SPEAKER VERIFICATION USING COMPOSITE REFERENCE

M. Vidalon*
Dept. of Electrical Engineering
University of Windsor
Windsor, Ontario N9B 3P4
Canada

M. Shridhar
Dept. of Electrical Engineering
University of Windsor
Windsor, Ontario N9B 3P4
Canada

M. Cañas
Instituto Venezolano de Investigaciones Cientificas
Caracas, Venezuela

Abstract—This paper examines the effectiveness of parametric representation of speech derived from the linear prediction model using the new concept of composite reference for speaker verification. In the proposed verification system the combined information of the identity of several speakers is given. That is, the existence of an overall reference contour is assumed which contains the information corresponding to the identities of several speakers in a given population. The verification task consists of authenticating an individual's claimed identity by an LPC analysis of his speech. A time warping procedure based strictly on correlation also has been incorporated. Automatic verification based on the predictor and parcor coefficients and using composite reference yielded a verification accuracy of about 96 percent.

I. INTRODUCTION

Most of the acoustic parameters of speech have two components: 1) information (or message) pertaining to the utterance, and 2) information about speaker-characteristics. Thus, it is possible that a given utterance may be dominated by the message and the speaker-defining information may be totally buried. Hence any distance measurement based on conventional methods may yield erroneous verification. The concept of composite reference (hereafter referred to as C.R.) is proposed to permit isolating the speaker characteristics from the message characteristics. This way the reliability of verification could be improved.

The method of linear prediction has proved quite popular and successful for use in speaker recognition [1]-[4]. In this work, an eight and six order linear prediction models were studied. For speech wave sampled at 8 kHz, eight predictor coefficients together with additional parameters such as autocorrelation and parcor coefficients were computed. Also, six predictor coefficients and related parameters were obtained from speech wave digitized at 4 kHz [5]. In the following we describe the basic procedure adopted for implementing a speaker verification system based on linear prediction analysis of speech using C.R. and provide a discussion of the results obtained.

II. FEATURE EXTRACTION

The predicted value of the nth speech sample in the linear prediction model is given by

$$\hat{S}_n = \sum_{k=1}^{p} a_k S_{n-k} \qquad (1)$$

where the a_k's are the predictor coefficients and S_n is the nth speech sample. For a particular frame of speech, the predictor coefficients are determined by minimizing the mean-squared prediction error between S_n and $\hat{S}_n$ [6]. The parcor coefficients k_i's are computed from the predictor coefficients by a backward recursion [7]. The parameter set chosen for this study included a_4 of the eight-order model and k_3 of the six-order model.

Before any of the selected parameter contours can be used for verification purposes, they must be brought into aligment with the reference contour with which is to be compared.

III. TIME REGISTRATION

In this operation, the speech events of the sample contour are brought into the best possible alignment with the corresponding speech events of the reference contour. At the beginning of the warping procedure the parameter contours are divided into an equal number of segments. Corresponding segments are then aligned in time and intermediate points are linearly interpolated. The description of the registration procedure follows.

Let S(t) be a function of time which is some parametric representation of the original speech wave form. If $S(t_1)$ represents S(t) during replication one and similarly R(t) represents a simple average of N replications of S(t), then it is desired to maximize, segmentwise, the similarity between $S(t_1)$ and R(t). The maximization of the similarity, say in segment i, is accomplished by:

1) a proper selection of the function $\lambda(t)$. λ has been chosen to be a shifting function; i.e.,

$$\lambda(t) = t+\Delta t \ , \quad t_i \leq t \leq t_{i+1} \qquad (2)$$

where Δt is an integer number. If $\Delta t = 0$, then λ represents the identity function,

---

* Presently, the author is with the Instituto Venezolano de Investigaciones Cientificas Apartado 1827, Caracas, Venezuela.

$\lambda = t$. If $\Delta t > 0$, then λ represents Δt units of shift to the right. If $\Delta t < 0$, then λ represents Δt units of shift to the left.

2) assigning values to Δt. In this work, $\Delta t = \{0, \pm 1, \pm 2, \pm 3\}$, and the coefficient of correlation, $\rho_{sr,i}$, between the shifted function $S(\lambda)$ and the function with which is to be registered, $R(t)$, is computed. The coefficient of correlation in segment i is defined as

$$\rho_{sr,i} = \frac{\overline{S(\lambda)R(t)}}{\overline{S^2(\lambda)R^2(t)}^{\frac{1}{2}}}, \quad t_i \leq t \leq t_{i+1} \qquad (3)$$

The overbar means time average computed over the range, $[t_i, t_{i+1}]$, where t and λ are defined.

3) determining the relative value of Δt for which the coefficient of correlation is the largest.

The shifted $S(\lambda)$ corresponding to the Δt obtained from step 3) becomes the final registered sample contour.

IV. CONSTRUCTION OF C.R. CONTOUR

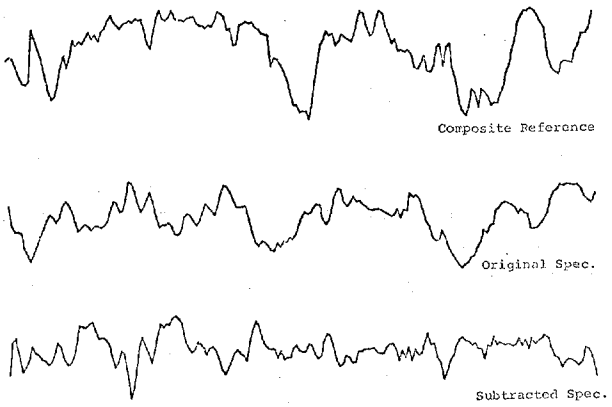
The construction of the composite reference file is done off-line, and it is updated from time to time using new set of analyzed data. The composite reference file is constructed using parameter contours of the design set of each speaker. Five specimens, for each speaker, were used in order to construct the speaker's initial reference file. The algorithm is similar to the one described by Lummis [8]. It is based on averaging all the sample contours, at each sample point, producing intermediate composite trial references. The final composite reference file is obtained at the end of the algorithm and kept on disk for subsequent use.

V. ALGORITHM FOR SPEAKER VERIFICATION USING C.R.

Once the composite reference contour is constructed, a subtracted reference contour for each of the speakers whose specimens entered in the construction of the C.R., is obtained. Suppose speaker Y is in C.R., then the subtracted reference contour of Y, say DYR, is formed by subtracting each of the parameter contours of the design set of speaker Y from C.R., and applying the technique described in section IV to these subtracted parameters. In a practical implementation of a verification system, suppose speaker X enters his identity Y and speaks his verification utterance. After a linear prediction analysis is performed on the offered voice sample, a parameter contour is constructed for X. Then, the task of verification is carried out by the following algorithm:

1) Subtract the parameter contour of speaker X, sample by sample, from C.R. The result is a new subtracted parameter contour, call it DX.
2) Register DX against the subtracted reference contour of speaker Y, namely, DYR.
3) Compute distances between DX and DYR.
4) Compute threshold and decide whether claim is valid or not.

The following figure illustrates the composite reference contour for a population of four speakers. Also a test parameter contour and its subtracted version are shown. The parameter is the parcor coefficient k_3.



Composite Reference

Speakers: JN, PE, MS, WCK

Test Contour: PE

Utterance: "My name is Miller, cash this bond, please"

VI. EXPERIMENTAL RESULTS

A. Data Collection

The utterances chosen for the experiment were:

1) Cash this bond, please
2) We were away a year ago
3) My name is Miller, cash this bond, please

The speech data consisted of 63 utterances, consisting of nine repetitions of the same utterance spoken by seven speakers. All the speakers were native male canadians with no noticeable speech defects. All the recordings were made over a period of one month in a studio environment. The stored samples were analyzed utterance by utterance. Each utterance was divided into equal number of segments. The duration of each segment was made proportional to the duration of the utterance. Thus, the resulting parameter contours were equal in length. The silent portions and pauses were removed automatically from the analysis interval.

B. Verification Procedure

Following the automatic time registration of the test sample contour to the reference contour, a set of measurements is applied to the test sample contour and compared with the same measurements applied to the reference contour. Two kinds of distances are calculated, those based on short segmental correlation and those computed over the whole contour. A detailed description of the distance computation carried out is given by Lummis [8]. In addition, a new set of distances, called exponential distance, was investigated. This measure was evaluated segment by segment, and also over the whole contour.
The segmental exponential distance is defined by

$$Ed_1 = \frac{1}{NSEG} \sum_{i=1}^{NSEG} \exp(1/\rho_{xr,i}^2) \qquad (4)$$

where $\rho_{xr,i}$ is the coefficient of correlation between the unknown sample contour x and the reference contour r in segment i. NSEG represents the number of segments.
The second exponential distance is defined over the whole contour and is given by

$$Ed_2 = \exp(1/\rho_{xr}^2) \qquad (5)$$

where ρ_{xr} is the correlation coefficient between the unknown sample and reference contours computed over the whole length of the parameter contour.

C. Results

Table I summarizes the results of the evaluation.

TABLE I

		Error Rates (%)	
Sent.	Para.	Spks. in C.R.	Spks. out of C.R.
1	a_4	4.5	None
2	k_3	4	0
3	k_3	0	4

The error rate obtained with the linear predictor coefficient a_4 for sentence 1 is 4.5 % for all the speakers whose specimens entered in the construction of the C.R. file. Speakers outside of C.R. were not used. The error rate obtained with the parcor coefficient k_3 for sentence 2 is 4 percent for speakers whose specimens entered in the construction of C.R. For the speakers that did not enter in the construction of C.R., all their sample contours were verified correctly. For sentence 3 and k_3, the error rate is in the same order of magnitude as that of sentence 2.
The number of speakers used to construct C.R. varied between 4 and 5. The error rate obtained here represent an improvement over the conventional approach[5], and further the error rates can be reduced by including more parameters in the overall computation.

VII. CONCLUSIONS

The results of this preliminary work show that verification with the use of composite reference is viable. Further investigation involving large population of speakers and new features for speaker verification are needed.

REFERENCES

[1] B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Amer., vol. 55, pp. 1304-1312, 1974.

[2] A.E. Rosenberg and M.R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoustic, Speech, and Signal Processing vol. ASSP-23, pp. 169-176, 1975.

[3] M.R. Sambur, "Speaker Recognition Using Orthogonal Linear Prediction," IEEE Trans. Acoustic Speech, and Signal Processing, vol. ASSP-24, pp. 283-289, 1976.

[4] A.E. Rosenberg, "Evaluation of an Automatic Speaker-Verification System Over Telephone Lines," The Bell System Technical Journal, vol. 55, pp. 723-744, 1976

[5] M. Shridhar and M. Vidalon, "An Algorithm for Speaker Verification," J. Acoust. Soc. Amer., vol. 60, S13(A), 1976.

[6] B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction," J. Acoust. Soc. Amer., vol. 50, pp. 637-655, 1971.

[7] J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, vol. 63, pp. 561-580, 1975.

[8] R.C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 80-89, 1973.