Self-Organizing Map Weights and Wavelet Packet Entropy for Speaker Verification

K.Daqrouq[1], A. Al-Qawasmi[1], O.Daoud[1] and W.Al-Sawalmeh[2]

[1]Department of Communication and Electronics, Philadelphia University, Jordan.

[2]Department of Communication Engineering, Al-Hussein Bin Talal University, Ma'an, Jordan.

Corresponding author: O. Daoud

***Abstract*—**With the growing trend toward distant security verification systems for telephone banking, biometric security measures and other remote access applications, Automatic Speaker Verification (ASV) has attracted a great attention in recent years. The complexity of ASV system and its verification time depends on the number of feature vector elements. Therefore, in this paper, we concentrate on optimizing dimensionality of feature space by selecting the weights of Self-Organizing Map (WSOM) Neural Network (NNT) for text-independent speaker verification system. This is accomplished by decreasing the number of feature vector elements of individual speaker obtained by using wavelet packet (WP) Shannon, Sure, and log energy in conjunction with energy indices (1020 elements) to 64 elements by WSOM. To investigate the performance of the proposed WSOM and wavelet packet entropies (SOMWPE) method, two other verification methods are proposed: Gaussian mixture model based method (GMMWPE) and K-Means clustering based method (KMWPE). The results indicated that a better verification rate for the speaker-speaker system was accomplished by SOMWPE. Better result was achieved (94.34%) in case of the speaker-imposter verification system. In case of white Gaussian noise (AWGN), it was observed that the SOMWPE system is generally more noise-robust than GMMWPE and KMWPE systems.

***Keywords*—**About four key words or phrases in alphabetical order, separated by commas.

I. Introduction

AUTOMATIC Speaker Verification (ASV) refers to the mission of verifying speaker's identity by means of the speaker-specific information contained in speech signal. Speaker verification methods are absolutely divided into text-dependent and text-independent applications. When the same text is used for both training and testing, the system is called to be text-dependent but for text-independent process, the text used to train and test of the ASV system is totally unconstrained. The text-independent speaker verification necessitates no restriction on the type of input speech. In contrast, the text-independent speaker verification generally gives less performance than text-dependent speaker verification, which requires test input to be the same utterances as training data [1,2].

Speaker verification has been the topic of active research for many years, and has many important applications where propriety of information is a concern [3]. Applications of speaker verification can be found in biometric person authentication such as an extra identity check in credit card payments over the Internet while, the potential applications of speaker identification may be utilized in multi-user systems. For example, in speaker tracking the task is to locate the segments of given speaker(s) in an audio stream [4-6]. It has also potential applications in an automatic segmentation of teleconferences and helping in the transcription of courtroom discussion.

There has been a wide spectrum of proposed approaches to speaker verification starting with very simplistic models such as those based on long term statistics [7]. The most sophisticated methods rely on large vocabulary speech recognition with phone-based HMMs [8].

Feature extraction is a key stage in speaker verification systems. Speech extracted features used in a speaker verification system drop within two classes based on their related space. One class includes features defined in an unconditional or absolute and irrelative space, while the other includes features defined in a relative space. For the first class, depiction of a speaker in the feature space is not related to any reference speaker [9]. While there is a momentous body of

Manuscript received December 26, 2011.

K. Daqrouq is with Department of Communication and Electronics, Philadelphia University, Jordan.

A. Al-Qawasmi. Department of Communication and Electronics, Philadelphia University, Jordan.

O.Daoud is Department of Communication and Electronics, Philadelphia University, Jordan., (corresponding author:962785583769; fax:96264799037; e-mail: odaoud@philadelphia.edu.jo).

W. AlSawalmeh is with [2]Department of Communication Engineering, Al-Hussein Bin Talal University, Ma'an, Jordan.

literature on features in the absolute space, very little research has been conducted for studying the properties of features extracted in the relative space. Mel-frequency cepstral coefficient (MFCC), Linear Prediction Cepstrum Coefficient (LPCC), wavelet coefficients, etc. are among the most common speech features in absolute space. In recent times, Campbell et al. used Maximum A Posteriori (MAP) adapted GMM mean super vectors as an absolute feature with Support Vector Machine (SVM) as a discriminative model for speaker verification [10-12]. For features defined in a relative space, each speaker in the feature space is described relative to some reference speakers. As well as, extracted features in the relative space can be applied in conjunction with any other set of techniques from the verification phase menu that are deemed more suitable. Merlin et al. proposed a new approach to speaker recognition and indexation systems, based on no directly-acoustic processing in the relative space[13]. In 2000 Kuhn et al. introduced the eigenvoices concept and represented each new speaker relative to eigenvoices [14,15]. Afterwards, other researchers used a different approach where they introduced the idea of space of anchor models to represent enrolled speakers in verification systems, and to verify a test speaker in a relative feature space [9, 16-18].

Speech features are often extracted by Fourier transform (FT) and short time Fourier transform (STFT). Unfortunately, they accept signals stationary within a given time frame and may therefore lack the ability to represent localized events properly. Recently, wavelet transform (WT) has been proposed for feature extraction. The particular benefit of wavelet analysis possesses is the characterizing signals at different localization levels in both time and frequency domains[19,20]. Furthermore, the WT is well suited to the analysis of non-stationary signals. It provides an alternative to classical linear time–frequency representations with better time and frequency localization characteristics. In earlier studies, these properties were applied in speaker recognition, particularly wavelet packet transform (WPT)[21-23].

Artificial neural network performance is depending mainly on the size and quality of training samples [24]. When the number of training data is small, not representative of the possibility space, standard neural network results are poor [25]. Incorporation of neural fuzzy or wavelet techniques can improve performance in this case, particularly, by input matrix dimensionality decreasing. Artificial neural networks (ANN) are known to be excellent classifiers, but their performance can be prevented by the size and quality of the training set.

In this paper, we improve effective feature extraction method for text-independent system, taking in consideration that the size of feature vector is very crucial issue. For this reason, the presented features extraction method offers a reduction of dimensionality of speech signal comparing with conventional methods. Three types of entropy coefficients of WPT in conjunction with energy indexes of WPT are utilized. To overcome data training difficulty by standard NNT, authors propose Self-Organizing Map (SOM) for speaker verification. The SOM is an unsupervised method for forming a representation of data [26,27]. It consists of local data models located at the nodes of the low dimensional map grid. The question remains if SOM can be developed for speaker verification. The specific aim of the present study was to address this question by developing and evaluating SOM for speech based text-independent verification of the speaker from imposters. For better investigation two other verification methods are proposed; K-Means clustering method and Gaussian Mixture Model method.

The rest of this paper is organized as follows. Section 2 presents a brief overview of features extraction by WP. SOM, GMM and K-Means based verification methods are described in Section 3. Section 4 reports computational experiments. It also includes a brief discussion of the results obtained. Finally, the conclusion is offered in the last section.

II. Features Extraction by wavelet Packet

A. Wavelet Packet

The wavelet packet method is a generalization of wavelet decomposition that offers a richer signal analysis. Wavelet packet atoms are waveforms indexed by three naturally interpreted parameters: position and scale as in wavelet transform decomposition, and frequency. In the following, the wavelet transform is defined as the inner product of a signal $x(t)$ with the mother wavelet $\psi(t)$:

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) \quad (1)$$

$$W_{\psi}x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi*\left(\frac{t-b}{a}\right)dt \quad (2)$$

where a and b are the scale and shift parameters, respectively. The mother wavelet may be dilated or translated by modulating a and b.

The wavelet packets transform performs the recursive decomposition of the speech signal obtained by the recursive binary tree (see Fig.1). Basically, the WPT is very similar to Discrete Wavelet Transform (DWT) but WPT decomposes both details and approximations instead of only performing the decomposition process on approximations. The principle of WP is that, given a signal, a pair of low pass and high pass filters is used to yield two sequences to capture different frequency sub-band features of the original signal. The two wavelet orthogonal bases generated form a previous node are defined as

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k-2/n) \tag{3}$$

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k-2/n) \tag{4}$$

where $h[n]$ and $g[n]$ denote the low-pass and high-pass filters, respectively. In equations (3) and (4), $\psi[n]$ is the wavelet function. Parameters j and p are the number of decomposition levels and nodes of the previous node, respectively [19].

B. Feature Extraction by Wavelet Packet

For a given orthogonal wavelet function, a library of wavelet packet bases is generated. Each of these bases offers a particular way of coding signals, preserving global energy and reconstructing exact features. The wavelet packet is used to extract additional features to guarantee higher recognition rate. In this study, WPT is applied at the stage of feature extraction, but these data are not proper for classifier due to a great amount of data length. Thus, we have to seek for a better representation for the speech features. Previous studies showed that the use of entropy of WP as features in recognition tasks is efficient [12]

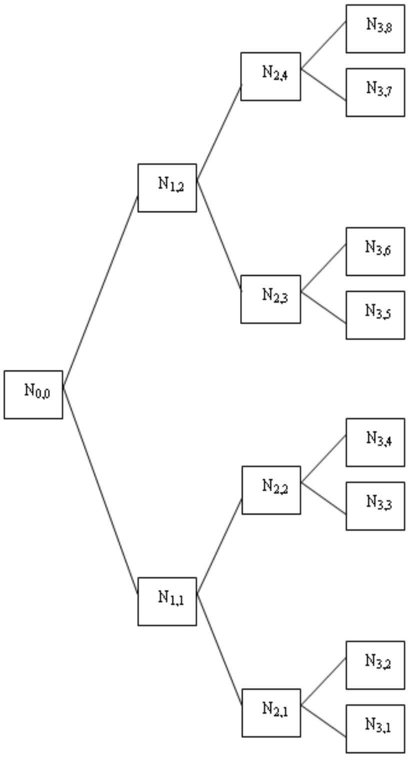


Fig. 1: Wavelet packet at depth 3

proposed a method to calculate the entropy value of the wavelet norm in digital modulation recognition. In [28] the author proposed features extraction method for speaker recognition based on a combination of three entropies types (sure, logarithmic energy and norm). Lastly, Avci in [23] investigated a speaker identification system using adaptive wavelet sure entropy.

As seen in above studies, the entropy of the specific sub-band signal may be employed as features for recognition tasks. In this paper, the entropy obtained from the WP will be employed for speaker identification. The wavelet packet features extraction method can be summarized as follows:

- Before the stage of features extraction, the speech data are processed by a silence removing algorithm followed by the application of a pre-processed by applying the normalization on speech signals to make the signals comparable regardless of differences in magnitude. In the present work, the signals are normalized by using the following formula (Wu & Lin, 2009):

$$S_{Ni} = \frac{S_i - \ddot{S}}{\sigma} \tag{5}$$

where S_i is the ith element of the signal S, $\ddot{S}$ and σ are the mean and standard deviation of the vector S, respectively, S_{Ni} is the ith element of the signal series S_N after normalization. Decomposing the speech signal by wavelet packet transform at depth 7 (level 7), with Daubechies type (db1).

- Calculating three entropy for all 256 nodes at depth 7 for wavelet packet using the equations [29,30]:

Shannon entropy:

$$E1(s) = -\sum_i s_i^2 \log(s_i^2) \tag{6}$$

Log energy entropy:

$$E1(s) = \sum_i \log(s_i^2) \tag{7}$$

Sure entropy:

$$|s_i| \le p \Rightarrow E(s) = \sum_i \min(s_i^2, p^2) \tag{8}$$

Where s is the signal, s_i are the WPT coefficients and p is a positive threshold. Entropy is a common concept in many fields, mainly in signal processing [31]. Classical entropy-based criterion describes information-related properties for a precise representation of a given signal. Entropy is commonly used in image processing; it posses information about the concentration of the image. On the other hand, a method for measuring the entropy appears as a supreme tool for quantifying the ordering of non-stationary signals. Fig.3 a shows Shannon entropy calculated for WP at depth 4 for three persons. For each person three different utterances were used.

Sure entropy was used at Fig.3 b and logarithmic energy entropy was used at Fig.3 c. we can notice that the feature vector extracted by Shannon entropy is more appropriate for speaker recognition. This conclusion has been obtained by interpretation the following criterion: the feature vector extracted should possess the following properties: 1) Vary widely from class to class. 2) Stable over a long period of time. 3) Should not have correlation with other features.

- For a better demonstration of the sub-band signals, the energy of speech is commonly computed. Previous investigations showed that the utilization of an energy index as features in recognition roles is efficient. In 2003, Kotnik et. al., in [31] proposed a robust speech recognition scheme in a noisy environment by means of wavelet-based energy as a threshold for de-noise estimation. In the biomedical field, Behroozmand and Almasganj are introduced a combination of genetic algorithm and wavelet packet transform used in the pathological evaluation, and the energy features are computed from a group of wavelet packet coefficients. Wu & Lin, in [19] mentioned that the energy indexes of WP were proposed for speaker identification. In this paper, the energy index of the WP is employed as additional features in conjunction with entropies for speaker verification tasks

C. A. Self-Organizing Feature Maps

Self-Organizing Feature Maps (SOFM) learn to classify input vectors along with how they are grouped in the input space (the architecture for this SOFM is shown at Fig.2). They differ from competitive layers in that neighboring neurons in the Self-Organizing Map learns to identify neighboring sections of the input space. Hence, SOM learn both the distribution as do competitive layers and topology of the input vectors they are trained on.

The neurons in the layer of an SOFM are set originally in physical positions along with a topology function. The function gridtop, hextop, or randtop can organize the neurons in a grid, hexagonal, or random topology. Distances between neurons are computed from their positions with a distance function. There are several distance functions, dist, boxdist, linkdist, and mandist. Link distance is the most popular. These topology and distance functions are illustrated in Topologies (gridtop, hextop, randtop) and Distance Functions (dist, linkdist, mandist, boxdist) [30].

A self-organizing feature map network identifies a winning neuron i^* by means of the same procedure as performed by a competitive layer. Though, in place of updating only the winning neuron, all neurons within a certain neighborhood $N_{i*}(d)$ of the winning neuron are updated, using the Kohonen rule [27]. Specifically, all such neurons $i \in N_{i*}(d)$ are adjusted as follows:

$$ {}_i w(q) = {}_i w(q-1) + \alpha(p(q) - {}_i w(q-1)) $$

or

$$ {}_i w(q) = (1-\alpha)\, {}_i w(q-1) + \alpha p(q) \qquad (9) $$

Here the neighborhood $N_{i*}(d)$ restrains the indices for all of the neurons that lie inside a radius d of the winning neuron i^*.

$$ N_i(d) = \{ j, d_{ij} \le d \} \qquad (10) $$

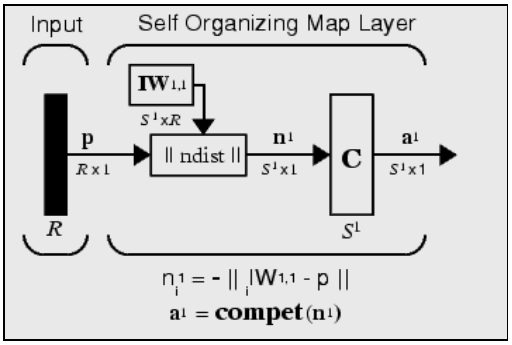


Fig. 2: The architecture for SOFM

D. Verification by Self-Organizing Map

We build a network having input vectors with two elements each presents on speaker features. This process defines variables used in learning phases. The default learning in a SOFM takes place in the batch mode (trainbuwb). The weight learning function for the SOM is learnsomb. First, the network detects the winning neuron for each input vector. Each weight vector in that case moves to the average position of all of the input vectors for which it is a winner or for which it is in the neighborhood of a winner. The distance that defines the size of the neighborhood is changed throughout training during two phases:

Ordering Phase, this phase lasts for the given number of steps. The neighborhood distance starts at a given initial distance, and decreases to the tuning neighborhood distance (1.0). As the neighborhood distance decreases over this phase, the neurons of the network typically order themselves in the input space with the same topology in which they are ordered physically.

Tuning Phase, this phase lasts for the rest of training phase. The neighborhood distance stays at the tuning neighborhood distance, which should contain only close neighbors, i.e., typically 1.0. The small neighborhood fine-tunes the network, while preserve the ordering learned in the previous phase stable.

As a result of training procedure the neurons have started to move toward the various training groups. Additional training is required to get the neurons closer to the various groups. The result is that neighboring neurons tend to have similar weight vectors and to be approachable to similar input vectors. In this paper the weights vectors taken from trained vectors by SOM are used for verification (WSOM approach). The decision is taken based on the following formula

$$S = 100 - [100 * \sqrt{(\sum (W(i,1) - (W(i,2))^2 / \sum (W(i,1))^2)}] \qquad (11)$$

Where S is the similarity percent between pattern weights vector $W(i,1)$ and a speaker signal needs to be verified weights vector $W(i,2)$. 70% is the empirical S threshold for deciding acceptance or rejection. In WP experiments, it was found that the recognition rates improved upon increasing the number of feature sets. However, the improvement implies a tradeoff between the recognition rate and extracting time. It is seen that the recognition rate has improved from 71.6% to 97.8%, but the number of feature sets has increased four times, from 32 to 128 [19]. On the other hand, the growth of extracting time indicated that the computational load has been burdened. In this paper, we investigate the use of SOM in by decreasing the feature vector dimensionality. It assists greatly in decreasing computational complexity by decreasing the feature vector dimensionality from 1020 to 64.

E. Verification by Gaussian Mixture Model

Gaussian Mixture Model GMM recently has become the dominant approach in text-independent speaker identification and verification. One of the influential attributes of GMMs is their capability to form smooth approximations to arbitrarily formed densities [31]. As a typical model based approach, GMM has been utilized to characterize speaker's voice in the form of probabilistic model. It has been reported that the GMM approach outperforms other classical methods for text-independent speaker recognition. We briefly review the GMM based speaker verification scheme:

Given

$$Y = \{Y_1 \; Y_2 .. Y_K\} \text{ where } Y = \{y_{t1} = T_1, y_{t2} = T_2, \ldots, y_{\mathrm{tj}} = T_j\}$$

is a sequence of T_j feature vectors in jth cluster R^j, the complete GMM for speaker model λ is characterized by the mean vectors, covariance matrices and mixture weights from all component densities. The parameters of the speaker model are denoted by

$$\lambda = \{p_{j,t} u_{j,t} \sum_{j,t} \}, i = 1, 2, \ldots, M_j \text{and} \, j = 1, 2, \ldots, K$$

Then, the GMM likelihood can be written as

$$p(Y \,|\lambda) = \prod_{t_1}^{T_1} p(y_{t_1} |\lambda) \ldots p(y_{t_K} |\lambda).$$

(12)

In this equation $p(y_{t_j} |\lambda)$ is the Gaussian mixture density for jth cluster and defined by a weighted sum of M_j component densities [22]. We use the Expectation Maximization (EM) algorithm to create an object of the Gaussian mixture distribution class restraining maximum likelihood estimates of the parameters in a Gaussian mixture model with k components for data in the n-by-d matrix X, where n is the number of observations and d is the dimension of the data. In this paper, verification is performed by building Gaussian mixture model by EM with 2 components of WP entropy and energy indices vectors of two speakers feature vectors GMMWPE. Then the GMM likelihood is used as the verification decision whether accept or reject. This is accomplished by determining empirical threshold for decision performing.

F. Verification by K-Mean Clustering Method

In this section, we introduce a brief outline of K-Means clustering algorithm and verification by this method.

Clustering in N dimensional Euclidean space R^N is the process of partitioning a given set of n points into a number, say K, of clusters based on some similarity metric which establishes a rule for assigning patterns to the domain of a particular cluster centroid as seen at Fig.3. Let the set of n points $\{x_1, x_2, \ldots, x_n\}$ be represented by the set S, and the *K* clusters is represented by $C_1, C_2, \ldots, C_K$ (Bandyopadhyay & Maulik, 2001). Then

$C_i \neq \phi$ for $i = 1, \ldots, K,$

$C_i \cap C_j = \phi$ for $i = 1, \ldots, K, j = 1, \ldots, K$, Where $j \neq i$,

$\bigcup_{i=1}^{K} C_i = S$.

K-Means [32,33] is one of the commonly used clustering techniques, which is an iterative hill climbing algorithm. It consists of the following steps:

1. Choosing *K* initial cluster centroids $z_1, z_2, \ldots, z_K$, randomly from the n points $\{x_1, x_2, \ldots, x_n\}$.
2. Assigning point $x_i, i = 1, 2, \ldots, K$ to cluster $C_j, j \in \{1, 2, \ldots, K\}$ where $\|x_i - z_j\| \leq \|x_i - z_p\|$, *p* =1,2,…,*k*, and $j \neq p$.
3. Calculating new cluster centroids: $z_i^*, z_2^*, \ldots, z_K^*$, where $z_i^* = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$, $i = 1, 2, \ldots K,$

where n_i is the number of elements belonging to cluster C_i.

4. If $z_i^* = z_i \quad \forall_i = 1,2,...,K$ then end. Otherwise continue from 2.

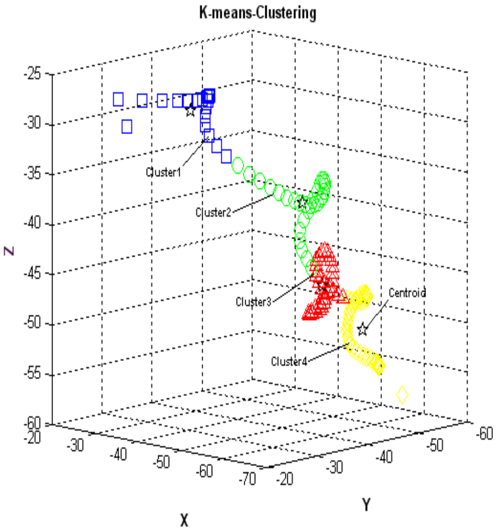


Fig.3: K-Means data clustering with K=4

K-means is a common clustering algorithm that has been used in a variety of application disciplines, such as image clustering and information retrieval, as will as speech and speaker recognition. Different types of clustering algorithms that are based on K-Means, are mentioned in [32, 33], such as the modified version for background knowledge, a genetic algorithm, and the syllable contour that is classified into several linear loci that serve as candidates for the tone-nucleus using segmental K-Means segmentation algorithm.

HERE IS AN INVESTIGATION OF A NEW SPEAKER VERIFICATION SYSTEM THAT BASED ON K-MEANS THREE TYPES ENTROPIES AND ENERGY INDICES FEATURES TAKEN FROM WAVELET PACKET TRANSFORM OF SPEECH SIGNALS. MORE SPECIFICALLY, THE PRESENTED VERIFICATION METHOD BY K-MEANS CLUSTERING CONSISTS OF TWO MAIN STAGES:

Partitions the points in the N1-by-P1 data matrix X1 (two WP features vectors for two speakers) into two clusters. Then we extract the two cluster centroid locations in the 2-by-P1 matrix consists of eight two elements columns. For each speaker four columns (8 coefficients) are preserved.

Partitions the points in the N2-by-P2 data matrix X2 (four WP features vectors of each speaker: three types entropies and energy indices) into four clusters. Then we extract coefficients as follows

Distances from each point to every centroid in the N-by-4 matrix D, afterwards we determine four coefficients: mean value, standard deviation, maximum and variance.

Four cluster centroid locations in the 4-by-P2 matrix C (16 coefficients).

Sums of point-to-centroid distances in the 1-by-4 vector M (4 coefficients).

The first 32 elements of N2-by-1 vector I containing the cluster indices of each point.

In total, 64 coefficients vector V are extracted by this method for each speaker. The verification decision is taken based on the Eq. 12:

$$S_{K-means} = 100 - [100 * \sqrt{(\sum (V1 - V2)^2 / \sum V1^2)}] \quad (13)$$

III. RESULTS AND DISCUSSION

A testing database was created from Arabic language. The recording environment is a normal office environment via PC-sound card, with spectral frequency 4000 Hz and sampling frequency 16000 Hz. These Arabic utterances are Arabic spoken digits from 0 to 15. In addition, each speaker read ten separated 30 seconds different Arabic texts. Total 29 individual speakers (19 to 40 years old) who are 19 individual male and 10 individual female spoken these Arabic words and texts for training by the SOM network. The total number of tokens considered for training was 725.

Experiments were conducted on a subset of our database consist of 19 male and 10 female speakers of different spoken words and texts. At first, feature vector was created by extracting WP entropies and energy indices from silence-removed data for each frame. Finally, verification process was performed using WSOM approach.

Speaker verification (SV) is a binary decision task to state whether a test utterance belongs to a speaker (target model) or not (hence, an outside imposter). Evaluations were carried out on the pool of 29 speakers, with the individual speaker features constructed using 100% of the data, and the imposter speaker model obtained from 100% of the utterances belonging to all speakers. In case of individual speaker to same speaker of different utterances verification (speaker-speaker system), 25 trials are applied for each speaker. In case of individual speaker to imposter verification (speaker-imposter system), 25 trials are also applied for each speaker. All our experiments were applied according to the text-independent system.

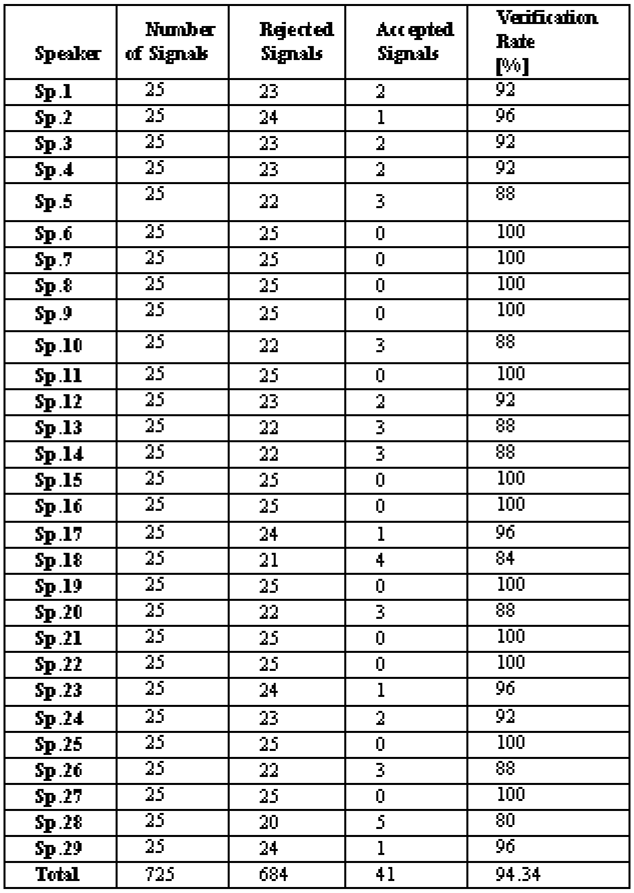
A single run of SV task consists of scoring test files against either the speaker model or imposter model. If the S score are greater than a threshold (see Fig.2), the test file is categorized as the target speaker, otherwise when the score is less than or equal to this certain threshold, the test file is classified as an outside imposter. The performance of presented verification system according to speaker-speaker verification system and speaker-imposter verification system for independent-text platform were reported at Table 1 and Table 2, respectively. In case of the speaker-speaker verification system 91.17% verification rate was accomplished. Better result was achieved (94.34%) in case of the speaker-imposter verification system.

Table 1**:** SOMWPE verification rate results for speaker-speaker system

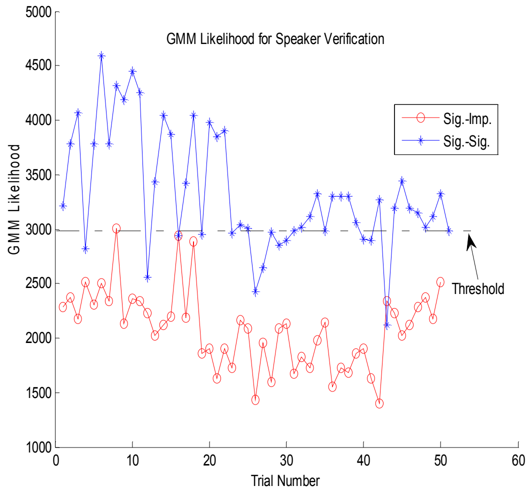
Speaker	Number of Signals	Accepted Signals	Rejected Signals	Verification Rate [%
Sp.1	25	24	1	96
Sp.2	25	21	4	84
Sp.3	25	22	3	88
Sp.4	25	23	2	92
Sp.5	25	24	1	96
Sp.6	25	22	3	88
Sp.7	25	25	0	100
Sp.8	25	24	1	96
Sp.9	25	21	4	84
Sp.10	25	23	2	92
Sp.11	25	25	0	100
Sp.12	25	23	2	92
Sp.13	25	24	1	96
Sp.14	25	20	5	80
Sp.15	25	24	1	96
Sp.16	25	23	2	92
Sp.17	25	23	2	92
Sp.18	25	20	5	80
Sp.19	25	25	0	100
Sp.20	25	18	7	72
Sp.21	25	22	3	88
Sp.22	25	24	1	96
Sp.23	25	23	2	92
Sp.24	25	24	1	96
Sp.25	25	22	3	88
Sp.26	25	22	3	88
Sp.27	25	25	0	100
Sp.28	25	21	4	84
Sp.29	25	24	1	96
Total	725	661	64	91.17

In the next experiment, the performances of the SOMWPE speaker verification systems in the speaker-speaker and speaker-imposter platforms were compared with the same of GMM and wavelet entropy method GMMWPE presented in section 3.4 and K-Means and wavelet entropy method KMWPE presented in section 3.3, under the recorded database. Fig. 4 demonstrates fifty trials verification results (S or likelihood) obtained for two speakers in the speaker-speaker and speaker-imposter systems for the GMMWPE, KMWPE and SOMWPE based on certain thresholds. These thresholds were determined empirically. The results of these experiments via recorded database are summarized in Table 3. These results indicate that under similar conditions, SOMWPE provides a better platform for speaker verification than GMMWPE and KMWPE. Moreover, the speaker-imposter system provides more accurate results than the speaker-speaker results for the all three methods.

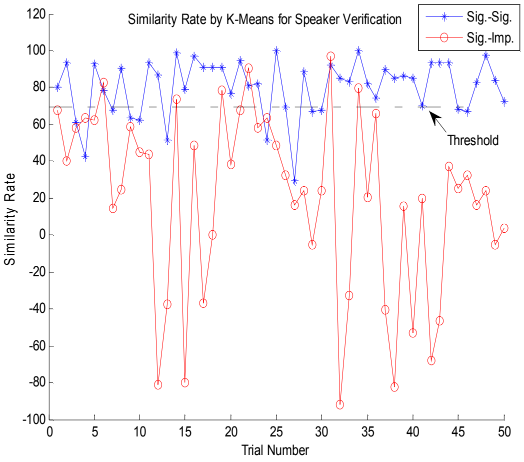
Table2: SOMWPE verification rate results for speaker-imposter system



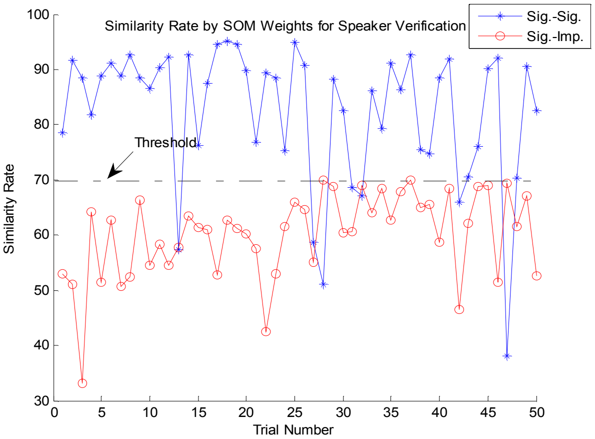
Speaker	Number of Signals	Rejected Signals	Accepted Signals	Verification Rate [%]
Sp.1	25	23	2	92
Sp.2	25	24	1	96
Sp.3	25	23	2	92
Sp.4	25	23	2	92
Sp.5	25	22	3	88
Sp.6	25	25	0	100
Sp.7	25	25	0	100
Sp.8	25	25	0	100
Sp.9	25	25	0	100
Sp.10	25	22	3	88
Sp.11	25	25	0	100
Sp.12	25	23	2	92
Sp.13	25	22	3	88
Sp.14	25	22	3	88
Sp.15	25	25	0	100
Sp.16	25	25	0	100
Sp.17	25	24	1	96
Sp.18	25	21	4	84
Sp.19	25	25	0	100
Sp.20	25	22	3	88
Sp.21	25	25	0	100
Sp.22	25	25	0	100
Sp.23	25	24	1	96
Sp.24	25	23	2	92
Sp.25	25	25	0	100
Sp.26	25	22	3	88
Sp.27	25	25	0	100
Sp.28	25	20	5	80
Sp.29	25	24	1	96
Total	725	684	41	94.34



(a)



(b)



(c)

Fig 4: Fifty verification trials obtained for two speakers in the speaker-speaker and speaker-imposter systems for the (a) GMMWPE, (b) KMWPE and (c) SOMWPE

Table 3: Speaker verification rate results for GMMWPE, KMWPE and SOMWPE

Verification Method	Number of Signals	Sig.-Sig. System Ver. Rate [%]	Sig.-Imp. Ver. System [%]	Verification Rate [%]
GMMWPE	725	87.23	93.76	90.49
KMWPE	725	74.65	77.53	76.09
SOMWPE	725	91.17	94.34	92.75

Table 4: Speaker verification rate under the condition of AWGN in speaker-imposter system for GMMWPE, KMWPE and SOMWPE

Verification Method	0 dB	5 dB	10 dB	15 dB	20 dB
GMMWPE	18	40	72	84	84
KMWPE	37	68	72	76	80
SOMWPE	20	92	92	84	84

Table 5: Speaker verification rate under the condition of AWGN in speaker-speaker system for GMMWPE, KMWPE and SOMWPE

Verification Method	0 dB	5 dB	10 dB	15 dB	20 dB
GMMWPE	12	20	40	64	76
KMWPE	56	64	68	52	78
SOMWPE	80	92	96	100	100

Subsequent to assessment in the normal condition, we conducted experiments to assess the speaker verification system in the speaker-imposter platform under abnormal noisy. The speaker-imposter system is more appropriate for such experiment, because a big amplitude noise added to the model and verified speech signals leads to artificial high similarity. To implement this experiment, additive white Gaussian noise (AWGN) was added to the clean speech samples of our recorded database with SNR values of 0, 5, 10, 15 and 20 dB. Which were then applied to GMMWPE, KMWPE and presented method SOMWPE. In this experiment, the performances of the speaker verification systems in speaker-imposter were compared via false-positive error (FPE), this error happens when the result of verification is acceptance in case of imposter speaker. Then verification rate is calculated from FPE (100 minus FPE percent). The results of these experiments are summarized in Table 4. These results indicate that the proposed verification system tackles additive white Gaussian noise condition more robustly than other speaker verification systems in case SNR values of 5, 10, 15 and 20 dB. In case of SNR value of 0 dB KMWPE showed better results.

In the next experiment, AWGN was added to the verified signal only. To implement this experiment, white noise was added to the clean verified speech samples with SNR values of 0, 5, 10, 15 and 20 dB. We conducted this experiment to assess the speaker verification system in the speaker-speaker platform. The speaker-speaker system is more appropriate for such experiment because a big amplitude noise added to the verified speech signals leads to artificial dissimilarity. The obtained results from this experiment are demonstrated in Table 5. These results indicate that the proposed verification system tackles AWGN condition more robustly and outperforms GMMWPE and KMWPE speaker verification systems. The artificial dissimilarity causes false-negative error (FNE) in the speaker-speaker system. This error appears when the result of verification is rejection when a test utterance belongs to a speaker (target model). The FNE results are demonstrated in Fig.5. The interpretation of Fig.5 concludes that the proposed verification system with AWGN is more robust to this noise than the other proposed methods.

IV. Conclusion

Weight vector of SOM based speaker verification system is proposed in this paper. This system was developed using a wavelet packet feature extraction method. In this study, effective feature extraction method for text-independent system is developed, taking in consideration that the computational complexity is very crucial issue. Three types of entropy coefficients of WPT in conjunction with energy indexes of WPT are utilized. The experimental results on a subset of recorded database showed that feature extraction method proposed in this paper is appropriate for text-independent verification system. Two other verification methods are proposed GMMWPE and KMWPE.

The results of the experiments conducted in this paper demonstrated a better performance of SOMWPE in text-independent verification task. Finally, the developed speaker verification system was employed with data obtained under abnormal conditions where AWGN noisy was added. In this case it was observed that the SOMWPE system is generally more noise-robust than similar systems with GMM and K-Means (GMMWPE and KMWPE).

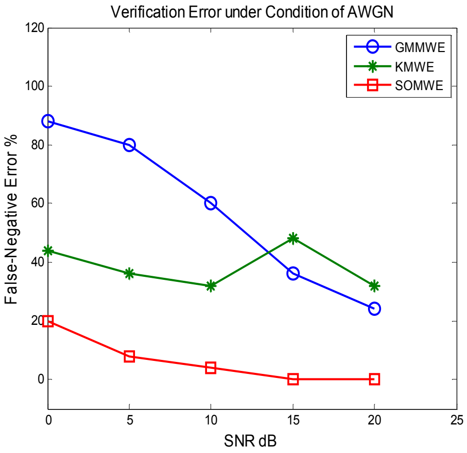


Fig.5: The FNE results for GMMWPE, KMWPE and SOMWPE

Another major contribution of this research is the development of a less computational complexity speaker verification system with weight vector of SOM capable of dealing with abnormal conditions for relatively good degree.

References

[1] Nemati, S., & Basiri, M. E. (2010), Text-independent speaker verification using ant colony optimization-based selected features. *Expert Systems with Applications*, doi:10.1016/j.eswa.2010.07.011.

[2] Xiang, B., & Berger, T. (2003). Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. *IEEE Transactions on Speech and Audio Processing*, 11(5).

[3] Lamel, L.F., Gauvain J.L. (2000), Speaker verification over the telephone, *Speech Communication* 31, 141-154

[4] Kwon, S., & Narayanan, S. (2002). Speaker change detection using a new weighted distance measure. *In Proceedings of international conference on spoken language processing (ICSLP 2002)*, Denver, CO (pp. 2537–2540).

[5] Lapidot, I., Guterman, H., & Cohen, A. (2002). Unsupervised speaker recognition based on competition between self-organizing maps. *IEEE Transactions on Neural Networks*, 13(2), 877–887.

[6] Martin, A., & Przybocki, M. (2001). Speaker recognition in a multi-speaker environment. In *Proceedings 7th European conference on speech communication and technology (Eurospeech 2001)*, Aalborg, Denmark (pp. 787–790).

[7] Furui, S., Itakura, F., Saito, S. (1972). Talker recognition by longtime averaged speech spectrum. *Trans. IECE* 55-A (1), 549-556.

[8] Newman, M., Gillick, L., Ito, Y., McAllister, Peskin, B., 1996. Speaker veri®cation through large vocabulary continuous speech recognition. In: Proceedings ICSLP'96, Philadephia, PA, pp. 2419-2422.

[9] Sadeghi A., Homayounpour M., Samani A. (2010), A real-time trained system for robust speaker verification using relative space of anchor models, *Computer Speech and Language*, 24, 545–561.

[10] Young, S. (1996). A review of large-vocabulary continuous-speech recognition. IEEE Signal Proc. Magazine 13 (5), 45–57.

[11] Rabiner, L., Juang, B.H. (1993). Fundamentals of Speech Recognition. Prentice Hall, New Jersey.

[12] Avci, E., Akpolat, Z.H. (2006), Speech recognition using a wavelet packet adaptive network based fuzzy inference system. *Expert Systems with Applications* 31 (3), 495–503.

[13] Merlin, T., Bonastre, J.F., Fredouille, C. (1999). Non directly acoustic process for costless speaker recognition and indexation. In: *Workshop on Intelligent Communication Technologies and Applications*.

[14] Kuhn, R., Junqua, J.-C., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. IEEE Transactions on Speech Audio Process 8 (6), 695–707.

[15] Thyes, O., Kuhn, R., Nguyen, P., Junqua, J.-C. (2000). Speaker identification and verification using eigenvoices. In: *International Conference on Spoken Language Processing (ICSLP)*, vol. 2. pp. 242–245.

[16] Mami, Y., Charlet, D. (2002). Speaker identification by location in an optimal space of anchor models. *International Conference on Spoken Language Processing (ICSLP)*, vol. 2. pp. 1333–1336.

[17] Mami, Y., Charlet, D. (2003). Speaker identification by anchor models with PCA/LDA post-processing. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. pp. 180–183.

[18] Mami, Y., Charlet, D. (2006). Speaker recognition by location in the space of reference speakers. *Speech Communication* 48 (2), 127–141.

[19] Wu, J.-D. & Lin B.-F. (2009), Speaker identification using discrete wavelet packet transform technique with irregular decomposition, *Expert Systems with Applications* 363136–3143.

[20] Zheng, H., Li, Z., & Chen, X. (2002). Gear fault diagnosis based on continuous wavelet transform. *Mechanical Systems and Signal Processing*, 16, 447–457.

[21] Wu J-D., Ye S-H., S-H. (2009), Driver identification based on voice signal using continuous wavelet transform and artificial neural network techniques, Expert Systems with Applications 36, 1061–1069.

[22] Lung S.-Y (2007), Efficient text independent speaker recognition withwavelet feature selection based multilayered neural network using supervised learning algorithm, *Pattern Recognition* 40 , 3616 – 3620

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

[24] Visser, E., Otsuka, M., & Lee, T. (2003). A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. *Speech Communication*, 41, 393–407.

[25] Kosko & Bart (1992). Neural networks and fuzzy systems: A dynamical approach to machine intelligence. *Englewood Cliffs*, NJ: Prentice Hall.

[26] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59–69.

[27] Kohonen, T. (1995). Self-organizing maps. Berlin: Springer.

[28] Avci, E. (2007), A new optimum feature extraction and classification method for speaker recognition: GWPNN, *Expert Systems with Applications* 32, 485–498.

[29] Coifman, R. R., & Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2), 713–718.

[30] MATLAB 5.3 version Wavelet Toolbox, MathWorks Company.

[31] Reynolds D.A., Rose R.C., Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. *Speech Audio Process*. 3 (1) (1995) 72–83.

[32] Bellot, P., & El-Beze, M. (1999), A clusterin method for information retrieval (Technical Report IR-0199). *Laboratoire d'Informatique d'Avignon*. France.

[33] Wagsta K., Cardie C. (2001), Constrained K- means Clustering with Background Knowledge, *Proceedings of the Eighteenth International Conference on Machine Learning*, p. 577-584, 2001