Speaker Verification System Using Discrete Wavelet Transform And Formants Extraction Based On The Correlation Coefficient

Tariq Abu Hilal , Hasan Abu Hilal, Riyad El Shalabi and Khalid Daqrouq

***Abstract*—In this paper, Discrete Wavelet Transform (DWT), and Logarithmic Power Spectrum Density (PSD) are integrated for speaker accurate formants extraction, afterward correlation coefficient is used for features classification, the correlation thresholding factor is adjusted. As the system works with the recorded samples, the features tracking capability was excellent with text dependant dataset; so the system can be applied in password, PINs identification, security system or mobile phones. The proposed system is simulated; the results show excellent performance, around 95 % Recognition Rate.**

***Index Terms*—Discrete Wavelet Transform, Power Spectrum Density, Speaker Identification.**

I. Introduction

THE applications of speech signal processing, such as speech recognition or speaker identification, improved rapidly in the last 5 years, because of the new technology of hardware and software; speaker identification structure can be utilized in suspect identification. The consciousness of speaker identification can be divided into two main parts: features extraction, followed by speaker's voices classification, based on the extracted features [1] [2]. Speaker Identification systems (SI) have been under progressing since more than six decades, a lot of researchers interested. From a commercial viewpoint, speaker identification system is a technology with potentially large market, due to the applications of broadly ranges of automation for operator assisted services [3] [4].

A basic question in speech recognition is how speech patterns are evaluated to settle on their likeness (the expanse between patterns). Depending on the particulars of the recognition systems, pattern association can be done in a broad selection of ways, basically the hypothesis is that the speech is comprised of a word or more and to be predictable as a complete unit with no unambiguity in the phonetic contented. Utterances detection algorithms consist of identical components (via time alignment), the calculated succession of spectral vectors of the spoken parts to be created, alongside for each position of the spectral patterns, and picking and building up the voice patterns. One more implied supposition is that every spoken utterance has a plainly distinct beginning and ending point, which could be found using some type of speech endpoint detector.

Manuscript received Decembe 08, 2010; revised December 22, 2010.

Tariq Abu Hilal is with the Department of Math/IT, Dhofar University, Salalah, Oman e-mail: tariq_abuhilal@du.edu.om.

Hasan Abu Hilal is with the Department of Math/IT , Dhofar University, Salalah, Oman e-mail: h_abuhilal@du.edu.om.

Riyad Al-Shalabi is with the Department of Information Technology, Petra University, Amman, Jordan, e-mail: ralshalabi@uop.edu.jo.

Khalid Dakrouq is with the Department of Electrical Engineering, Philadilphia University, Amman, Jordan, e-mail: haleddaq@yahoo.com.

As a consequence, pattern corresponding could be dependably completed, without need to be concerned about uncertainties in the endpoints of the patterns being compared. For many requests, particularly those referred to as command and control claims, in which the user is obligatory to speak the command words one at a time (i.e. with separate gaps and command words), this model's function is totally suitable. Nevertheless, some applications where the speech to be well-known, those consist of a sequence of words from the recognition vocabulary, such a paradigm is often improper for practical reasons.

In this paper, we consider the performance and simplicity of the correlation process, to be applied on the wavelet transform and power spectral calculated coefficients, to be applied for identical text and different speakers; we investigate the similarities of the manipulated signal. The organization of this paper is as follows. In Section 2, the speaker recognition process is briefly described. In Section 3, we show the model of the proposed system. In Section 4, we depict and discuss the results of the proposed method. Finally, conclusions are stated out in section 5.

II. Speaker Recognition Process

The speaker recognition model starts with generating a speech signal by speaking a complete given words, the spoken production is decoded into speech signal as a vector of values, the process of speaker recognition is a combination of the input step, signal processing step, then classification and recognition step. First, a stored data set is used to be processed , afterward the manipulation stage, finally, features of each speech signal are stored as reference features, for training and validation sets, verification is to minimize the error rate and to achieve precise recognition rate.

The feature vectors of speech are used to create a pattern for each speaker, the number of reference models that are required for effective speaker recognition application depends upon the type of features and methods that the system uses for identifying any speaker, the purposeful features those are the same as stored are extracted from an input voice wave of the speaker to be authenticated, afterward the acceptance depends upon the comparison and the relationship between the stored model and the extracted

ISBN: 978-988-19251-2-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

ones from the entered signal.

In speaker identification and recognition, the difference between an input voice waves and all other recorded data set patterns is computed. The pattern of the listed user, whose difference with the input voice wave's model is the smallest will be accepted as the same speaker of the input voice waves. Many methods were proposed to do so in the literature. In case of speaker verification the similarity is computed only between the input signal and the stored patterns of the other recorded speakers. If the result is less than profound threshold, then the speaker will be accepted, otherwise will be considered as an imposter and discarded.

The drawback of speaker recognition can be divided into two main sub problems; speaker verification and speaker identification. Speaker identification is considered as the assignment of identifying who is talking from a dataset of known voices of speakers. A given utterance based on the information restricted in speech signals is the process of deciding who is speaking. The unknown voice comes from an unchanging set of known speakers; therefore we need an identification task to be referred as a closed set identification, the process of accepting or rejecting the speaker to be the authenticated one is called speaker verification, since it is assumed that imposters those who pretend as valid users to fake the systems are not known, this is referred to as the open set duty, this derives two essential choices to the closed set identification task, which enable a combination of the two tasks, and it is called open set identification.

A failure that may take place in speaker identification is the false identification of a speaker, and the faults in speaker verification can be classified into the following; false rejections, which is an actual speaker is rejected as an imposter who fakes the system, and false acceptances, which is a false speaker or imposter is accepted as a true one, where indeed, he is not recorded in the dataset [5] ;

False rejection: an actual speaker is rejected as an imposter who fakes the system. False acceptances: a false speaker or imposter is accepted as a true one, where indeed, he is not recorded in the dataset.

In the majority of speaker recognition systems, a categorization towards stored speaker's label is processed and compared with pre-classified threshold. The verification and the rejection processes depend on a predefined threshold value, if the computed deference is below the threshold, the speaker is confirmed, if not the speaker is discarded as an imposter.

The decision threshold is placed at the point where both errors are equiprobable, the speaker identification techniques can also be separated into text dependent and text independent systems. In case of text dependent systems a speaker is required to produce a predefined set of words, sentences, passwords, or numbers. Features of any voice can be extracted from the same signals, there is no predefined set of words or sentences regarding the text independent methods, and the speakers could not even be aware about being recognized. Both the text dependent and independent systems face the same problems. These

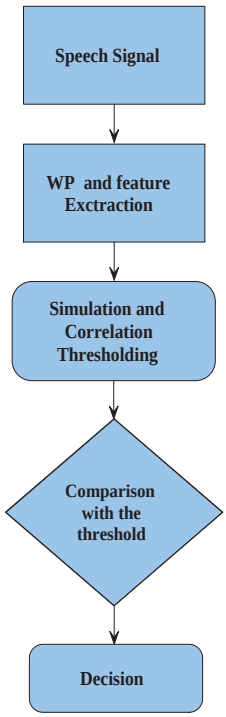

Fig. 1. Block diagram of the proposed system algorithm

methods can be misleading because of someone who plays back a recorded voice of an actual speaker uttering, the key words or passwords can be verified as the registered speaker.

The use of recorded dataset, which is arbitrarily chosen at all times, can be repeated in the requested order, by sophisticated electronic recording tools, therefore a text will be provoked as a computer driven and text dependent speaker recognition system. With the unification of speaker and speech recognition systems, over the development in speech recognition precision, the characteristic between text dependent and independent systems will ultimately decrease. The text dependent speaker recognition is the most commercially feasible and useful application, even though there is a lot of research which carried out on both of them. On the other hand, due to the promises offered, more awareness is being given to the text dependent methods of speaker recognition neglecting their complexity.

III. The Proposed System

In this paper wavelet transform system is constructed and presented, it consists of three main parts; signal transforming, features extracting and similarity comparison, reducing the complexity of using Neural Network. The advantage of the system than the other classification techniques is the accumulate features tracking and fast identification (i.e. the neural networks complexity). The transform depends on discrete wavelet function's density with a selected level, and chosen function type, to better differences tracking of the signal's behavior, processing and calculating the exact variation in the frequency. That what exactly happens in non-stationary signals, such as speech signal, then the correlation coefficient is used on the manipulated signal to classify and verify.

ISBN: 978-988-19251-2-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
CORRELATION COEFFICIENT BETWEEN THE FIRST SPEAKER 10 UTTERANCES

Sample	1	2	3	4	5	6	7	8	9	10
1	1.00000	0.98521	0.98465	0.98521	0.99655	0.98689	0.99429	0.99917	0.99716	0.99880
2	0.98521	1.00000	0.99208	0.97031	0.99603	0.99809	0.99505	0.99137	0.99510	0.99241
3	0.98465	0.99208	1.00000	0.97853	0.99191	0.99795	0.99755	0.98906	0.99300	0.99015
4	0.98521	0.97031	0.97853	1.00000	0.98318	0.97628	0.98548	0.98656	0.98488	0.98627
5	0.99655	0.99603	0.99191	0.98318	1.00000	0.99690	0.99837	0.99910	0.99986	0.99941
6	0.98689	0.99809	0.99795	0.97628	0.99690	1.00000	0.99836	0.99220	0.99604	0.99327
7	0.99429	0.99505	0.99755	0.98548	0.99837	0.99836	1.00000	0.99835	0.99894	0.99770
8	0.99917	0.99137	0.98906	0.98656	0.99910	0.99220	0.99835	1.00000	0.99935	0.99996
9	0.99716	0.99510	0.99300	0.98488	0.99986	0.99604	0.99894	0.99935	1.00000	0.99963
10	0.99880	0.99241	0.99015	0.98627	0.99941	0.99327	0.99770	0.99996	0.99963	1.00000

TABLE II
CORRELATION COEFFICIENT BETWEEN THE SECOND SPEAKER 10 UTTERANCES

Sample	1	2	3	4	5	6	7	8	9	10
1	1.00000	0.97314	0.98809	0.97314	0.99306	0.98277	0.99343	0.99825	0.99488	0.99757
2	0.97314	1.00000	0.99213	0.97991	0.99347	0.99791	0.99139	0.98506	0.99130	0.98681
3	0.98809	0.99213	1.00000	0.98641	0.99685	0.99814	0.99898	0.99422	0.99719	0.99514
4	0.97314	0.97991	0.98641	1.00000	0.98978	0.98520	0.98942	0.98982	0.99012	0.99007
5	0.99306	0.99347	0.99685	0.98978	1.00000	0.99726	0.99913	0.99828	0.99980	0.99883
6	0.98277	0.99791	0.99814	0.98520	0.99726	1.00000	0.99726	0.99172	0.99629	0.99305
7	0.99343	0.99139	0.99898	0.98942	0.99913	0.99726	1.00000	0.99912	0.99953	0.99857
8	0.99825	0.98506	0.99422	0.98982	0.99828	0.99172	0.99912	1.00000	0.99908	0.99994
9	0.99488	0.99130	0.99719	0.99012	0.99980	0.99629	0.99953	0.99908	1.00000	0.99949
10	0.99757	0.98681	0.99514	0.99007	0.99883	0.99305	0.99857	0.99994	0.99949	1.00000

The wavelet series is just a sampled version of Continuous Wavelet Transform (CWT) and its calculation may devour important amount of time and assets, depending on the resolution requisite. The Discrete Wavelet Transform (DWT), which is based on sub-band coding, is started to yield a fast computation of the wavelet transform. It is easy to put into practice and lessens the totaling time and resources required [6]. Filters are one of the mainly extensively used signal processing functions. Wavelets can be comprehended by iteration of filters with rescaling. The resolution of the signal, which is a gauge of the amount of feature information in the signal, this is taken by the filtering operations, and the scale is determined by up sampling and down sampling (sub sampling) operations [7].

In statistical signal processing and physics, the spectral density, Power Spectral Density (PSD), or Energy Spectral Density (ESD), is a positive real function of a frequency changeable linked with a motionless stochastic process, or a deterministic function of time, which has magnitudes of power per Hz, or energy per Hz. It is frequently called just the spectrum of the signal.

Instinctively, the spectral density detains the frequency substance of a stochastic process and assists recognize periodicities. An often more practical substitute is the Power Spectral Density (PSD), which explains how the power of a signal or time series is circulated with frequency. Here power can be the definite physical power, or more often, for expediency with theoretical signals, which can be distinct as the squared value of the signal, this is as the real power. The power spectral density of a signal exists if and only if the signal is a wide-sense stationary process. If the signal is not stationary, then the autocorrelation function must be a function of two variables, so no PSD exists, but similar ways may be used to estimate a time-varying spectral density [8].

Formants are the frequency parts of speech signal those are related to the human distinct vocal tract anatomy form, which is distinguishable for each person's resonance. We use these formants as the basic speaker features carriers [9], which is determined by PSD which is estimated using the Yule-Walker Autoregressive (AR) method. This method, also called windowed method, it fits an AR linear prediction filter model to the signal by minimizing the forward prediction error in the least squares sense. This formulation leads to the Yule-Walker equations, which have been solved by the Levinson-Durbin recursion [10] [11]. The spectral estimate returned by method is the squared magnitude of the frequency response of this AR model. Then N vector of the speaker's formants is represented by logarithmic scale as in equation 1.

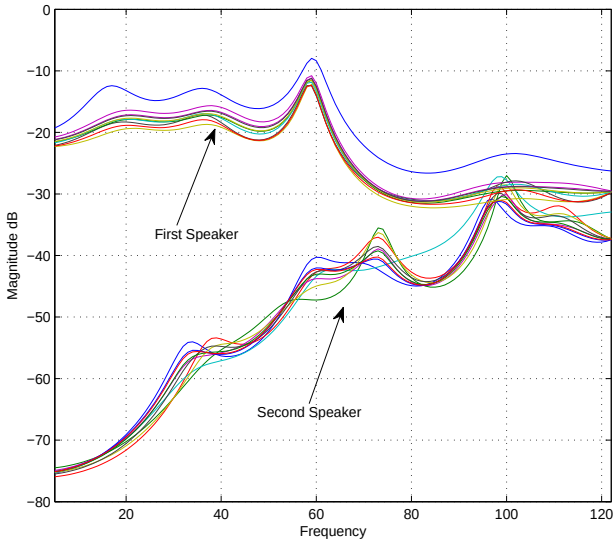

Fig. 2. Two speakers extracted features (10 utterances each)

$$F(n) = \sum_{n}^{N} 10log10(P), \tag{1}$$

We use the discrete wavelet transform approximation coefficients, denoted as a_j of multiple scales as an input to P, at that time P output is denoted by WP as shown in equation 2.

$$a_{j+1}(t) = \sum_{m} h(m-2t)a_j(m), \tag{2}$$

Where, the set of numbers $a_j(m)$ represents the down sampled approximation of the signal at the resolution 2^{-j}, $h(n)$

ISBN: 978-988-19251-2-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE III
CORRELATION COEFFICIENT BETWEEN THE FIRST AND THE SECOND SPEAKER 10 UTTERANCES

Sample	1	2	3	4	5	6	7	8	9	10
1	-0.62131	-0.68382	-0.66463	-0.71982	-0.65747	-0.67528	-0.65667	-0.64058	-0.65339	-0.64393
2	-0.61945	-0.69154	-0.67069	-0.71866	-0.66049	-0.68216	-0.66051	-0.64117	-0.65603	-0.64503
3	-0.61416	-0.68769	-0.66745	-0.71375	-0.65590	-0.67862	-0.65508	-0.63623	-0.65150	-0.64019
4	-0.57502	-0.64407	-0.61992	-0.67844	-0.61420	-0.63290	-0.61159	-0.59573	-0.60938	-0.59928
5	-0.62272	-0.69010	-0.67003	-0.72193	-0.66137	-0.67917	-0.65989	-0.64325	-0.65709	-0.64685
6	-0.61808	-0.69102	-0.67042	-0.71767	-0.65954	-0.68177	-0.65844	-0.64001	-0.65510	-0.64393
7	-0.62013	-0.68980	-0.66974	-0.71968	-0.65993	-0.68082	-0.65994	-0.64123	-0.65562	-0.64498
8	-0.62254	-0.68750	-0.66786	-0.72148	-0.65995	-0.67873	-0.65935	-0.64244	-0.65577	-0.64592
9	-0.62235	-0.68970	-0.66974	-0.72162	-0.66099	-0.68077	-0.66063	-0.64287	-0.65672	-0.64648
10	-0.62257	-0.68812	-0.66840	-0.72160	-0.66029	-0.67931	-0.65974	-0.64262	-0.65609	-0.64614

is the coefficient of the linear combination that approximates the wavelet scaled version function [12] [13] , and the correlation coefficient is calculated as:

$$\rho = \frac{E[(X - \overline{X})(Y - \overline{Y})]}{\sigma_X \sigma_Y}, \tag{3}$$

where ρ is the Correlation coefficient and $E[.]$ denotes the expectation of the product of the speech signal model, vector **X** is about the mean value, and the speech signal vector **Y** is about the mean value that related to the product of the Standard Deviation of $X(\sigma_X)$ and Standard Deviation of $Y(\sigma_Y)$. ρ is efficient likeness or similarity tool judgment between the two vectors X and Y in terms of unity (out of one hundred percent similarity).

IV. SIMULATION RESULTS

Discrete Wavelet Transform (DWT) based speaker feature extraction method is investigated by the use of the correlation coefficients. The introduced system depends on two steps: features extraction by PSD over DWT level, due to its better capability of formants illustration over different bandpass of signal frequency, which is more suitable for non stationary signals, this is shown step by step in Fig. 1, and classification based statistical analysis of each speaker vector, in other words, the correlation value of the new input vector is compared with a threshold, the system will reject any imposter if the result has a low correlation coefficient, and will accept true speaker who has high values as shown in Tables III, III and III.

The system works with excellent capability of features tracking, simulation has been conducted on 100 different speakers, tables show only two different speakers with 10 voice samples for each. The threshold value between the users is adjusted to minimize the error and maximize the recognition rate. This system reduces the complexity of using other techniques like Neural Networks, with high verification results. Text - dependant system is used, with MATLAB package simulation tools, the results show excellent performance, approximately 95% classification rate.

For two different speakers with ten audio recordings, each for the same text uttering, processed and classified; Figure 2 shows the voice signals after processing with DWT and PSD, afterward classification results will be done by correlation coefficient. Table III shows the high recognition rates between the same first speaker's trials, and Table III shows the high recognition rates between the same second speaker's trials. Table 3 shows the low recognition rate between the first speaker's trials and the second speaker's trials, since the recognition rate in both Table 1 and 2 is around 98% in this case for the same speaker, and the recognition rate in Table III is around 66% between the two different speakers, so the high recognition rate means that the same speaker, but the low recognition rate means different speakers (an imposter). If the recognition rate is above the predefined threshold, in any verification trial, then the acceptance decision will be taken, otherwise rejection decision is considered.

V. CONCLUSIONS AND FUTURE WORK

Throughout this paper, Discrete Wavelet Transform, and Power Spectrum Density based speaker feature extraction method is investigated by the use of the correlation coefficient. The introduced system depends on two steps features extraction over fixed approximation DWT level, due to its outperforming capability of formants illustration over the signal frequency, and classification using the correlation coefficient.

After processing the first speaker 10 utterances by the system, Table III compares between the same ten trials for the first speaker; it shows very high recognition rates.

After processing the second speaker 10 utterances by the new system, Table III compares between the same ten trials for the second speaker; it shows very high recognition rates.

After processing two speakers utterances by the system, Table III compares between ten trials for each speaker, for both the first and the second speaker's trials; speaker one and two signals does not have high recognition rate, which means no similarity. The system works with excellent capability of classification and similarity detection, the system can be applied for several speech classification problems.

REFERENCES

[1] K. Daqrouq, T. Abu Hilal, M. Sherif, S. El-Hajjar, and A. Al-Qawasmi,"Speaker Identification System Using Wavelet Transform and Neural Network," IEEE 2009 International Conference on Advances in Computational Tools for Engineering Applications.- 2009 Advances in Computational Tools for Engineering Applications, Lebanon

[2] J.Wu -D., Lin B.-F. , "Speaker identification using discrete wavelet packet transform technique with irregular decomposition Expert Systems with Applications", 36 (2009) 3136-3143.

[3] D. Avci , "An expert system for speaker identification using adaptive wavelet sure entropy", Expert Systems with Applications, 36 (2009) 6295-6300.

[4] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). "Speaker verification using adapted gaussian mixture models". Digital Signal Processing," 10(1-3), 19-41.

[5] B. Imperl, "Speaker recognition techniques, Laboratory for Digital Signal Processing", Faculty of Electrical Engineering and Comp. Sci., Smetanova 17, 2000 Maribor, Slovenia.

[6] D. Ranjan Panda and Chittaranjan Nayak, "Eye Detection Using Wavelets And Ann", A Thesis Submitted In Partial Fulfillment Of Department Of Electronics and Instrumentation Engineering National Institute Of Technology Rourkela-769008 - (2007).

[7] N. Rao and A. Govardhan, "Comparative Study of Visible Reversible Watermarking Algorithms" Image Security Paradigm, Engineering College, Hyderabad, A P, India, Vol. 7, No. - 177 (2, April 2010).

ISBN: 978-988-19251-2-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[8] W. Al-Sawalmeh, Khaled Daqroug and Tareq Abu Hilal, "The use of wavelets in speaker feature tracking identification system using neural network," WSEAS Transactions on Signal Processing archive Volume 5 , Pages: 167-177 Year of Publication: 2009 ISSN:1790-5022 Issue 5, (May 2009).
[9] J. M. Naik, L. P. Netsch, and G. R. Doddington. "Speaker verification over long distance telephone lines." In IEEE Proceedings of the 1989.
[10] S. Marple, L., *Digital Spectral Analysis*, Prentice-Hall, 1987, Chapter 7.
[11] P. Stoica,, and R.L. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, 1997.
[12] P. Goupillaud, A. Grossmann, J. Morlet, "Cycle-octave and related transforms in seismic signal analysis", Geoexploration, 23, 85-102, 1984-1985.
[13] A. Grossmann and J. Morlet, Decomposition of Hardy. "functions into square integrable wavelets of constant shape," SIAM J. Math. Anal, Vol. 15, pp 723-736, 1984.
[14] Meyer, Wavelets, Ed. J.M. Combes et al., Springer Verlag, Berlin, p. 21, 1989.

ISBN: 978-988-19251-2-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)