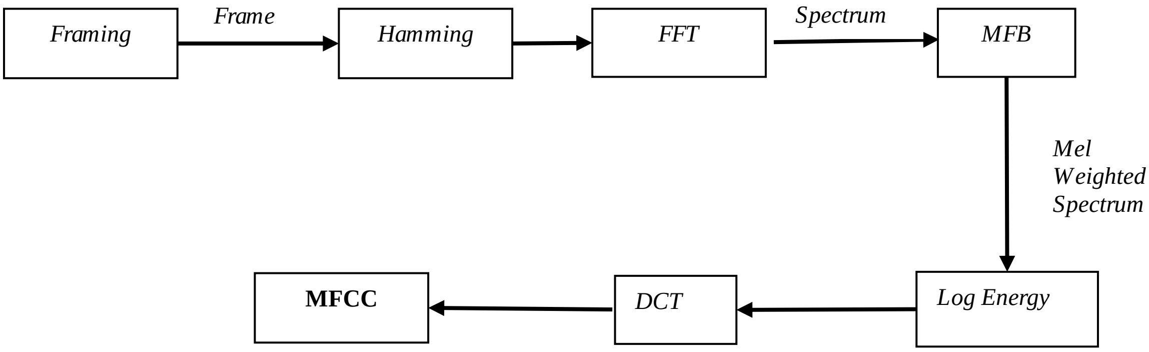


FIG.2. CLARIFICATION OF MFCC PHASES

2. The Four Features (FF)

Amplitude (A) of sound is the volume of sound that is generated. The Mean (M) is the average used to add up all the numbers and then divided by the number of digits. Zero-Cross (ZC) is the rate at which the signal's sign shifts during the frame. Standard Deviation (SD) is a statistic that measures a dataset's distribution of an average. Through measuring each data point's deviation of an average, the standard deviation is measured as the square root of variation.

3. Vector Quantization (VQ)

(VQ) can be described as the method applied to get a big group of feature vectors and created a little collection of the feature vectors, which are describing the distribution's centroids, and we can express it in another way, points spaced to minimize average distance for all one of the other points. The reason for employing VQ is useful to store each one of the feature vectors which are generated from training speech. Though the VQ needs time to calculate, it saves time during the testing stage. The features extraction can be optimized by employed (VQ) in order to quickness the recognition method [17].

4. Quadratic Discriminant Analysis (QDA)

Quadratic discriminant analysis (QDA) is a multivariate, (SML) technique that is employed to cluster anonymous instances in a collection based on recognized training variables. QDA is achieved by maximizing the posterior probability below the assumption that observations support a normal distribution, QDA assumes the most general situation with classes possessing various means and covariance [18]. (DA) proposes to classify new notes into collections identified a priori. QDA, as an expansion of (LDA), assumes multivariate regular distributions of the data, an advantage of QDA is that it is strong to breaches of the assumption of regularity and often better than other classification techniques. QDA needs information of the Gaussian parameters for all class. This can be achieved by

Received 6/7/2021; Accepted 26l8l2021

DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

determining these parameters from the available training points employing the highest likelihood evaluation, a method that should be efficient if the number of training instances is enough high. When the number of training instances is little compared to their fields, maximum likelihood covariance matrix estimates can be defectively conditioned, reaching high misclassification error rates. One common method to determine the ill-posed calculation consists of regularizing the covariance calculation.

5. K- Fold Cross-Validation

To evaluate and confirm the achievement of the (SML) model, resampling techniques are selected. This technique determines the prediction capability of the (SML) classifier on new hidden input data. The k fold cross-validation is one of the resampling methods used in this research to validate the (SML) models on the dataset sample. The ‘k’ describes the number of times the data model is divided. Each division of the dataset sample is called a subsample or sampling group. These sub instances are employed to validate the training dataset. In this research, the ‘k’ value is preferred as 8. Therefore, it can be described 8-fold cross-validation resampling technique. The 8-fold cross-validation method intends to decrease the bias of the prediction system.

IV. PROPOSED SYSTEM EVALUATION

Certain parameters are used to evaluate a model's actions when assessing its output. The findings are influenced by the size of the training data, the consistency of the files, and, most importantly, the type of supervised machine-learning algorithm used. The efficacy of the models is evaluated using the following parameters:

Accuracy (Acc): The percentage of examples correctly classified out of all those given. It is determined as follows [19]:

$$Accuracy = \frac{p+q}{p+q+f+g} \tag{1}$$

Where

p = True positives are the number of instances that were predictable to be positive but mutated out to be positive.

f = False positives the number of instances that were predictable to be positive but mutated out to be negative.

q =True negatives are the number of instances that were predictable to be negative but mutated out to be negative.

g=False negatives the number of instances that were predictable to be negative but mutated out to be positive.

Precision (Pre): For all those classified as class x, the percentage of true x-class instances. It is determined as follows [20]:

$$\text{Precision} = \frac{\text{p}}{\text{p+f}} \tag{2}$$

Recall (Rc): The proportion of examples classified as class x out of all examples classified as class x. It is determined as follows [20]:

$$Recall = \frac{p}{p+q} \tag{3}$$

F- measure: precision and recall have a harmonic mean. It's worked out as follows [20]

$$F_1 = 2 * \frac{Pre*Rc}{Pre+Rc} \tag{4}$$

Error Rate (ER): An error is basically a misclassification, a case is presented to the classifier, and it incorrectly classifies the case, as shown in Eq. (5) below [20]:

Received 6/7/2021; Accepted 26l8l2021

$$Error\ Rate\ = \frac{f+g}{p+q+f+g} \qquad (5)$$

Specificity (NTR): The tendency of a test to be negative when the condition is not present is measured. It is also referred to as the false-positive rate, accuracy, Type I error, error, commission error, or null hypothesis [20].

$$Specificity(NTR) = \frac{q}{p+q} \qquad (6)$$

Mean Absolute Error (MAE) and RMSE (Root Mean Square Error): These measures are commonly used to assess the accuracy of a recommender system, and are calculated as shown in Eq. (7) and (8) [20]:

$\text{MAE} = \sum |r^n - rn|\ N\ n{-}1\ /N \ldots$ (7)

$\text{RMSE} = \sqrt{}\sum (r^n - rn)\ \text{N 2 n}{-}1\ /\text{N}$ (8)

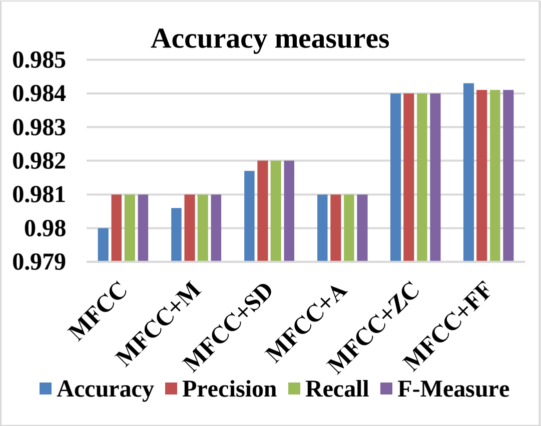
Where, r^n means the expectation rating; rn means the true rating in testing data set; N is the number of rating expectation pairs among the testing data and expectation result.

V. EXPERIENTIAL FINDINGS

The QDA technique is used to estimate our datasets in this research. Table 1 shows the voice dataset information. Table 2 and Figs. numbered (3, 4) illustrate the results of accuracy and error measurements by implementing a classifier (QDA) when used with different feature extraction methods.

TABLE 2. RESULTS OF QDA CLASSIFIER

Features Extraction	QDA							
	Accuracy	Precision	Recall	F-Measure	NTR	ER	MAE	RMSE
MFCC	0.98	0.981	0.981	0.981	0.995	0.02	0.012	0.08
MFCC+M	0.9806	0.981	0.981	0.981	0.995	0.0194	0.011	0.078
MFCC+SD	0.9817	0.982	0.982	0.982	0.995	0.0183	0.011	0.077
MFCC+A	0.981	0.981	0.981	0.981	0.995	0.019	0.012	0.08
MFCC+ZC	0.984	0.984	0.984	0.984	0.996	0.016	0.009	0.073
MFCC+FF	0.9843	0.9841	0.9841	0.9841	0.9961	0.0157	0.008	0.07



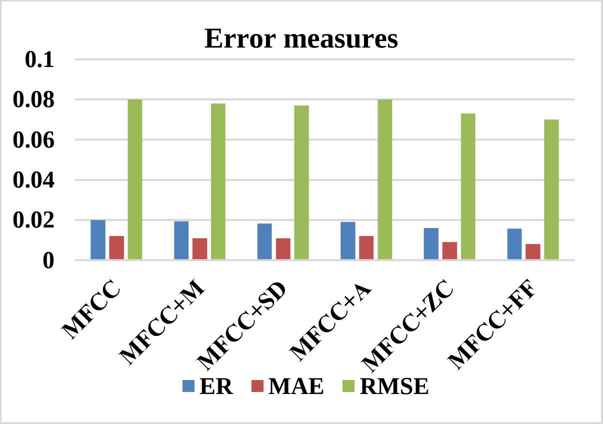


FIG.3. ACCURACY MEASURES FOR CLASSIFIER

FIG.4. ACCURACY MEASURES FOR CLASSIFIER

Received 6/7/2021; Accepted 26l8l2021

DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

VI. COMPARISON RESULTS WITH RELATED WORK

Table 3 compares the propose QDA to the relevant work, in terms of feature extraction method and the approved classification technique.

TABLE 3. RESULTS AND TECHNIQUES COMPARISON

Reference Number	Year	Features Extraction Method	Technique Used	Accuracy
Ref.[9]	2018	MPNCCs	GMM-UBM and SVM	97.5%
Ref.[10]	2018	MFCC+GMM	MLE	93%
Ref. [11]	2019	MFCC+LPC+ZCR	SVM	79.1%
			ANN	93.1%
Ref. [12]	2020	MFCC	MLP	93.33%
Ref. [13]	2020	MFCC+DD	KNN	68.1%
Ref. [14]	2021	MFCC+RPC	RF	97%
Proposed QDA	2021	MFCC+FF	QDA	98.43%

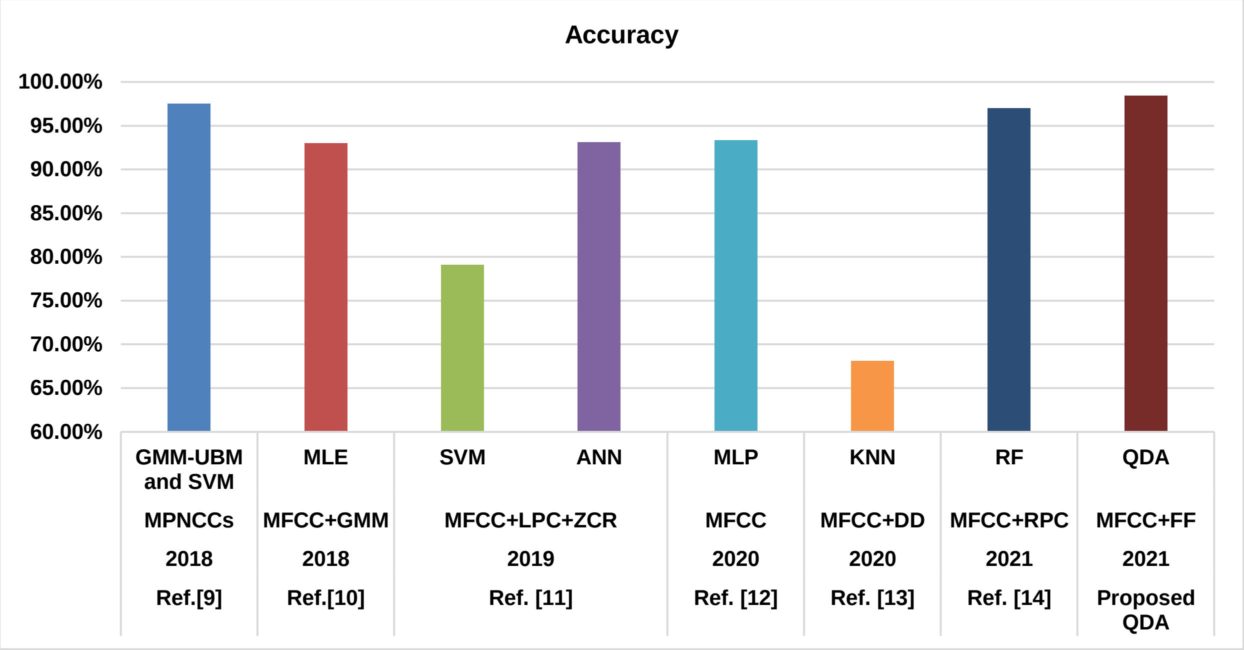


FIG.5. COMPARISON RESULTS WITH RELATED WORK

The comparison results show that the proposed QDA outperformed the overall literature survey techniques in terms of accuracy of speaker recognition using the QDA algorithm, illustrate in Table 3 and *Fig.* 5.

VII. CONCLUSIONS

The biometric features of humans are a crucial factor in identifying a person, and one of the most important of these features is the voice. The method of extracting features (MFCC) is an important and effective factor in extracting voice features. In addition, the strength and reliability of (MFCC) increased by adding features (FF) as the results of accuracy by employing a (QDA) classifier reached

Received 6/7/2021; Accepted 26l8l2021

DOI: https://doi.org/10.33103/uot.ijccce.21.4.8

98.43%. The evaluation method (K-fold cross-validation) is an effective way to validate the results, specifically when choosing a value (K) equal to 8.

REFERENCES

[1] R.K.Hasoun, and R.S. Ali. “A COMPREHENSIVE REVIEW ON IRIS RECOGNITION METHODS”, Iraqi Journal for Computers and Informatics 46, no. 2, pp. 12-16, (2020).

[2] K.J.Taher, and H.D. Majeed. “Recognition of Handwritten English Numerals Based on Combining Structural and Statistical”, Iraqi Journal for Computers and Informatics 22, no. 1,pp. 73-83, (2021).

[3] J.Salman, T.R,Saeed, and A.H. Ali. “Improve the Recognition of Spoken Arabic Letter Based on Statistical Features”, Iraqi Journal for Computers and Informatics 18, no. 3, pp. 26-32 (2018).

[4] H.Nurdiyanto, H. Kurniawan, and S. Karnila. “Human Voice Recognition Using Artificial Neural Networks”, Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 9, pp. 1070-1077, (2021).

[5] S. Joshi, A. Kumari, P. Pai, S. Sangaonkar, and M. D'Souza. "Voice recognition system”, Journal for Research 3, no. 01, pp. 6-9, (2017).

[6] Y. A. Ibrahim, J.C. Odiketa, and T. S. Ibiyemi. “Preprocessing technique in automatic speech recognition for human computer interaction: an overview”, Annals. Computer Science Series 15, no. 1, pp. 186-191, (2017).

[7] M. Malik, M.K. Malik, K. Mehmood, and I. Makhdoom. “Automatic speech recognition: a survey”, Multimedia Tools and Applications 80, no. 6, pp. 9411-9457, (2021).

[8] T.MF.Taha, A.Adeel, and A.Hussain. “A survey on techniques for enhancing speech." International Journal of Computer Applications 179.17, pp. 1-14 (2018).

[9] M.S Athulya., and P. S. Sathidevi. “Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers." Digital Investigation 25, pp. 70-77 (2018).

[10] N. Chauhan, T. Isshiki, and D. Li. “Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database”, In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), pp. 130-133. IEEE, 2019.

[11] A. Herrera-Camacho, A. Zúñiga-Sainos, G. Sierra-Martínez, J. Trangol-Curipe, M. Mota-Montoya, and A. Jarquín-Casas. “Design and Testing of a Corpus for Forensic Speaker Recognition Using MFCC, GMM and MLE”, In Proceedings of the 2019 International Conference on Video, Signal and Image Processing, pp. 105-110. 2019.

[12] K.J. Devi, N. H. Singh, and K. Thongam. “Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network”, Microprocessors and Microsystems 79 (2020): 103264.

[13] M.S. Rao, Y. Haritha, and A. H.V. Reddy. “AUTOMATIC SPEAKER RECOGNITION SYSTEM USING K-NEASREST NEIGHBOR ALGORITHM”, Journal of Social Sciences, Vol 11, pp 576-583. Issue 4, April/2020.

[14] K. Nawas, M. K. Barik, and A. N. Khan. “Speaker Recognition using Random Forest”, In ITM Web of Conferences, vol. 37, p. 01022. EDP Sciences, 2021.

[15] S. A. Kadhum, A. B. Muslim, and A. Y. Al-Sultan. “Survey of Features Extraction and Classification Techniques for Speaker Identification”, Journal of University of Babylon for Pure and Applied Sciences, pp. 43-54, (2020).

[16] L. A. Al-Hashime, S. M. Abdul Satar, and G. A. AL-Suhail. “PAPR Reduction in Coherent Optical OFDM System using Modified Sliding Norm Transformer”, Iraqi Journal Of Computers, Communications, Control And Systems Engineering 18, no. 3, pp. 33-42, (2018).

[17] A. Bejaoui, K. Elkhalil, A. Kammoun, M. S.Alouni, and T.Al-Naffouri. “Improved design of quadratic discriminant analysis classifier in unbalanced settings”, arXiv preprint arXiv:2006.06355 (2020).

[18] Ling, Shaoshi, and Yuzong Liu. “Decoar 2.0: Deep contextualized acoustic representations with vector quantization”, arXiv preprint arXiv:2012.06659, pp. 1-10 (2020).

[19] A. Albayati, Q. Abdulhakeem, and S. H. Ameen. “A Method of Deep Learning Tackles Sentiment Analysis Problem in Arabic Texts”, IRAQI JOURNAL OF COMPUTERS, COMMUNICATIONS, CONTROL AND SYSTEMS ENGINEERING 20, no. 4, pp. 9-20, (2020).

[20] A.T. Ali, H.S. Abdullah, and M. N. Fadhil. “Voice recognition system using machine learning techniques”, Materials Today: Proceedings (2021).

Received 6/7/2021; Accepted 26l8l2021