Speaker Identification based on Hybrid Feature Extraction Techniques

Feras E. Abualadas[1]
Computer Science
International Islamic University Malaysia
Ajloun National University, Jerash, Jordan

Akram M. Zeki[2]
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Kuala Lumpur, Malaysia

Muzhir Shaban Al-Ani[3]
College of Science and Technology- Computer Science
University of Human Development
Sulaymaniyah-KRG, Iraq

Az-Eddine Messikh[4]
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Kuala Lumpur, Malaysia

***Abstract*—One of the most exciting areas of signal processing is speech processing; speech contains many features or characteristics that can discriminate the identity of the person. The human voice is considered one of the important biometric characteristics that can be used for person identification. This work is concerned with studying the effect of appropriate extracted features from various levels of discrete wavelet transformation (DWT) and the concatenation of two techniques (discrete wavelet and curvelet transform) and study the effect of reducing the number of features by using principal component analysis (PCA) on speaker identification. Backpropagation (BP) neural network was also introduced as a classifier.**

***Keywords*—Speaker identification; biometrics; speaker verification; speaker recognition; text-independent; text-dependent**

I. Introduction

A biometric system is considered one of the most important patterns recognition that authenticates a person based on features extracted from physiological or behavioral characteristics [1].

Biometric identification method is preferred in comparison with conventional identification methods that contain passwords for different reasons; the speaker to be identified is needed to be physically present at the point-of-identification [2]. The identification based on biometric techniques does not require to carry a token, a smartcard and remember a password [3].

Human voice is one of the biological characteristics used to distinguish a person from his/her voice, thus we indicate to voice recognition systems [4]. Speaker recognition system is a process that used individuals sound to recognize/discriminate purposes, where it differs from speech recognition since it is concerned with the identity of a person while speech recognition is concerned with recognizing the word [4].

Speaker recognition systems have many applications for security purpose such as keys or passwords and database access [5].

Automatic Speaker recognition can be divided basically into two types: speaker identification (SI) and speaker verification (SV) [5].

Speaker verification is the task of verifying the identity of speakers based on information that contains in the speech signal to make sure that the person is the one who claimed [6]. Basic structure related to the speaker verification as shown in (Fig. 1).

On the other hand, speaker identification refer to the task that is interested in finding identity of the anonymous speakers by one-to-many (1: n) comparisons, where the speaker's voice is compared to the voice of speakers listed in a database, in which basic structure is concerned with speaker identification as explained in (Fig. 2). While in speaker verification the comparison is one-to-one (1:1) and a person is authenticated if it is the one who claims to be [7].

Speaker recognition system can be a text-independent and text-dependent system, depending on the speech used by a system. Text dependent systems are those systems that have prior knowledge of the text to be spoken where the same text is used in (training, testing) phase [8].

While in a text-independent system, recognition system does not possess previous knowledge concerned with spoken text (the text system is unconditional with the used text) [9].

This work concerns on the problem of speaker identification; we proposed a speaker identification system, which deals with defining the speaker's identity based on features extraction (discrete wavelet transformation and (curvelet) including principal component analysis (PCA). For speaker identification, various recognizers can be used such as Hidden Markov Models (HMM), Random Forest (RF), Self-Organizing Map (SOM), statistical approaches, etc. In this research Backpropagation (BP) neural network is used as a classifier.

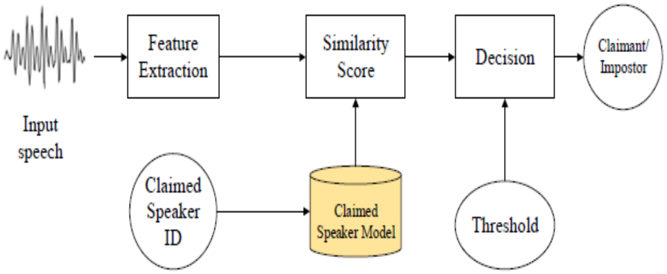

Fig. 1. The basic Structure Related to Speaker Verification.

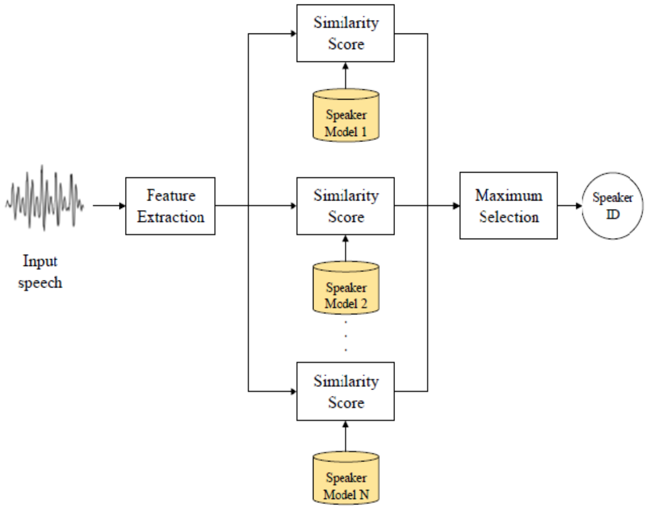

Fig. 2. The basic Structure Related to Speaker Identification.

The rest of the paper is organized as: the next section describes some of the researches in the last few years related to speaker recognition. In Section 3 presents the design and implementation of the developed system. Section 4 presents the result for speaker identification. Finally, paper concludes in Section 5.

II. Related Work

There have been many studies and researches on speaker recognition, where some strategies have been suggested in the last few years.

For speaker recognition systems, these techniques can achieve high performances. Generally, some of these researches were summarized below with different techniques.

They presented a feature extraction method that depends on wavelet analysis for speaker identification system (SIS). Two Techniques combined with each other (Stationary Wavelet Transform (SWT) and Mel-Frequency Cepstral Coefficient) are used to overcome the discrete wavelet Transform drawbacks, these features are used as an input for a classifier, the data set consists of 200 speakers, where K-nearest neighbors (Knn) are used as a classifier. The experimental result showed that suggestion approach achieved better performance rate by using (SWT), where the drawbacks of DWT reduced [10].

They study a speaker identification system, where two systems designed and compared in terms of (computational time and gender and identification rate). A Mel-frequency Cepstral Coefficients (MFCC) and bark frequency Cepstral coefficient (BFCC) as feature extraction are used with Gaussian Mixture Model (GMM), where the impact of filter coefficients number and speaker number are investigated. The experimental results showed that when the number of speaker and coefficient are increased the time of computational increases and the results show that use of MFCC with GMM is better than GMM based on BFCC [11].

They compared different features for text-dependent speaker recognition, where they used wavelet transforms under stressed condition; these conditions have been adopted from SUSAS database, Question, Neutral, Lombard and Anger. Vector quantization is used as a classification method with wavelet. Experimental results showed that Linear Predictive Cepstral Coefficients (LPCC) provide the best result among many features such as, Linear Frequency Cepstral Coefficients (LFCC), ARC, Log Area Ratio (LAR), CEP and Male Frequency Cepstral Coefficients (MFCC), where the improvement achieved in Neutral and Lombard case to 93% and 94% [12].

They study a Linear Predictive Cepstral Coefficients (LPCC) and Mel-frequency Cepstral Coefficient (MFCC) methods independently for speaker identification system and proposed a new feature based on the concatenation Linear Predictive Cepstral Coefficients and Mel-frequency Cepstral Coefficient (LMACC), where each of them recorded in clean and noisy environment based on multi-layer perceptron (MLP) neural network as classifier. The experimental result showed that concatenation of both features LMACC-MLP achieved height performance recognition rate in comparative with each method independently reach to 85% [13].

Authors have presented an approach for speaker identification based on fusion via samples and statistical approach. The data set collected by recorded from 20 male speakers of 5 samples each of 7 words, these samples are passing through a preprocessing phase which includes resizing, noise removal and windowing to be adequate for processing. Features vectors defined each speaker was generated by employing a statistical approach after principal component analysis (PCA) for feature extraction is used for performance evaluation. Features vector was partitioned into overlapping segments of feature vectors, where the percentage of the correct identified segments over all tested segment was used to calculate the performance. The Experimental results showed that this approach obtained a good result of recognition reaching 95% similarity [14]. developed a model for a text-independent speaker identification to obtain a features vector without losing information based on Mel Frequency Cepstral Coefficients (MFCC) with Vector Quantization (VQ), for speaker recognition a Gaussian Mixture Model (GMM) is adopted. To increase the efficiency of feature extraction the signal is passing through the preprocessing phase, which includes Pre-emphasis, silence removal and downsampling. To reduce the number of speaker during a test stage, gender detection algorithm is used. Experimental results show that the suggested algorithm reduced the time testing to almost half 0.l0 51sec and gives 91% accuracy in comparison with VQ and GMM gives 88% and take 0.2242sec [15].

They presented an automatic speaker identification system (SID) based on Gaussian Mixture Model and Support Vector Machines (GMM-SVM), where data set consist of 360 speakers. Each one has 10 sentences adopted from TIMIT phone labeled database corpus. The extracted Features, Mean Hilbert Envelope Coefficients (MHEC) and Gammatone Frequency Cepstral Coefficients (GFCC) are modeled by Gaussian Mixture Model (GMM). Support Vector Machine (SVM) was used to train the corresponding super vectors. Experimental results showed that MHEC features are better in comparison with RASTA-MFCC and GFCC features in different noisy conditions [16].

Authors have used Mel Frequency Cepstral Coefficient (MFCC) as a feature extraction technique to study the performance of speaker recognition system in noisy environments, where noise is considered as one of the factors that affect the sound of a person. Three different techniques Back-Propagation Neural Network (BPNN), Euclidean distance and Self Organizing Map (SOM) are used as classifiers with different windowing, Hamming window, and Blackman window. Experimental results showed that SOM gives better performance in comparison with BPPNN and Euclidean distance [17].

They study and compare different techniques for feature extraction and feature classification to get an optimal choice for automatic speaker recognition system (ASR). The experimental result showed that Mel frequency cepstral coefficients (MFCC) are preferred in comparison with Linear Predictive Cepstral Coefficients (LPCC), Linear Predictive Coefficients (LPC) and Wavelet decomposition techniques and Gaussian Mixture Model (GMM) gives better accuracy and less memory usage for feature classification in comparison with Vector Quantization (VQ), Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) techniques [18].

Authors have presented a method based on Mel-Frequency Cepstral Coefficients (MFCCs) and Discrete Wavelet Transform (DWT) to design a speaker identification system that minimizes the probability of identification errors. For noisy speech signals (MFCC) based feature extraction with additive white Gaussian noise (AWGN), and to enhance the representation of signal features extracted from DWT vector synthesized features in MFCC. Vector Quantization using the Linde-Buzo-Gray (VQLBG) with MFCC is used to enhance the MFCC performances and to recognize the noisy speech. Experimental results showed that the proposed technique in comparison with MFCC gives a better performance where the use of DWT of degraded signal obtains more features and reduces the noise effect when dealing with signals like AWGN, this leads to higher identification rates and improves the recognition rate [19].

They Developed an automatic speaker recognition based on discrete wavelet transformation (DWT) for feature extraction with Back Propagation Network as classifier, where features extracted from approximation and detailed coefficients. Experimental results showed that wavelet is Appropriated for feature extraction where Discrete Meyer wavelet provides higher inter-class variance and lesser intra-class variance in comparison with various wavelets such as Haar, Symlet and Reverse Biorthogonal [20].

Feature extraction technique Mel Frequency Cepstral Coefficient (MFCC), Dynamic Mel-Frequency Cepstral Coefficient (DMFCC) and the Plural between these two features are used to evaluate the performance of text-independent, multilingual speaker identification using Gaussian Mixture Model (GMM) as a classifier. The data set created consists of 120 speakers; each one has five sessions recorded for 20 seconds with a 16 KHz sampling rate using Gold Wave software in English and Tamil languages. The system performance was tested using a different length of segment. The experimental result showed that combination between DMFCC and MFCC features achieved better performance rate in comparison with using each one individually, where the Error Rate obtained for MFCC, DMFCC and (MFCC+ DMFCC) is 5.8%, 2.9% and 1.2% with MFCC respectively [21].

The effect of combining the features extracted from Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC) in comparison with using features individually for speaker identification system in case of cross, mono and multilingual. To do that, data of 30 speakers were created. Each one recorded his/her voice in three different languages (English, Hindi, and Canada) languages. The number of speakers identified by MFCC is 18 and 20 speakers by LPC while the number of speakers with a combination of features (MFCC and LPC) is 22. This concludes that using MFCC and LPC features combined instead of using (one at a time) improves the speaker identification performance about 30% for created dataset [22].

This work concerns with sound recognition techniques to identify individuals speakers. Each human has a singular feature in his sound; it is helpful to distinguish between person and another one using their own sound. The concept of voice recognition that is completely unlike speech recognition is to identify the person speaker versus a store sound pattern, to not perceive what's being aforesaid. Within the domain of sound recognition, several ways are developed like Neural Network, Hidden Markov Models, Genetic algorithms and Fuzzy logic.

III. Methodology of Developed System

This research focuses on creating the speaker identification system and then assessing the system performance and capabilities in speaker recognizing area of research where Back Propagation Neural Network is used to develop the system.

The determination of individuals who speak is the main attribute of any speaker recognition system, where it consists of different modules in an addendum to the classification engine. In this research, we proposed a system that includes (4 Modules) and these Modules are depicted in Fig. 3.

First, the resulted voice after processing goes over to the feature extraction module to extract features that are used to construct the dataset, where the resulted features passed to selection module include principal analysis component (PCA).

Finally, the extracted features resulting from two modules (feature extraction & feature selection) will be passed to the final module, which is the recognition module, with separated forms. The testing and training phases together compose the recognition module, where the system is used to discriminate a speaker sound after it trained.

The dataset (sounds) as an input, where these sounds will pass through sample resizing operation during preprocessing module because there is no ability to control the amount of sample sounds during the recording process, wherein this work the number of samples was resized to 40000.

After reading the sound file, discrete wavelet transformation utilized then each voice fed to the NN. The speaker recognition module is called after extracting features, where these features represent the (coefficients) of different DWT, where BP Neural Network was designed to use the features extracted from singly and combined to train the designed classifiers. Two various datasets were formed are involved: a dataset of discrete wavelet transformation (DWT) features only, and a dataset of (DWT + Curvelet). Fig. 4 shows the classification operation, where it consists of two phases mainly, (training and testing) phase.

Initially, a group of patterns that represent discrete wavelet coefficients that was extracted from three various levels of wavelet are used to train the classifier at the first time. To divide space of feature in a method that allows maximizing the ability of recognition for Neural Network. To build appropriate weight vectors that can classify the training set in correctly within defined some error rate. The trained classifier: that uses these weight vectors that result from training phase are used for appoint the unknown input pattern to one of the class (speaker) depending on the feature vector that was extracted.

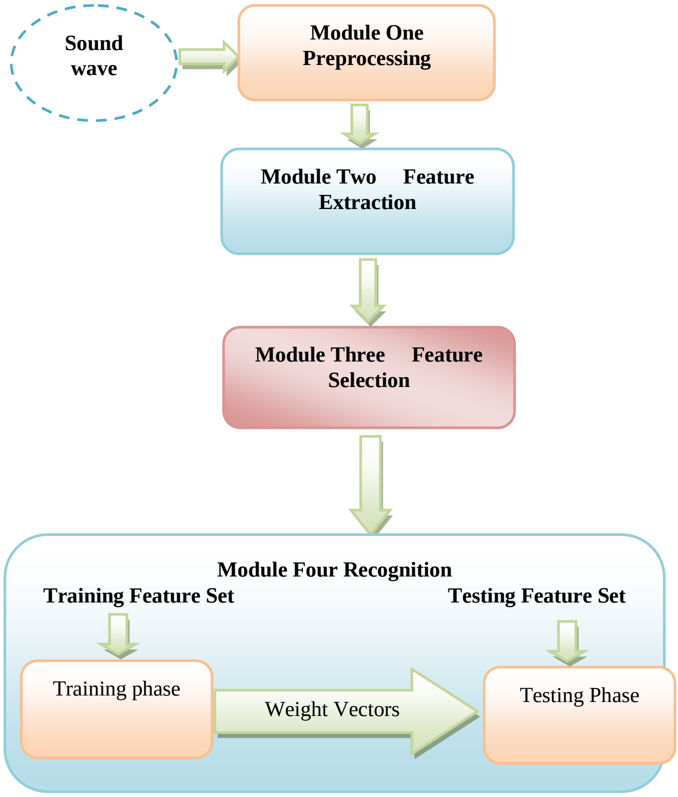

Fig. 3. Developed System Flow Control.

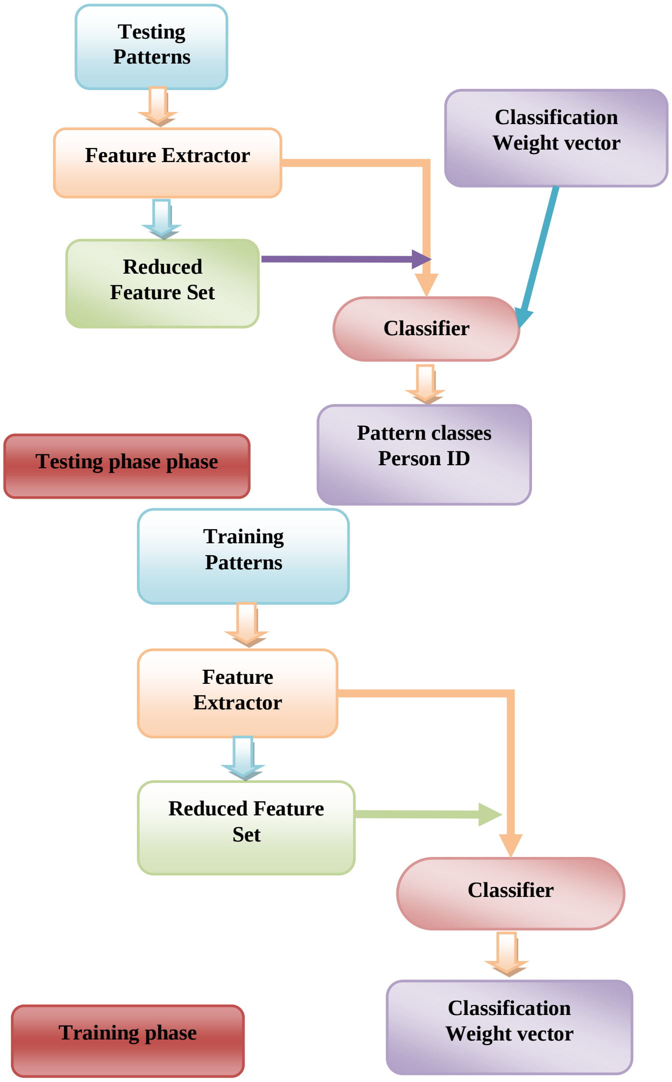

Fig. 4. Recognition System Phase.

IV. Assessment of Results

In this research, the dataset involves various sounds download from (http://www.voxforge.org) for fifty various persons (male and female), each one spoke ten various statements, where the dataset consists of 500 samples in total.

To minimize the loss of information in a signal speech, the parameters gaining from data must be chosen according to the speech signal nature to being processed. Signal of speech is used within this work, quantized with 16-bit quantization level and sampled with Fs=8 KHz.

For performance measurement of proposed system (use part of the data for training and whole it for testing, were the features that extracted from each level independently was used for training Neural Network that was suggested, to determine the level with best classification ability.

The accuracy of classification when a set of discrete wavelet coefficients extracted independently for different three wavelet levels that are used to train the BP NN is tabulated in Table I.

TABLE I. ACCURACY OF CLASSIFICATION FOR BP (DIFFERENT DISCRETE WAVELET LEVEL)

# of levels	Level 1	Level 2	Level 3
Testing accuracy	95	94	100

From Table I, it is inferred that level three achieved the best accuracy, where level one and level two shows acceptable discrimination ability. Fig. 5 shows the accuracy of classification for different discrete wavelet levels.

It is very necessary to study, the impact of concatenation (features extracted from various levels with other technique) to investigate if there is any impact on the classification ability in a form that can be positively or negatively done, where in this work curvelet transform is used with discrete wavelet transformation.

The accuracy of classification, in which the classifier can be trained over set of patterns, represent set of Discrete wavelet coefficients + curvelet), as shown in Table II.

From Table II, it is inferred that the accuracy was increased in level one and two when curvelet is concatenation with discrete wavelet transformation and gives the best classification accuracy, which is equal to the classification accuracy of level three. Fig. 6 shows the accuracy of classification when DWT and curvelet were combined.

The output of feature extraction has many features of which none is important for speaker discrimination, and the number of features should be also relatively low.

The process of feature selection is to select the best features that describe the speaker when dealing with hundreds of features that lead to increasing the workload of recognition. Selecting the best features set leads to reducing the classifier training time and as well as increasing the classification accuracy [23].

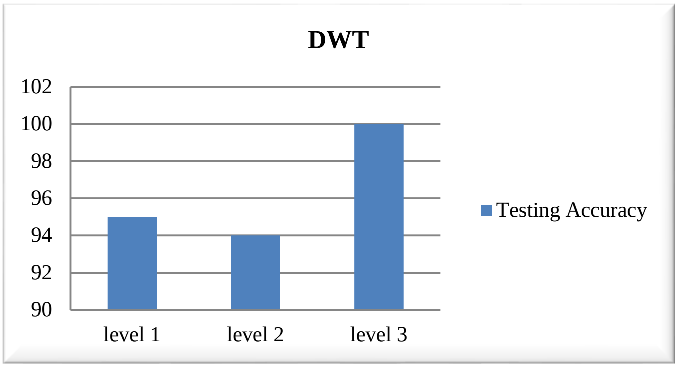

Fig. 5. Accuracy of Classification of different Discrete Wavelet Level.

TABLE II. ACCURACY OF CLASSIFICATION USING (FEATURES EXTRACTED FROM CONCATENATION DWT AND CURVELET)

# of levels	Level 1	Level 2	Level 3
Testing accuracy	100	100	100

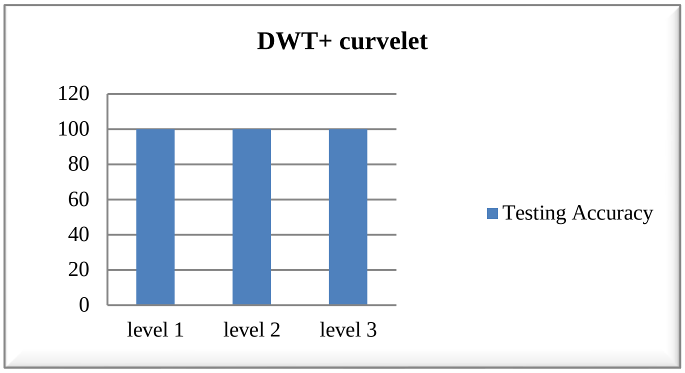

Fig. 6. Accuracy Classification of Combined different DWT Level with Curvelet.

The accuracy of classification using principal component analysis in addition to discrete wavelet + curvelet is shown in Tables III and IV.

From Tables III and IV, it is inferred that the accuracy was impacted positively and it is clear that reducing the features by using PCA did not affect the classification accuracy where the classification accuracy of level one and level two was increased to achieve the best classification and the accuracy of level three still 100%. Fig. 7 shows the accuracy of classification when PCA applied to DWT and (DWT + curvelet), respectively.

TABLE III. ACCURACY OF CLASSIFICATION USING PCA WITH (DWT + CURVELET)

# of levels	Level 1	Level 2	Level 3
Testing accuracy	100	100	100

TABLE IV. ACCURACY OF CLASSIFICATION USING PCA WITH DWT

# of levels	Level 1	Level 2	Level 3
Testing accuracy	100	100	100

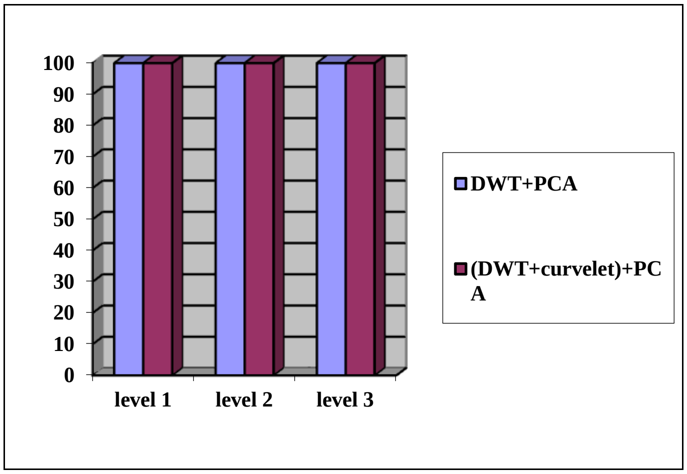

Fig. 7. The Accuracy of Classification of Applying PCA.

V. Conclusion

This paper concentrates on analyzing and studying the effect of using various levels of discrete wavelet transformation in addition to concatenation of different DWT level with curvelet technique to perform speaker identification. Also, the behaviors of classifier (Backpropagation Neural Network) were studied within the field of speaker recognition.

The practical results showed that level three of DWT gives the best accuracy where achieved to 100% and the accuracy was improved in level (1 and 2) when applying (DWT + curvelet).

In this approach, it is clear that introducing PCA with BP networks improved the accuracy. This approach is an effective method for speaker identification system, where it keeps the effective information and reduces the redundancy of characteristic parameters. Fig. 8 shows the effect of using different techniques of feature extraction using three different levels of discrete wavelet transformation.

Future work will focus on integrated some techniques with each other to increase the accuracy of speaker identification system such as DWT&LPC, LPC&MFCC were the development can be occurs in this stage that concentrated on reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable recognition accuracy.

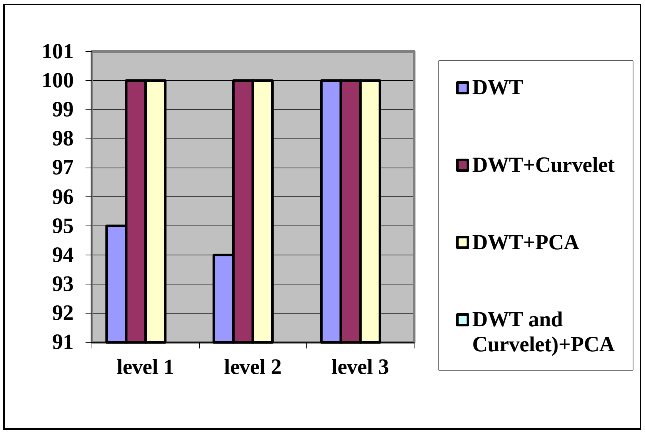

Fig. 8. Classification Accuracy of different Feature Extraction Technique.

References

[1] Prabhakar, S., S. Pankanti, and A.K. Jain, Biometric recognition: Security and privacy concerns. IEEE security & privacy, 2003. 99(2): p. 33-42.

[2] Tripathi, K., A comparative study of biometric technologies with reference to human interface. International Journal of Computer Applications, 2011. 14(5): p. 10-15.

[3] Woo, S.C., C.P. Lim, and R. Osman. Development of a speaker recognition system using wavelets and artificial neural networks. in 2001. Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, . 2001. IEEE.

[4] Togneri, R. and D. Pullella, An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine, 2011. 11(2): p. 23-61.

[5] Shah, S.M. and S.N. Ahsan. Arabic speaker identification system using combination of DWT and LPC features. in 2014 International Conference on Open Source Systems and Technologies (ICOSST),. 2014. IEEE.

[6] Nemati, S. and M.E. Basiri, Text-independent speaker verification using ant colony optimization-based selected features. Expert Systems with Applications, 2011. 38(1): p. 620-630.

[7] Powar, S.M. and V. Patil, Study of Speaker Verification Methods.

[8] Singh, N. and R. Khan, Speaker Recognition and Fast Fourier Transform. International Journal, 2015. 5(7).

[9] Gbadamosi, L., Text independent biometric speaker recognition system. International Journal of Research in Computer Science, 2013. 3(6): p. 9.

[10] Sekkate, S., M. Khalil, and A. Adib. Fusing wavelet and short-term features for speaker identification in noisy environment. in, 2018 International Conference on Intelligent Systems and Computer Vision (ISCV). 2018. IEEE.

[11] Kumar, C., et al. Analysis of MFCC and BFCC in a speaker identification system. in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), . 2018. IEEE.

[12] Biswal, K.T. and J. Rout, Speaker Recognition System using Wavelet Transform under Stress Condition. IJSEAT, 2017. 4(12): p. 745-751.

[13] Omar, N.M. and M. El-Hawary. Feature fusion techniques based training MLP for speaker identification system. in 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE),. 2017. IEEE.

[14] Al-Shayea, Q.K. and M.S. Al-Ani, Speaker Identification: A Novel Fusion Samples Approach. International Journal of Computer Science and Information Security, 2016. 14(7): p. 423.

[15] AboElenein, N.M., et al. Improved text-independent speaker identification system for real time applications. in 2016 Fourth International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC),. 2016. IEEE.

[16] Dhineshkumar, R., A.B. Ganesh, and S. Sasikala, Speaker identification system using gaussian mixture model and support vector machines (GMM-SVM) under noisy conditions. Indian Journal of Science and Technology, 2016. 9(19).

[17] Sukhwal, A. and M. Kumar. Comparative study of different classifiers based speaker recognition system using modified MFCC for noisy environment. in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT),. 2015. IEEE.

[18] Kaur, K. and N. Jain, Feature Extraction and Classification for Automatic Speaker Recognition System-A Review. International Journal of Advanced Research in Computer Science and Software Engineering, 2015. 5.

[19] Maged, H., A. AbouEl-Farag, and S. Mesbah. Improving speaker identification system using discrete wavelet transform and awgn. in 2014 5th IEEE International Conference on Software Engineering and Service Science (ICSESS), . 2014. IEEE.

[20] Albin, A.J., N. Nandhitha, and S.E. Roslin. Text independent speaker recognition system using Back Propagation Network with wavelet features. in 2014 International Conference on Communications and Signal Processing (ICCSP). 2014. IEEE.

[21] Nidhyananthan, S.S. and R.S.S. Kumari, Language and text-independent speaker identification system using GMM. WSEAS Transactions on Signal Processing, 2013. 9(4): p. 185-194.

[22] Nagaraja, B. and H. Jayanna, Combination of Features for Multilingual Speaker Identification with the Constraint of Limited Data. International Journal of Computer Applications, 2013. 70(6): p. 1-6.

[23] Srinivasan, V., V. Ramalingam, and V. Sellam, Classification of Normal and Pathological Voice using GA and SVM. International Journal of Computer Applications, 2012. 60(3): p. 34-39.