Speaker Identification using Multi Methods of Features Extraction

[1]Dr. Khadim Mahdi Hashim ; [2]Hasan karam

[1,2]Thi-Qar University College of Education for pure science Department of Computer
e-mail: Kadhimmehdi63@utq.Edu.Iq; e-mail: hsnkrmd20@gmail.com

Abstract: The primary challenge in identifying speakers is extracting recognition features from speech signals to optimize classification algorithms' performance. Several methods are proposed in this article for extracting differential features from an audio signal in order to classify the speaker. The following methods were used to obtain the features of the audio signal in this approach: Power Spectral Density (PSD), Short Term Energy (STE), Fast Fourier Transform (FFT), Hue Seven Moment Invariants method (HSMI), Mel Frequency Cepstrum Coefficients (MFCC), cross-correlation estimates of MFCC (XCORR), and Linear predictive coding (LPC). The classification methods in this paper are the artificial neural network (ANN), the Euclidean distance, and the autocorrelation, where the results obtained from the experiments showed that the accuracy rate is more than 96%.

Index Terms : Speaker identification, pattern recognition, FFT, STE, LPC, Hue seven moments, MFCC, cross-correlation, ANN, Euclid's distance, and autocorrelation

1. Introduction

Speaker identification is a technique for recognizing the speaker from a given expression by analyzing the voice biometrics of the utterance and previously collected expression models[1]. It is also known as Speaker Recognition, which automatically analyzes and detects who is speaking based on the unique details found in sound waves and verifying individuals' identities. Speaker identification requires the extraction of information to recognize a speaker [2].

For Speaker identification, these studies employ feature-based frequency slope coefficients (MFCCs), Gaussian mixture models (GMM) [3], and vector quantization[4]. Following that, basic machine learning classifiers are fed these features[5]. MFCC, cepstral coefficient of linear prediction (LPCC), energy-dependent cepstral coefficient, spectral characteristics, and time-domain properties have been proposed as feature extraction techniques. However, in complex datasets, the above features are inadequate for speaker identification and have poor classification accuracy [6]. The MFCC and LPCC rating outputs are degraded by channel fluctuations induced by ambient noise and magnetic interference in phones or microphones[7]. Artificial intelligence is increasingly expanding and involves designing and implementing many applications such as speech recognition, decision-making, Facial Recognition, etc. Voice biometrics has recently been used to authenticate a person's identity since the human voice is the most practical mode of communication due to its versatility, individuality, and universality. Compared to other biometric identification methods, the advantages of hearing aid recognition include simple access to speech, ease of use, low and easy-to-obtain costs, and comparatively more explicit Recognition of individuals [8].

The proposed work in this paper is the incorporation of features derived from the speech

input signal using various techniques. These methods were used with specific procedures. These methods are as follows: PSD, STE, FFT, Hue Seven Moments, XCORR for MFCC ANN, Euclidean distance, and autocorrelation were used to classify the attribute vector obtained. Moreover, the vector-based speaker identification system containing all the features of the above methods, which was built on an actual database of 13 speakers of different ages and genders, was evaluated due to the strength of the features obtained through the proposed method, in addition to To the dependence of the neural network on reducing the error using the mean square error computation function, the results obtained using the neural network are more reliable than those obtained using similarity measurement functions such as autocorrelation and Euclidean distance.

Related Works

Feature extraction is vital within the speaker recognition process because it impacts the efficiency of the speaker recognition classification model. Several experts within the area of speaker identification have proposed new features in recent years that have proved helpful in successfully classifying human voices. A speech recognition system must be ready to withstand environmental variations like transmission channels, speaker inconsistencies, and noise. Researchers benefit from filtering the signal with some ways to process the speech signals to get a simple classification idea behind noise reduction is to make a filter that will effectively eliminate noise while retaining valuable information. As a result, several researchers are investigating how to reduce the impact of noise to classify speech signals correctly.

Fong et al. [9] conducted an alternative study to classify speakers with various statistical features of time-domain and machine learning classifiers, which achieved the highest accuracy of the multi-layer perceptron classification of about 94%. While the experiment results completed a high degree of accuracy, the results cannot be further applied. The authors only used 16 speakers from the PDA language data set in the experiment. In addition, the report used some speakers' speeches in the training tests.

Ali et al. [10] suggested a model for recognizing the speakers in the Urdu language to classify ten individual speakers. In this study, deep learning and MFCC functions were fused to organize speakers using an SVM algorithm. The test results obtained a rating accuracy of 92 percent. The results are, therefore, promising. However, the data set used in the experiments has some flaws. For starters, the tests only used ten-speaker utterances. Second, each statement contained the authors propose only one term. Thus, ineffective and inefficient fusion characteristics for complex human voices.

Soley manpower et al. [11] have investigated MFCC clustering functions combined with an ANN classification to categorize 22 ELDSR speakers. 93% classification accuracy was obtained by the experimental findings.

II.Proposed Technique

As illustrated in Figure (1), this section details the proposed Technique for recognizing speakers, including the following: First, the experiments' various speaker data were gathered. Second, the proposed methods extract and group useful features into vectors, which are then labeled and used as inputs to the ANN to construct a trained network model capable of identifying unknown speech signals. Finally, the accuracy and similarity scales are used to measure the efficiency of the compound model; the details of these methods are discussed in

the following sections.
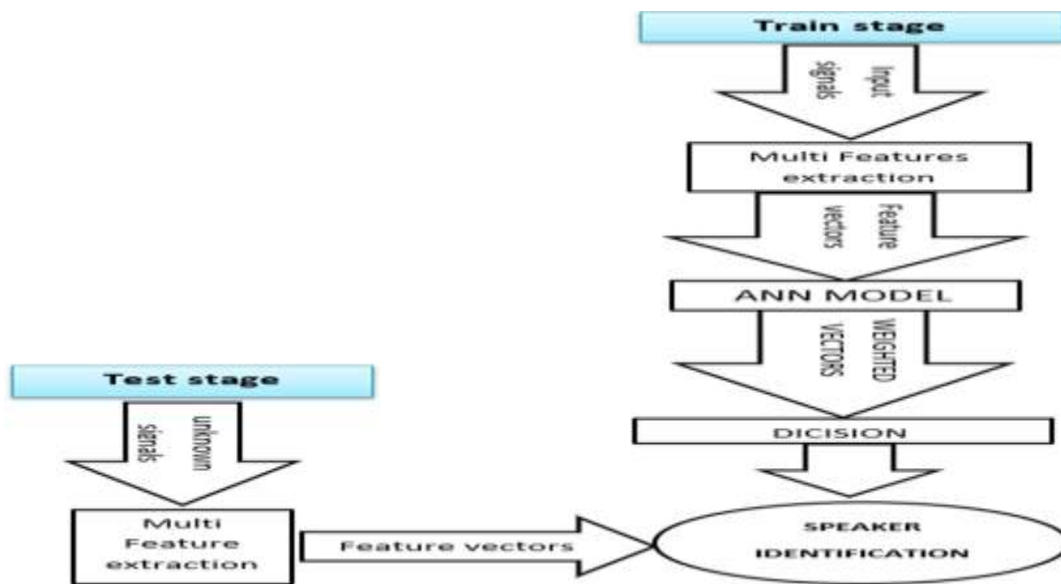
Figure (1) The general outline of the proposed Technique

A. Dataset

An actual data set was used, as it was collected from the speech recording of 13 Arabic speaking subjects, and each person scored 20 samples. A person says a sentence of varying lengths ranging from 3 to 15 seconds in. Each sentence is repeated twice, recorded according to the specifications mentioned in Table No. (1).

Table (1) showing the voice recorder specifications

speakers	mal	female	Total size	Platform Support	NumChannels	SampleRate	BitsPerSample
13	6	7	49,696,768 bytes	WAVE (.wav)	1	16000 Hz	16

Where NumChannels is the Number of audio channels encoded in the audio file. SampleRate is the Sample rate of the audio data in the file in hertz. BitsPerSample is the Number of bits per sample encoded in the audio file.

B. Speech preprocessing

The bandwidth filter is used according to the settings in the default Sound Recorder Pro version 3.7.2.0 mobile app. The signal is processed with a bandpass filter path waveform compensating for filter delay with a low-rank filter.

C. Features Extract

Extracting features is more complex than characterizing classification because they are more industry-specific than the general essence of taxonomic customization. Therefore, the

proposed method for extracting the valuable features is implemented in this study, as it was designed by adopting different methods, which are as follows: FFT, STE, LPC, Hue seven minutes, MFCC, Cross-Correlation, and LPC. The MATLAB program was used to implement the proposed method. The use of all suggested methods of feature extraction will be detailed below:

1-Power Spectral Density (PSD)

the power spectral density (PSD) is the spectral energy distribution observed per unit of time. Summation or integration of the spectral elements, according to Parseval's theorem, yields total power (for a physical process) or variance (for a statistical process) [12]. Using Matlab software, calculate each PSD estimate by equal noise bandwidth (Hz) of the window to obtain an estimate of the power at each frequency. Then the arithmetic mean was calculated as an independent feature, as shown by the algorithm (1).

Algorithm (1): Mean of PSD Algorithm

Begin
Step1: Read data from audio files in vectors.
Step2: Calculate the length of that vector.
Step3: Apply the hanning function type periodic on the vector as eq. 1.
Step4: Find periodogram power.
Step5: Normalized the power vector
Step 6: Calculate the mean of the vector X.
End

---

2 - Fast Fourier Transform (FFT)Method

The Fast Fourier transform divides the original time-based waveform into a series of sinusoidal terms, each with a unique volume, frequency, and phase where this process creates an energy spectrum and is the result of the initial waveform response in the frequency domain, and so on, using Fourier transforms to find the frequency components of a signal buried in the noise [13]. By the algorithm (2), the maximum value of the envelope constructed from the maximum values of the Fourier spectrum of the signal was used to feature this study.

Algorithm (2): Max of FFT Algorithm

Begin
Step1: Read data from audio files in vectors.
Step2: Calculates the fft of the data.
Step3: Find nonzero absolute values[14].
Step4: Find and keep the maximum value of the vector.
End

3- Short Term Energy (STE) Method

The energy of a short speech segment is referred to as short-time energy [35]. Short-time energy is a quick and efficient segmentation parameter for both voiced and unvoiced pieces[15]. Energy is often used to detect utterance endpoints [16]. The short-time energy

of the signal can be determined from the following expression[17]:

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \quad , \quad n - N + 1 \leq m \leq n$$

…………………………..(1)

n=0, 1T, 2T, ..., N is the window length, and T is the frameshift.

Where w (n-m) is the window, n is the sample on which the analysis window is focused, and N is the length of the window. The specified window defines the timer for processing and slices by progressing the values squared. Here high energy is classified as acoustic and low energy as non-noisy. In this study, the mean was calculated from the short power vector of the speech signal as an independent feature of the total energy of the sample and as shown by the algorithm (3).

Algorithm (3) Mean of STE
Begin Step1: Read data from audio files in vectors. Step1: Normalize data. Step2: Framing rectangular data with a duration of 25 seconds and no overlaps. Step4: Calculate the total energy of all the frames in (ste) vector as eq1. Step5: Normalize the energy in the vector [18]. Step6: The mean of the vector is calculated and kept. End

4- Hue Seven Moments Invariants

Hu's definition of momentary constants was first proposed, where Hu extracts six absolute orthogonal variables and one orthogonal deflection in [19] using algebraic constants. The seven moments are the statistical expectation of certain energy functions of a random variable, The most expected moment is the mean which represents only the expected value of a random variable; invariant moments are beneficial in pattern recognition problems and it constant in scaling, rotation, and transformation [20]. In this study, using the MATLAB program, only five of the seven moments suggested by He in [19] were used as valuable features to identify the speaker in the proposed Technique, as shown in the algorithm (4):

$$\phi_1 = \eta_{20} + \eta_{02} \quad \text{.. (6)}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \quad \text{.. (7)}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \mu_{03})^2 \quad \text{...................................... (8)}$$

$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} + n_{03})(\eta_{21} + n_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{12} + \eta_{03})^2] \text{…... (9)}$$

$$\phi_6 = (\eta_{20} - 3\eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + n_{03})$$
…………………………………………………………………..….. (10

Algorithm (4): Seven Moments Invariants
Begin Step1: Read data from audio files in vectors. Step2: Convert the vector into a square matrix. Step3: Apply the fft of a square matrix. Step 4: get the real portion of the matrix. Step5 the maximum values are extracted. Step6: getting five moments [M1 M2 M3 M5 M6]. End

5- MFCC Method

MFCC related features are helpful when identifying speakers [21]; the MFCCs are a compact representation of the audio signal's spectrum. The MFCC coefficients contain information about rate changes in the various spectrum bands; if the cepstral coefficient is positive, most spectral energy is concentrated in the low-frequency regions. On the other hand, if the cepstral coefficient has a negative value, the most spectral power is concentrated at high frequencies. The function $mfcc_m[n]$ as:

$$mfcc_m[n] = \frac{1}{R} \sum \log \left(MF_m[r] \right) \left[\frac{2\pi}{R} (r + \frac{1}{2}) n \right] \quad \text{...............................} \quad (11)$$

Typically, $mfcc_m[n]$ is evaluated for n= 1, 2, ..., $Nmfcc_m[n]$, where N mfcc is less than the number of Mel-filters, e.g., $Nmfcc_m[n]=14$ and R=24.

The MFCC algorithm consists of the following steps: pre-confirmation, framing, cavitation, fast Fourier transform, Mel filter bank processing, and discrete cosine transform [22],[23]. In this study, the MFCC function divides the whole speech signal data into parts, with a window length of 30 ms for each segment, and overlaps between 20 ms. The idea is to get a logarithm vector of energy followed by 13 vectors, then the sum of each vector is taken as an independent feature as shown in the algorithm (5).

Algorithm (5) The Sum of MFCC Coefficients
Step1: Read data from audio files in vectors. Step2: The MFCC is applied to the input signal usi Step3: Keep to the log-energy value follo Step4: Find the sum of each v Step5: Finding the End

6 -Cross-Correlations of MFCC features.

In signal processing, cross-correlation is a measure of similarity of two series as a function of displacement of one relative to the other; the term cross-correlation refers to probability and statistics to correlations between the inputs of two random vectors. The correlations between the various time instances of a random vector over time are called autocorrelations; if X and Y are independent random variables, then the different probability density is given formally

through cross-correlation, the meaning signal processing the skew provides the probability density function of the sum[24],[25]. In this study, automatic correlations and cross-correlation sequences are found for all columns obtained using the MFCC algorithm on speech signal and illustrated by the algorithm (6).

Algorithm (6) XCORR Of MFCC Algorithm

Input: The Matrix obtained in the step2 of the algorithm (5)
Outputs: 196 maximum values in a vector
Begin
Step1: input Matrix M by 14 derived from an algorithm (5) step 4
Step2: Compute XCORR of MFCC matrix.
Step3: Get 196 maximum value.
End

---

7 LPC feature.

One of the essential methods of speech analysis is linear prediction (LP), in which a speech sample can be approximated as a linear series of previous instances, and a particular set of predictor coefficients can be determined by linear reduction of the total squared differences between natural speech samples and predicted samples over a limited period. The LPC uses the Levinson-Durbin iteration to solve the standard equations that derive from the least-squares formulation. The LPC steps are Pre-emphasis, Frame Blocking, Windowing, Autocorrelation, LPC Analysis, and LPC Parameter Conversion to LPC Coefficient using the equations [26]. Using MATLAB, the linear predictor coefficients with the FIR filter of order seven are calculated, and then their sum is calculated in this study, as shown in the algorithm steps (7)

Algorithm (7) sum of LPC Algorithm

Begin:
Step1: Read data from the audio file in vector.
Step2: Obtaining LPC Coefficients with the FIR filter of order seven.
Step3: Sum the seven LPC coefficients.
**End**

D-Speaker Identification

1-   Artificial Neural Network (ANN)

A neural network is a type of information processing device that is used to simulate the human brain. Multiple layers of essential computing elements known as neurons comprise an artificial neural network. Its processing components are a large number of densely connected neurons. Learning occurs when these strengths are adjusted to allow the overall network to produce the desired results. The artificial neural network applications include diagnostic systems, biochemical analysis, image analysis, and drug creation. Artificial neural networks are a perfect way to learn. Learning with the help of an artificial neural network entails training the machine to recognize the input data. Unknown inputs are fed into the neural network to

evaluate its output. The training phase is repeated until the network responds as predicted[27].

_ANN Training

The ANN was built using the Backpropagation algorithm for computing Jacobian vectors, which tests the weight matrix's output and the bias variables and the Levenberg-Marquardt algorithm. The network is divided into two hidden layers, each with 50 "logsig" neurons. After extracting the features from the speech signal using the algorithms as mentioned earlier, they are grouped into a single vector for each sample, which is then arrayed and serves as the input to the neural network, from which the neural network is trained, and from which a trained network capable of classifying and distinguishing the vectors is obtained. Figure 3 illustrates the procedure. Using MATLAB, the vector array will be the input of the neural network created from three primary stages, the first is the input unit, the second is the processing unit, and it is two hidden layers, each one contains fifty neurons, while the third: is the output unit, and it consists of four outputs, And it represents a binary number that in turn indicates the classification of the sample vector, as shown by the algorithm (8)
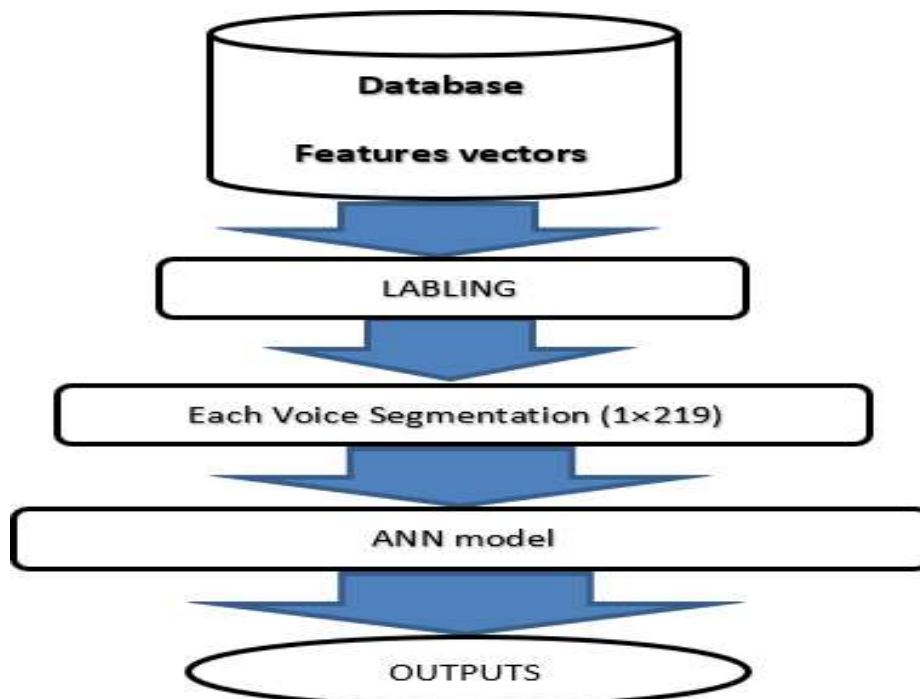
Figure (3) ANN Training for the speaker Identification System

Algorithm (8) ANN Training models Algorithm

Limitations: uses the Jacobian function: Levenberg-Marquardt: Backpropagation algorithm
Input: Features matrix
Outputs: Matrix of a binary number
Begin:
Step1: Input Features matrix
Step2: Initialize the ANN model.
Step3: Classification process using ANN
Step4: the neural network model is kept
End

2- Testing the Classifier

The database in this section was divided into two parts: training and the other for testing, and each of them was divided according to the type of test. The first type was the 20% test, and in that, four samples were tested from each unknown speaker on a trained network on 16 samples, then the second test was of the 70% type where six unknown samples were used on a trained network on 14 sample amplifiers. Figure (4) shows the stages the test sample goes through until a decision is made to determine the speaker's identity. The algorithm (8) describes the overall processes used in this study from the beginning, where the recorded samples are collected, then processed, features extracted, then trained, and the training network obtained. Then the recorded examples are tested for unknown speakers on the acquired network.
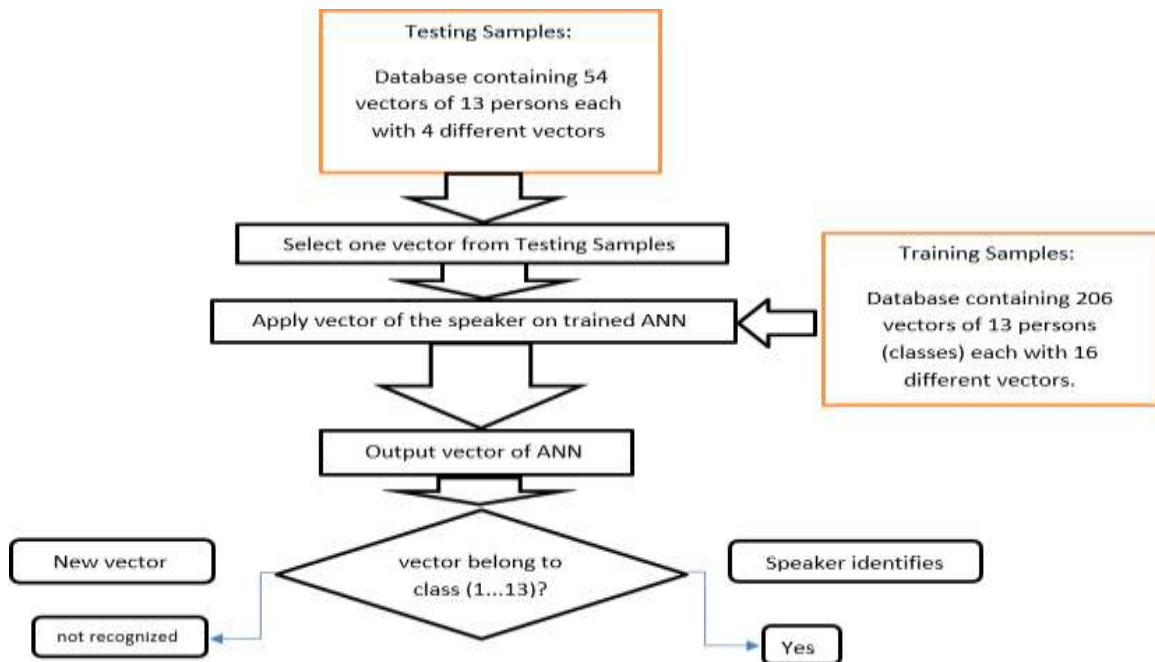

Figure (4) Diagram of Testing the Classifier model

E-EXPERIMENTAL RESULTS

This section presents the results, as the features extracted using the different methods that were explained in the previous sections were raised, and Tables (2) shows the comparison of the results obtained in the ANN test in different proportions where 20%, 30% were tested 40% and 50% on Straight. After that, by comparing it to some measures of similarity, the results were perfect for testing the neural network, where the 20% test reached a success rate of more than 96%, as the network showed its efficiency and technology to predict the result and determine the identity of the speaker, and the results were optimistic, due to the strength of the features extracted by the methods Which were previously detailed in this study. The Tables showing the comparison of the ANN results with some similarity measures too.

The use of ANN to classify speaker identity was effective in this study, with the proposed features achieving an overall accuracy of over 96 percent based on rigorous experimental results. Moreover, ANN was found suitable for identifying speakers using the features of the proposed methods. Experience shows that the proposed speaker identification system is effective, accurate, and robust in renewing the speaker's identity. It is possible to benefit from

the proposed approach in access control, security, and other applications.

When using the Euclid scale, which is one of the most common methods of measuring similarity, the results showed in Table (2) that the more training samples increase, the accuracy of the classification increases, which is a natural behavior and indicates the extent of the rigor and strength of the methods through which the features were extracted from the speech signal.

The Kendall Rank Correlation Coefficient was used, which is a nonparametric measure of correlation. Kendall's tau is based on the number of pairs [58]. According to the results shown in Table (2), the Kendall correlation coefficient gives an impression of the correlation strength of the traits extracted by the proposed methods. It is still shown that the network's behavior in classification is more robust and systematic.

Table (2)

Testing / Methode	TEST 20%	TEST 30%	TEST 40%	TEST 50%
ANN	96.15%	92.3%	85.57%	79.23%
EUCLDIAN	88.46%	80.07%	77.88%	71.54%
CORRELATION	92.31%	93.59%	93.27%	90.%

References

[1]. Boujelben et al.,(2009), ―Robust text-independent speaker identification using hybrid GMM- SVM System‖, International Journal of Digital Content Technology and its Applications, Volume 3,103– 110.

[2]. Verma, G. K.(2011)‖Multi-feature fusion for closed set text-independent speaker identification‖, International conference on information intelligence, systems, technology, and management Springer, 170–179.

[3] A. Maurya, D. Kumar, and R. K. Agarwal, "Speaker recognition for Hindi speech signal using MFCC-GMM approach," Procedia Comput. Sci., vol. 125, pp. 880–887, Jan. 2018.

[4] D. A. Reynolds, T. F. Quartieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process., vol. 10, nos. 1–3, pp. 19–41, Jan. 2000.

[5] W.-C. Chen, C.-T. Hsieh, and C.-H. Hsu, "Robust speaker identification system based on two-stage vector quantization," J. Sci. Eng., vol. 11, no. 4, pp. 357–366, 2008.

[6] D. B. A. Mezghani, S. Z. Boujelbene, and N. Ellouze, "Evaluation of SVM kernels and conventional machine learning algorithms for speaker identification," Int. J. Hybrid Inf. Technol., vol. 3, pp. 23–34, Jul. 2010.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audiobooks," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2015, pp. 5206–5210.

[8] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," PLoS ONE, vol. 11, no. 7, Jul. 2016, Art. No. e0158520.

[9] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," IEEE Signal Process.Lett., vol. 13, no. 1, pp. 52–55,

Jan. 2006.

[10]   S. Fong, K. Lan, and R. Wong, "Classifying human voices using hybrid SFX time-series preprocessing and ensemble feature selection," BioMed Res. Int., vol. 2013, Oct. 2013, Art. No. 720834.

[11]   H. Ali, S. N. Tran, E. Benetos, and A. S. D. A. Garcez, "Speaker recognition with hybrid features from a deep belief network," Neural Comput. Appl., vol. 29, pp. 13–19, Mar. 2018.

[12] M. Soleymanpour and H. Marvi, "Text-independent speaker identification based on a selection of the most similar feature vectors," Int. J. Speech Technol., vol. 20, no. 1, pp. 99–108, Mar. 2017.

[13] STOICA, Petre, et al. Spectral analysis of signals. 2005.

[14] VAN LOAN, Charles. Computational frameworks for the fast Fourier transform. Society for Industrial and Applied Mathematics, 1992.

[15] PROAKIS, John G. Digital signal processing: principles algorithms and applications. Pearson Education India, 2001.

[16] YANG, Xiaoling, et al. Comparative study on voice activity detection algorithm. In: 2010 International Conference on Electrical and Control Engineering. IEEE, 2010. p. 599-602.

[17] ENQING, Dong, et al. Voice activity detection based on short-time energy and noise spectrum adaptation. In: 6th International Conference on Signal Processing, 2002. IEEE, 2002. p. 464-467.

[18] https://drive.google.com/file/d/0B3qx_fO_3y2AaHdpZHVzYURGclk/view

[19] HU, Ming-Kuei. Visual pattern recognition by moment invariants. IRE transactions on information theory, 1962, 8.2: 179-187.

[20] HUANG, Zhihu; LENG, Jinsong. Analysis of Hu's moment invariants on image scaling and rotation. In: 2010 2nd International Conference on Computer Engineering and Technology. IEEE, 2010. p. V7-476-V7-480.

[21] TIRUMALA, Sreenivas Sremath, et al. Speaker identification features extraction methods: A systematic review. Expert Systems with Applications, 2017, 90: 250-271.

[22] Rabiner, Lawrence R., and Ronald W. Schafer. Theory and Applications of Digital Speech Processing. Upper Saddle River, NJ: Pearson, 2010.

[23]   Auditory   Toolbox.   https://engineering.purdue.edu/~malcolm/interval/1998-010/AuditoryToolboxTechReport.pdf

[24] MOUNCE, S. R., et al. Predicting combined sewer overflows chamber depth using artificial neural networks with rainfall radar data. Water science and technology, 2014, 69.6: 1326-1333.

[25] GUPTA, Anjalika; RAIBAGKAR, Pankaj; PALSOKAR, Anup. Speech Recognition Using Correlation Technique. International Journal of Current Trends in Engineering & Research (IJCTER) e-ISSN, 2017, 2455-1392.

[26] DENG, Li; O'SHAUGHNESSY, Douglas. Speech processing: a dynamic and optimization-oriented approach. CRC Press, 2018.

[27] ZUPAN, Jure. Introduction to artificial neural network (ANN) methods: what they are and how to use them. Acta Chimica Slovenica, 1994, 41: 327-327.