J. King Saud Univ., Vol. 17, *Comp. & Info. Sci*., pp. 23-41 (A.H. 1425/2004)

Robust Text-independent Speaker Recognition with Short Utterance in Noisy Environment Using SVD as a Matching Measure

Rabah W. Aldhaheri[*] and Fuad E. Al-Saadi[**]
[*]Department of Electrical and Computer Engineering, King Abdulaziz University, P.O.Box 80204, Jeddah 21589, Saudi Arabia
[**]Department of Communication, Jeddah College of Electronics and Communication, P.O.Box 16947, Jeddah 21474, Saudi Arabia

(Received 22 September 2003; accepted for publication 11 February 2004)

Abstract. A new technique for text-independent speaker recognition for noisy speech is presented. This technique is based on finding the ratio of the singular values of the feature vectors of the unknown speaker and each of the *N* reference features stored in the constructed database. The i^{th} reference feature that gives the largest ratio is considered the feature of the unknown speaker.

An overall correct recognition accuracy of 94% for clean speech and 32% for noisy speech of 0 dB SNR was obtained. A further step was conducted to enhance the noisy features by series expansion. The improvement in the recognition rate using the proposed SVD-based algorithm is compared with other distance measure algorithms. It is found that the proposed technique when cepstral features are used outperforms the conventional matching measure such as the Euclidean, the Weighted and the Mahalonobis distances, respectively.

1. Introduction

Speaker recognition is the process of automatically recognizing the identity of the speaker on the basis of information obtained from his/her speech waves. This technique will make it possible to verify the identity of persons accessing systems, that is, access control by voice, in various services. These services include voice dialing, banking transactions over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information areas, and remote access to computers. Speaker recognition can be divided into speaker identification and speaker verification. Speaker identification is the process of identifying a speaker from a group of *N* registered speakers. Speaker verification is the process of accepting or rejecting a person claimed identity from his voice.

In other words, speaker identification system attempts to answer the question, "Who are you?" Speaker verification system attempts to answer the question, "Are you whom you claim to be?"

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former requires the speaker to provide utterances of the key words or sentences having the same text, whereas the latter does not rely on a specific text being spoken. The text-dependent methods are usually based on template matching techniques in which the time axis of an input speech sample and each reference template or reference model or registered speakers are aligned, and the similarity between them, accumulated from the beginning to the end of the utterance, is calculated. The structure of text-dependent recognition systems is, therefore, rather simple. Since this method can directly exploit the voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent method.

An important step in the speaker identification process is how to extract sufficient information for good discrimination, and at the same time, the size of this information should be amenable to effective modeling. This process is called *feature extraction*. After feature extraction, a classification technique is used to compare between the test feature and the registered features in the database.

Different techniques are used for classification and we can split them into two broad types: template matching and probabilistic algorithms [1-3]. In the template matching, or termed statistical features averaging, we mean the comparison of an average computed on test data to a collection of stored averages developed for each of the speakers in the database [4-8]. In probabilistic, the speakers are modeled by probability distribution rather than by average features and in this case a log-likelihood score is computed instead of distance measure [9-10]. The common techniques used in this type are: Hidden Markov Model (HMM) [8-11], Artificial Neural Network (ANN) [12-13], Linear Vector Quantization (LVQ) and others [3,14-15].

The matching algorithms are much simpler and less expensive than the probabilistic algorithms. Moreover, the time required for training the models is much longer. But, the recognition accuracy is better to some extent. In our study, we consider the first type, and the comparison is made with the same matching measure techniques.

In previous research on speaker recognition, researchers used the same features used in speech recognition such as linear prediction and cepstral coefficients [1-3]. Atal [4] studied the effectiveness of prediction coefficients, impulse response, autocorrelation, and cepstrum coefficients for automatic speaker identification and verification. He concluded that the cepstrum coefficients give better overall recognition accuracy. The weighted cepstral distance measure for a speaker-independent isolated

word recognition system using dynamic time warping was tested in [5]. It is found that the weighted cepstral distance outperformed the Euclidean cepstral distance and the log-likelihood distance measure. In [6], a comparison between four distance measures for text-independent speaker identification was presented and it was found that the weighted Euclidean distance performed better than the others. On the other hand, the Mahalonobis distance measure was inferior to the other methods despite the fact that it was computationally more complex. In [7], different distance measures were compared for Multidimensional Autoregressive (MAR) model instead of the one dimensional that is often used. It is shown that the optimal order of AR process is approximately 2 or 3. In the previous techniques [4-7], the Euclidean, Mahalonobis and/or the weighted distances are used as a pattern matching.

In [8], the SVD of the energy and the zero crossing are used as a pattern matching for text-dependent speaker identification. Only one sentence, uttered by 3 male and 2 female speakers, is used in both the training and the test sessions. An overall identification score of 80% was obtained for clean speech.

In [9], two identification algorithms, based on LPC and LPC-cepstral feature extractors, followed by a Continuous Density Hidden Markov Model (CD-HMM) classifier, have been implemented and tested on the Italian database. This database consists of 360 phone calls made by 20 speakers. The performance of closed set text-independent speaker identification is evaluated. It is found that the LPC-cepstral based system performs better than the LPC-based one.

Although speaker recognition has reached the state of launching commercial products, operational systems still face the problem of maintaining high recognition performance in adverse environment. The degradation to recognition performance is typically attributed to the mismatch between training and testing conditions.

Robust recognition methods include signal enhancement techniques as a front-end and/or feature space transformations that reduce variability due to noise are addressed in [11, 16-18]. The effect of noise is still an open problem and some extra work in this direction must be conducted to overcome this problem and this is what we will try to do in this paper.

In this paper, a robust closed set text-independent speaker identification algorithm based on LPC and/or cepstral coefficients is presented. The pattern matching used here depends on the ratio of singular values of the average test feature vector $\boldsymbol{x}$ and each of the N reference features $\boldsymbol{z}^{(i)}$ stored in the database. The i^{th} reference feature that gives the largest ratio is considered closest to the unknown speaker.

The robustness of the proposed technique is evaluated, in terms of the recognition scores, by adding white noise to the test speech. The overall correct accuracy varies between 94% for clean speech (recorded in an office environment) and 32% for noisy speech of 0 dB SNR. The experimental results show that the template-matching algorithm based on SVD is superior to those algorithms based on distance metrics such as Euclidean, Weighted and Mahalonobis distances. The result of this paper is an extension to a previous work [19], where here, we doubled the database population size. Also, the time duration of the test utterance is investigated and how can this duration affects the recognition rate when it is short. Moreover in this paper, we propose an algorithm to enhance the noisy features of the test speakers using series expansion.

This paper is organized as follows: In Section 2 some preliminaries regarding the LPC and cepstral coefficients are presented. Also, in this section, we showed that the coefficients vary nonlinearly with the noise power. In Section 3, the noisy coefficients are enhanced or modified by Taylor series expansion to obtain an estimate of almost free noise coefficients. The singular value decomposition as a matching measure between the test and the template vectors is presented in Section 4. Moreover, in this section a simple procedure is presented to compute the singular values. In Section 5, the proposed algorithm is evaluated on a constructed database. A comparison between the proposed algorithm and the other distance measures algorithms is given also in this section. Finally, Section 6 presents the conclusions.

2. Preliminaries and Problem Formulation

Figure 1 shows the basic structure of speaker identification system. The speech signal is band-limited with a 6^{th} order Butterworth bandpass filter, with [60 Hz – 4 kHz] passband. Then it is sampled at a rate of 8 kHz with 8 bits/sample. This sampled signal is processed by high frequency pre-emphasis filter $(1-0.95z^{-1})$, and then partitioned into frames of 32 ms using Hamming window with 50% overlap. From the experiments that we conducted in this study, we found that the energy and zero crossing result in satisfactory classification of the speech frame into voiced and unvoiced. In the feature extraction, the LPC and/or the LPC derived cepstral coefficients are used as the speaker specified feature. To determine the LPC coefficients, the clean speech $s(n)$ can be modeled as an Autoregressive (AR) model:

$$s(n) = G\, e(n) - \sum_{i=1}^{p} a_i s(n-i) \qquad (1)$$

Or equivalently, its z-transform is:

$$S(z) = \frac{G\ E(z)}{1 + \sum_{i=1}^{p} a_i z^{-i}} \tag{2}$$

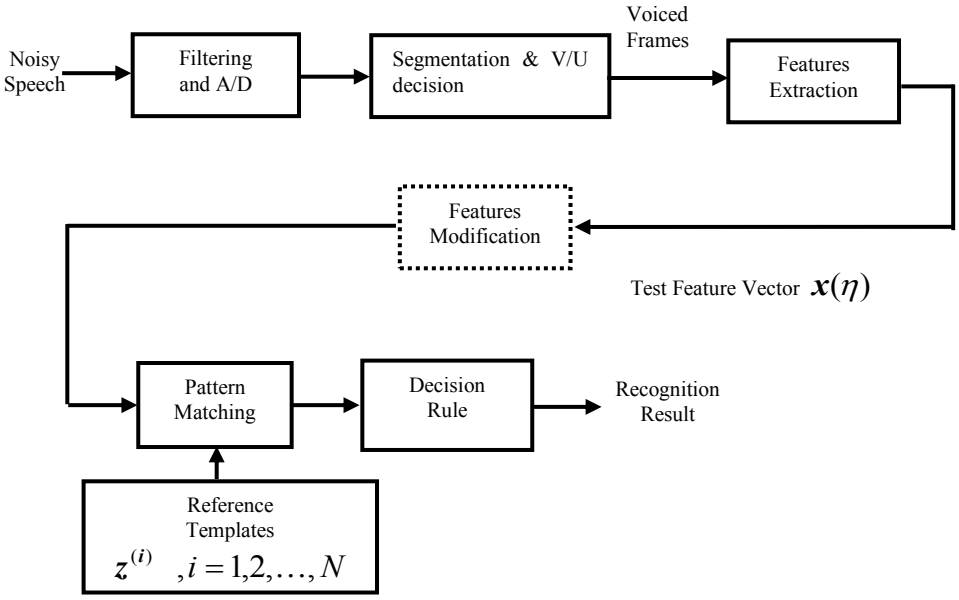

Fig. 1. Basic structure of speaker identification system.

Where p is the prediction order, a_i, $i = 1,\ldots,p$ are the linear prediction coefficient; $e(n)$ is the excitation and G is a gain scaling factor.

The LPC coefficients can be obtained by using standard LPC analysis [1,2] such as the autocorrelation method. Thus, the LPC coefficients are determined by solving the p linear equations, which can be written in a matrix form as:

$$\begin{bmatrix} r_0 & r_1 & r_2 & \cdots\cdots & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots\cdots & r_{p-2} \\ r_2 & r_1 & r_0 & \cdots\cdots & r_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots\cdots & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_p \end{bmatrix} \tag{3}$$

Where r_τ are the time-averaged estimates of the autocorrelation at lag τ, and it can be expressed as:

$$r_\tau = \sum_{n=0}^{M-1-\tau} s(n)\, s(n+\tau), \qquad \tau = 0,1,\ldots,p-1 \tag{4}$$

where M is the frame size. Now equation (3) can be solved efficiently using Durbin's algorithm [1-2].

The cepstral vector $\boldsymbol{c} = \lfloor c_1 \quad c_2 \quad \cdots \quad c_p \rfloor$ can be obtained by solving the recursive equation:

$$c_n = a_n + \frac{1}{n}\sum_{i=1}^{n-1} (n-i)\, c_{n-i}\, a_i \quad , n = 1,2,\ldots,p \tag{5}$$

Now, suppose that the noisy speech $x(n)$ is composed of the original clean speech $s(n)$ and an additive uncorrelated white noise $w(n)$ with zero mean and power ζ, then:

$$x(n) = s(n) + w(n) \tag{6}$$

The autocorrelation of $x(n)$ is defined as:

$$r_x(\tau\tau = \sum_{n=0}^{M-1-\tau} x(n)\, x(n+\tau) = \begin{cases} r_0 + \zeta & \text{for } \tau = 0 \\ r_\tau & \text{for } \tau \neq 0 \end{cases} \tag{7}$$

Therefore the autocorrelation matrix of the noisy speech is:

$$R_x(\zeta\zeta = \begin{bmatrix} r_0 + \zeta & r_1 & r_2 & \cdots\cdots & r_{p-1} \\ r_1 & r_0 + \zeta & r_1 & \cdots\cdots & r_{p-2} \\ r_2 & r_1 & r_0 + \zeta & \cdots\cdots & r_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots\cdots & r_0 + \zeta \end{bmatrix} = \lfloor R + \zeta I \rfloor \tag{8}$$

The LPC vector of the noisy speech,

$$a(\zeta) = \lfloor a_1(\zeta) \quad a_2(\zeta) \quad \ldots \quad a_p(\zeta) \rfloor \tag{9}$$

is determined by solving the linear equation:

$$[R + \zeta I]\, a(\zeta) = r \tag{10}$$

where $\boldsymbol{R}$ is a *p*-by-*p* matrix and $\boldsymbol{r}$ is a *p*-by-1 vector as defined in (3).
Again equation (10) can be solved by Durbin's algorithm. Similarly, the noisy cepstral vector is given by:

$$c_n(\zeta) = a_n(\zeta) \frac{1}{n} \sum_{i=1}^{n-1} (n-i)\, c_{n-i}(\zeta)\, a_i(\zeta) \qquad , n = 1,2,\ldots,p \tag{11}$$

In this paper, we will assume that the reference feature vectors (LPC or cepstral coefficients) are noise free, while the test feature is noisy.

3. Noisy Feature Modification Using Taylor Series Expansion

In this section the noisy feature $a(\zeta)$ and $c(\zeta)$ derived in the previous section is enhanced from noise by estimating the almost free noise feature. This is done as follows:

First, let us rewrite equation (8) as:

$$R_x(\zeta) = \begin{bmatrix} \sigma_s + \zeta & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & \sigma_s + \zeta & r_1 & \cdots & r_{p-2} \\ r_2 & r_1 & \sigma_s + \zeta & \cdots & r_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & \sigma_s + \zeta \end{bmatrix} = \begin{bmatrix} \sigma_s(1+\eta) & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & \sigma_s(1+\eta) & & \cdots & r_{p-2} \\ r_2 & r_1 & \sigma_s(1+\eta) & \cdots & r_{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \ldots & \sigma_s(1+\eta) \end{bmatrix} \tag{12}$$

where $\sigma_s = r_0$, the speech signal power and $\eta = \dfrac{\zeta}{\sigma_s}$ the noise to signal ratio (NSR).

Thus, the LPC vector $\boldsymbol{a}$ is a function of NSR. Now both the correlation matrix $\boldsymbol{R}_x(\zeta)$ and the vector $\boldsymbol{a}$ can be rewritten as a function of η as follows:

$$R_x(\eta)\, a\,(\eta) = r \tag{13}$$

Now, if $\boldsymbol{a}(\eta)$ is differentiable with respect to η , then Taylor series expansion of $\boldsymbol{a}(\eta)$ in the neighborhood of $\eta = 0$ is given by:

$$\mathrm{a}(\eta(= \mathrm{a}(0) + \eta\,\mathrm{a}^{[1]}(0) + \frac{1}{2!}\eta^2\,\mathrm{a}^{[2]}(0) + \frac{1}{3!}\eta^3\mathrm{a}^{[3]}(0) + \mathrm{O}(\eta^4) \tag{14}$$

where $\boldsymbol{a}(0)$ is the estimated feature vector at $\eta = 0$, $\boldsymbol{a}^{[k]}(0)$ is the k^{th} derivative of $a(\eta)$ when $\eta = 0$ and $O(\eta^4)$ is the error of order 4 .

To find the derivative terms in equation (14), take the differentiation of both side of equation (13). Now, since $\frac{d}{d\eta}\boldsymbol{R}_{\boldsymbol{x}}(\eta) = \sigma_{\boldsymbol{s}}\boldsymbol{I}$ and $\frac{d\,\boldsymbol{r}}{d\eta} = 0$, we have

$$\mathrm{R}_{\mathrm{x}}(\eta)\frac{\partial\,\mathrm{a}(\eta)}{\partial\,\eta} = -\sigma_{\mathrm{s}}\;\mathrm{a}(\eta) \tag{15}$$

Repeat the differentiation of both sides of (15) to obtain:

$$\mathrm{R}_{\mathrm{x}}(\eta)\;\frac{\partial^2\,\mathrm{a}(\eta)}{\partial\eta^2} = -2\sigma_{\mathrm{s}}\;\frac{\partial\,\mathrm{a}(\eta)}{\partial\,\eta} \tag{16}$$

The k^{th} order can be computed as:

$$\mathrm{R}_{\mathrm{x}}(\eta)\,\mathrm{a}^{[\mathrm{k}]}(\eta) = -\mathrm{k}\,\sigma_{\mathrm{s}}\;\mathrm{a}^{[\mathrm{k}-1]}(\eta) \tag{17}$$

Thus, the derivatives at $\eta = 0$ can be computed recursively. Notice again that equation (17) can be solved efficiently using Durbin's algorithms. Now, the Taylor series can be approximated as:

$$\mathrm{a}(\eta) = \mathrm{a}(0) + \eta\;\mathrm{a}^{[1]}(\eta)\Big|_{\eta=0} + \frac{1}{2!}\eta^2\;\mathrm{a}^{[2]}(\eta)\Big|_{\eta=0} + \frac{1}{3!}\eta^3\;\mathrm{a}^{[3]}(\eta)\Big|_{\eta=0} + \mathrm{O}(\eta^4)$$

Thus, the estimated vector, $\widetilde{a}$ is given by:

$$\widetilde{\mathrm{a}} = \mathrm{a}(0) \approx \left[\mathrm{I} - \eta\,\sigma_{\mathrm{s}}\;\mathrm{R}^{-1} + \eta^2\;\sigma_{\mathrm{s}}^2\;\mathrm{R}^{-2} - \eta^3\;\sigma_{\mathrm{s}}^3\;\mathrm{R}^{-3}\right]^{-1}\;\mathrm{a}(\eta) \tag{18}$$

Equation (18) is the formula to modify the prediction coefficients up to the third order with respect to the noise to signal ratio, η . Higher order is possible, but it is found that the third order gives sufficient approximation.

Now, the estimated cepstral coefficients $\widetilde{\boldsymbol{c}} = \left[\widetilde{c}_1 \quad \widetilde{c}_2 \quad \cdots \quad \widetilde{c}_{\boldsymbol{p}}\right]$ are given by:

$$\tilde{c}_n = \tilde{a}_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)\,\tilde{c}_{n-i}\,\tilde{a}_i \qquad n = 1,2,\ldots,p \tag{19}$$

Where $\tilde{a}_i$ is the i^{th} element of the vector, $\tilde{\boldsymbol{a}}$. These modified LPC and cepstral features are used as a test features and the recognition rate are recalculated and the result is compared with the case of unmodified features.

4. Singular Value Decomposition (SVD) as A Matching Measure

In this section we will show how the singular value decomposition is used as a measure of matching instead of the conventional distance measure.

For an $m \times n$ real matrix $\boldsymbol{A}$ of rank r, the SVD is defined as:

$$A = U\ A\ V^T = \sum_{j=1}^{r} \sigma_j\ u_j\ v_j^T \tag{20}$$

where U and V are orthogonal matrices of dimensions $m \times m\ and\ n \times n$, respectively. The singular values, σ_j, are ordered in a descending order, $\sigma_1 \geq \sigma_2 \geq ..... \geq \sigma_r > 0$. The column vectors $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the j^{th} left and right singular vectors, respectively.

In our experiment, let us define $\boldsymbol{A}^{(i)}$ as:

$$A^{(i)} = \left[z^{(i)} \quad x(\eta(\right], \quad i = 1, 2, ..., N \tag{21}$$

where N is the number of references, $\boldsymbol{z}^{(i)}$ represents the i^{th} reference speaker features $a\ or\ c$ and $\boldsymbol{x}(\eta)$ represents the test speaker features $\boldsymbol{a}(\eta)$ or $\tilde{\boldsymbol{a}}$ and $\boldsymbol{c}(\eta)$ or $\tilde{\boldsymbol{c}}$.

Since $\boldsymbol{A}^{(i)}$ is a $p \times 2$ matrix, the singular values are computed rather simpler than what we did in [19], in which we applied the general algorithm for computing the SVD [20]. Here, we compute the singular values as follows:

First, let us drop the supscript i from the vector z and the matrix $\boldsymbol{A}$. Thus, $\boldsymbol{A} = \left[z \quad \boldsymbol{x}(\eta)\right]$.

Assume that,

$$\|z\| = \|\boldsymbol{x}(\eta)\| = L \tag{22}$$

Where $\|.\|$ denotes the norm. Notice that equation (22) can always be met by scaling the two vectors.

$$\sigma_j^2 = \lambda_j(A^T A), \qquad j = 1, 2 \tag{23}$$

where $\lambda_j(\cdot)$ denotes the j^{th} eigenvalue . Therefore,

$$\sigma_1^2 = L^2 + z^T x(\eta) \tag{24}$$

and

$$\sigma_2^2 = L^2 - z^T x(\eta) \tag{25}$$

From equations, (24) and (25)

$\frac{\sigma_1^2}{\sigma_2^2} = \frac{1+\cos\theta}{1-\cos\theta}$, where θ is the angle between the two vectors, z and $x(\eta)$.

Since $\sigma_1 \geq \sigma_2 > 0$, θ takes values over the range $\left[0, \frac{\pi}{2}\right]$.

Thus the ratio,

$$\rho = \frac{\sigma_1}{\sigma_2} = \sqrt{\frac{1+\cos\theta}{1-\cos\theta}} = \cot(\theta/2) \tag{26}$$

The ratio ρ defined by equation (26) is calculated for each reference vector, $z^{(i)}$, i.e. $\rho^{(i)}$ for all i. The decision rule that we have considered for classification is to find, the i^{th} speaker that maximizes the following function:

$$\underset{i}{\operatorname{argmax}}(\rho^{(i)}) \tag{27}$$

5. Experimental Results

The performance of the SVD matching algorithm is tested on a constructed database of twenty speakers (four females and sixteen males). Those speakers are used in the training data as reference templates. The training data consists of three sessions. Each session contains 4 different Arabic sentences, recorded through a desktop microphone at approximate nominal time interval of 2-3 weeks, the duration of the sentences is about 3-6 seconds. In the test, each of the twenty speakers utters 20 different, from the one recorded in the training, Arabic sentences of duration of 1-3 seconds. Each sentence is segmented into a frame of 32 ms with (16 ms overlap). The voiced frames are retained and LPC and cepstral coefficients are extracted from these frames. The average test feature vector $x(\eta)$ is computed for each sentence. Thus, twenty test vectors are computed for each speaker for clean speech (an office environment) and for the noisy speech of 0 to 20 dB SNR. The noisy speech is obtained manually by adding white noise to the clean speech. The total number of tests is, therefore, 400 tests for clean speech and these numbers of tests are repeated at different noise power (SNR of 0 to 20 dB). Notice here that, based on the background noise

associated with the recording process, the clean speech is considered to be of 30 dB SNR.

The proposed algorithm is compared with other template-matching algorithms such as Euclidean, Weighted and Mahalonobis distances. The Euclidean, Weighted and Mahalonobis distances to the i^{th} reference speaker are defined, respectively as

$$d_{ED}^{(i)} = \sqrt{\sum_{j=1}^{p} (z_j^{(i)} - x_j(\eta))^2} \tag{28}$$

$$d_{WD}^{(i)} = \sqrt{\sum_{j=1}^{p} (z_j^{(i)} - x_j(\eta))^2 / w_j} \tag{29}$$

$$d_{MD}^{(i)} = \sqrt{(z^{(i)} - x(\eta))^T \Sigma^{-1} (z^{(i)} - x(\eta))} \tag{30}$$

Where $z_j^{(i)}$ and $x_j(\eta)$ are the j^{th} element of the feature vectors $\boldsymbol{z}^{(i)}$ and $\boldsymbol{x}(\eta)$, respectively. w_j is the variance of the j^{th} element of $\boldsymbol{z}^{(i)}$ and $\boldsymbol{\Sigma}$ is the *p*-by-*p* covariance matrix of the template features. The test sentences duration is made short intentionally because it is known that template matching algorithm depends on averaging the feature vectors extracted from the utterance. As it is known, the accuracy of the average-estimate is dependent on the utterance duration. The recognition rate would have been better than what we obtained in this paper if we had increased the duration of the utterance. It is worth to notice that the computational complexity of Mahalonobis is more than the other algorithms. Yet, the proposed SVD-based algorithm gives better result for the noisy speech when cepstral coefficients are used.

Table 1 shows the overall correct recognition using Euclidean distance (ED), Weighted distances (WD), Mahalonobis distance (MD) and the SVD-based algorithm. Notice that the Mahalonobis distance is performed better in case of LPC coefficients while the proposed algorithm outperforms the other algorithms in case of cepstral coefficients. The difference in performance gets better in favor of the SVD-based algorithm as the noise power increases.

Table 1. Results of the text-independent experiments using Euclidean distance (ED), weighted distance (WD), mahalonobis distance (MD) and SVD-based algorithm for clean and noisy speech

	Recognition rate using ED		Recognition rate using WD		Recognition rate using MD		Recognition rate using SVD	
	LPC	Cepstral	LPC	Cepstral	LPC	Cepstral	LPC	Cepstral
Clean speech	88.25%	91.75%	87.75%	92.50%	93%	94.75%	88.75%	94%
SNR=20 dB	28.50%	78.50%	29.25%	80.25%	87.50%	81.75%	71.25%	88.25%
SNR=15 dB	18.25%	55.25%	19%	52.75%	65.50%	57.25%	56.75%	80%
SNR=10 dB	15.25%	31%	11.50%	34.50%	35.75%	32.25%	36%	64.50%
SNR=5 dB	13.75%	21.25%	5.50%	25.25%	22%	11.25%	12%	45.25%
SNR=0 dB	10%	11.5%	5%	21%	5.5%	5%	5.25%	32%

Table 2 illustrates the overall recognition rate when the procedure of section 3 is employed to enhance the noisy features. Again, Mahalonobis distance outperforms the others when the LPC features are used. For cepstral features the proposed SVD-based algorithm gives better recognition rate than the others. Moreover, the modification features algorithm improves the recognition rate when the distance measures are used. But for the SVD-based algorithm, the improvement is marginal which means that the extra modification step, which requires extra computational work, is not needed. Notice

Table 2. Results of the text-independent experiments using Euclidean distance (ED), weighted distance (WD), Mahalonobis distance (MD) and SVD-based algorithm for noisy speech after feature modification

	Recognition rate using ED		Recognition rate using WD		Recognition rate using MD		Recognition rate using SVD	
	LPC	Cepstral	LPC	Cepstral	LPC	Cepstral	LPC	Cepstral
SNR=20 dB	29.75 %	80.25%	30%	82.25%	89.50%	82.50%	72.25%	88.25%
SNR=15 dB	18%	60.25%	19.5%	57.25%	73.75%	60%	57.50%	80%
SNR=10 dB	16%	36.25%	13%	39.50%	50.25%	37.25%	37%	65.50%
SNR=5 dB	13.75 %	26.25%	7.5%	30.75%	33.75%	20.25%	13.75%	49.50%

SNR=0 dB	12%	17.25	5.5%	23.75%	5.5%	5%	5%	35.50%

that, even with this extra step, the results of Table 1 for the proposed algorithm is better than what is obtained by the conventional algorithms with modified features. Figs. 2 and 3 illustrate the correct recognition rate (%) versus the SNR for cepstral and LPC coefficients, respectively.

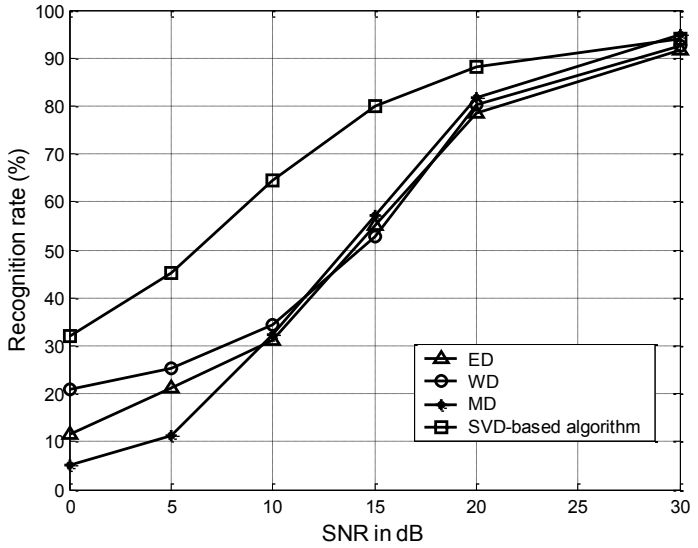

Fig. 2. Recognition rate (%) of ED, WD, MD and SVD-based algorithm using noisy cepstral coefficients.

Figures 4 and 5 are the same as 2 and 3, but with applying the features modification of Section 3. As we notice in Fig. 1, the features modification is optional and as mentioned before, it will not add that much to the recognition rate when the proposed algorithm is used. Comparison between Figs. 2 and 4 confirms the fact that the proposed SVD-based algorithm gives better performance than what is achieved by the conventional distance measure algorithms even with modified features. On the other hand if we compare Figs. 3 and 5, it is clear that the Mahalonobis distance performs better with LPC coefficient. Also, applying the modification algorithm of Section 3 would give substantial improvement in the recognition rate. The question now is, why is the Mahalonobis distance performs better with LPC coefficients? This is a difficult question, but from the observation that we noticed during our study, we may pinpoint to the following reasoning: In Mahalonobis distance, the underlying assumption is that the features of the speakers are Gaussian distributed and the covariance matrix is a sort of weighting, or compensating the variability of features in the training and testing. So, if the distribution of the features in the training and testing is not Gaussian, then we should

not expect a good performance with Mahalonobis. So, it seems to me that the distribution of the LPC fits the Gaussian distribution and even with adding white noise to

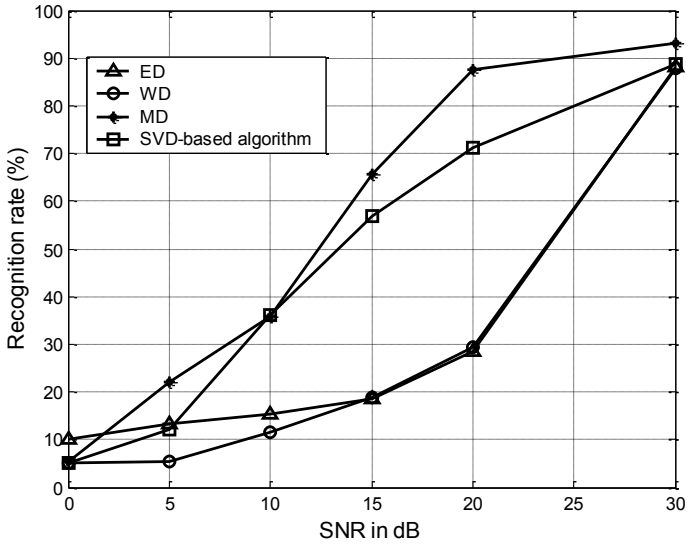

Fig. 3. Recognition rate (%) of ED, WD, MD and SVD-based algorithm using noisy LPC coefficients.

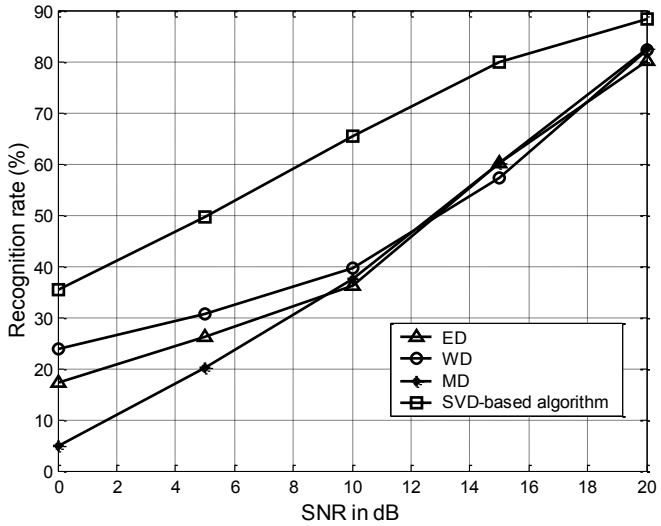

Fig. 4. Recognition rate (%) of ED, WD, MD and SVD-based algorithm using modified cepstral coefficients.

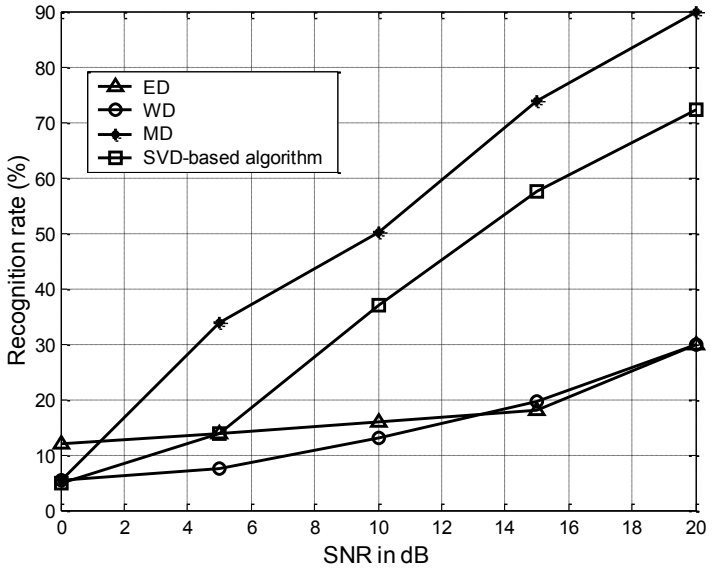

Fig. 5. Recognition rate (%) of ED, WD, MD and SVD-based algorithm using modified LPC coefficients.

the speech signal, the features are still fitting the Gaussian distribution. Future works must be conducted to investigate this phenomenon.

Finally, it is worth mentioning here that in a previous work [19], we obtained better recognition rate when the test utterance is longer than what is considered here (2-4 seconds versus 1-3 seconds here, in this paper). This confirms the assumption that the accuracy of the features and their relevance to the speaker is dependent on the number of the voiced frames obtained from the test utterance.

6. Conclusion

In this paper a new technique for text-independent speaker recognition in noisy environment is presented. This technique is based on finding the ratio of the singular values of a matrix formed from the test feature and the average reference features of every speaker in the constructed data base. The proposed SVD-based algorithm is compared with the conventional distance measure algorithms in case of clean speech and noisy speech of 0 dB to 20 dB SNR. It is found that the proposed algorithm outperforms

the conventional algorithms and it is more robust against noise. Moreover, it is found that the features extracted from the voiced frames give better overall recognition rate than the ones extracted from the whole frames. This means that the voiced frames carry more precise speaker information. In Section 3, we attempted to enhance the noisy features by series expansion in order to obtain a better recognition rate. Again, the comparison with the other algorithms is conducted to show the significance of the proposed algorithm.

References

[1] Deller, J. R., Proakis, J. G. and Hansen, J. H. *Discrete Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.

[2] Campbell, J. P. "Speaker Recognition: A Tutorial". *Proc. IEEE,* 85 (1997), 1437-1462.

[3] Furui, Sadaoki. "Recent Advances in Speaker Recognition". *Pattern Recognition Letters,* 18, No. 9 (1997), 859-872.

[4] Atal, B. S. "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", *J. Acoust. Soc. Amer*., 55, No. 6 (1974), 1304-1312.

[5] Tohkura, Yoh'ichi. "A Weighted Cepstral Distance Measure for Speech Recognition", *Proc. Int. Conf. Acoust. Speech and Signal Process. (ICASSP-86),* 761-764.

[6] Ong, S., Sridharan, S., Yang, C.-H. and Moody, M.P. "Comparison of Four Distance Measures for Long Time Text-independent Speaker Identification". *Proc. ISSPA,* (Aug. 1996), 369-372.

[7] Griffin, C., Matsui, T. and Furui, S. "Distance Measures for Text-independent Speaker Recognition Based on MAR Model". *Proc. Int. Conf. Acoust. Speech and Signal Process.(ICASSP-94),* I309-312.

[8] Goplan, K. and Mahil, S.S. "Speaker Identification and Verification via Singular Value Decomposition of Speech Parameters". *Proc. 33rd Midwest Symposium on Circuits and Systems,* Calgary, Alberta, Canada, (Aug. 1990), 725-728.

[9] Caini, C., Salmi, P. and Corali, A.V. "CD-HMM Algorithm Performance for Speaker Identification on an Italian Database". *Proc. IEEE Int. Conf. Inform. Comm., and Signal Process. (ICICS'97),*2 (September 1997), 1003-1006.

[10] Gales, M.J.F. "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition". *Computer Speech and Language,* 12 (1998), 75-98.

[11] Matsui, T., Kanno, T. and Furui, S. "Speaker Recognition Using HMM Composition in Noisy Environments". *Computer Speech and Language,* 10 (1996), 107-116.

[12] Castellano, P. and Sridharan, S. "A Two Stage Fuzzy Decision Classifier for Speaker Identification". *Speech Communication,* 18, No. 2 (1996), 139-149.

[13] Misra, H., Ikbal, S. and Yegnanarayana, B. "Speaker-specific for Text-independent Speaker Recognition". *Speech Communication,* 39 (2003), 301-310.

[14] Song, F.K., Rosenberg, A.E., Rabiner, L.R. and Juang, B. H. "A Vector Quantization Approach to Speaker Recognition". *Proc. Int. Conf. Acoust. Speech and Signal Process. (ICASSP-90),* 281-284.

[15] Reynolds, D.A. and Rose, R.C. "Robust Text Independent Speaker Identification Using Gaussian Mixture Models". *IEEE Trans. Speech Audio Process.,* 3, No. 1 (1995), 72-83.

[16] Gong, Y. "Speech Recognition in Noisy Environment: A Survey", *Speech Communication*, 16, No. 3 (1995), 261-291.

[17] Cowling, M. and Sitte, R. "Comparison of Techniques for Environmental Sound Recognition". *Patter Recognition Letters*, 24 (2003), 2895-2907.

[18] Dampar, R.I. and Higgins, J.E. "Improving Speaker Identification in Noise by Subband Processing and Decision Fusion". *Pattern Recognition Letters*, 24 (2003), 2167-2173.

[19] Aldhaheri, R. W. and Al-Saadi, F. E. "Text -independent Speaker Identification in Noisy Environment Using Singular Value Decomposition". *Proc. 4th Int. Conf. on Inform. Comm, and Signal Processing (ICICS-PCM 2003),* 3 (December 2003), 1624-1628.

[20] Stewart, G. W. *Matrix Algorithms, Volume 1: Basic Decompositions,* SIAM, (1998).

التعّرف آلياً على متحدث في بيئة ضوضائية باستخدام تجزيء القيمة المفردة كقياس للتطابق حينما تكون العبارة المنطوقة قصيرة

د. رباح واصل الظاهري* و م. فؤاد عيد الصاعدي**

**جامعة الملك عبد العزيز – قسم الهندسة الكهربائية وهندسة الحاسبات*

ص.ب ٨٠٢٠٤ – جدة ٢١٥٨٩

***كلية جدة للالكترونيات والاتصالات، قسم الاتصالات*

ص.ب ١٦٩٤٧ – جدة ٢١٤٧٤

(قدّم للنشر في ٢٠٠٣/٠٩/٢٢م؛ وقبل للنشر في ٢٠٠٤/٠٢/١١م)

ملخص البحث. قدمت في هذه الورقة طريقة جديدة للتعرف على متحدث في بيئة ضوضائية دون النظر إلى العبارة المنطوقة.

تبدأ هذه الطريقة بتحليل العبارة المنطوقة وإيجاد سمات تعتمد على معاملات التنبؤ الخطية لكل متحدث مسجل في قاعدة البيانات. هذه السمات تخزن على أنها دلائل على المتحدثين وعند عملية اختبار أي متحدث والتعرف عليه تحلل أيضا العبارة التي ينطقها ويستخرج منها السمات الخاصة بالمتحدث وتقارن هذه السمات بكل السمات الموجودة في قاعدة البيانات.

المقارنة تتم بإيجاد نسبة القيمة المفردة الكبرى على القيمة المفردة الصغرى والمتحدث الذي يعطي أكبر نسبة يعتبر هو المتحدث المجهول.

طبقت هذه الطريقة على عشرين متحدثٍ (١٦ رجلا و ٤ نساء). يقوم كلاً منهم بنطق عشرين عبارة اختبار وقد أثبتت هذه الدراسة على أن الطريقة المقترحة أفضل من الطرق الأخرى التي أجريت عليها الدراسة تحت كل الظروف المختلفة من الضوضاء حيث حصلنا على نسبة تعرف تصل

إلى ٩٤% في حالة عدم وجود ضوضاء، وحينما تكون الضوضاء عالية جداً (نسبة الإشارة إلى نسبة الضوضاء = ٠ dB)، فإن النسبة تصل إلى ٣٢%.