Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech

S. Furui*
A. E. Rosenberg

Bell Laboratories
Murray Hill, New Jersey 07074

ABSTRACT

New techniques for automatic speaker verification are described which are based on cepstral analysis of fixed sentence-long utterances. Utterances are represented by time functions of cepstral coefficients expanded by an orthogonal polynomial. Sample utterance representations are compared with similarly constituted reference patterns to decide whether to accept or reject an identity claim. A dynamic programming technique is used to bring sample utterance representations into time registration with reference patterns. A local distance measurement is summed along the optimal path associated with the time registration to provide an overall distance which is compared with a threshold to effect the decision.. Results of an evaluation using prerecorded utterance sets including telephone speech are presented. Verification error rates less 1% are obtained.

Speaker verification is a process in which the identity claim of an individual is accepted or rejected by comparing a set of measurements of the individual's spoken utterances with a reference set of measurements for the utterances of the person whose identity is claimed. Research on an automatic system for speaker verification at Bell Laboratories has been reported in previous papers [1-4]. The system is based on an acoustic analysis of a fixed, sentence-long utterance resulting in a function of time or contour for each feature analyzed. Features selected in previous implementations have included pitch, intensity, the first three formants and selected linear prediction coefficients. This paper describes new techniques for this system based on the selection of cepstral coefficients, expanded by an orthogonal polynomial representation, as feature contours. In addition a new method of overall distance computation is employed.

SYSTEM DESCRIPTION

A block diagram indicating the principal operation of the system is shown in Figure 1. There are two inputs to the system, an identity claim and a sample utterance. The identity claim, which may be provided by a keyed-in identification number, causes reference data corresponding to the claim to be retrieved. The second input is activated by a request to speak the sample utterance. The speech wave, bandlimited typically from 100 Hz to 3.2 kHz and digitized at a 6.67 kHz rate, is scanned forward from the beginning of the recording interval and backward from the end to determine the actual endpoints of the sample utterance. This is accomplished by means of an energy calculation. The delimited speech is pre-emphasized and analyzed every 10 msecs. over a 30 msec. Hamming window to extract the first 10 linear predictor coefficients by the autocorrelation method. These are then transformed to cepstrum coefficients using a recursive relationship [5]. The cepstrum coefficients are averaged over the duration of the entire utterance and the average values are subtracted from the coefficients of each frame to compensate for fixed spectral distortions introduced by the transmission system. The time function of each normalized cepstrum coefficient is then expanded by a second order orthogonal polynomial representation over a 90 msec interval every 10 msecs. (In actuality, the zero-th order polynomial coefficients are replaced by the original normalized cepstrum coefficients.) Since a second order polynomial is applied to each of 10 cepstrum coefficients, the result is a representation of the utterance by a 30-dimensional vector every 10 msecs through the utterance. From these 30 elements, a fixed reduced set (typically 18), is selected which is most effective in separating the populations of customer and impostor sample utterances. The criterion for selection is based on an inter-to-intra-speaker variability ratio for each element calculated over a population of training utterances.

A crucial property of the system is automatic time registration of the feature contours extracted from the sample utterance to the feature contours retrieved as the reference template corresponding to the claimed identity. This is accomplished by means of a dynamic programming technique which makes use of a local measure of similarity or distance between the n-th frame of the reference contour $R(n), l \leqslant n \leqslant N$, and the m-th frame of the test contour, $T(m)$, $l \leqslant m \leqslant M$, to establish an optimum mapping, $m = w(n)$, between the two contours together with an overall distance, D_T, associated with this mapping. Representing the local distance as $d(R[n], T[m])$, the optimum path $w(n)$ is that which minimizes the accumulated local distance, yielding

$$D_T = \min_{w[n]} \sum_{n=1}^{N} d(R[n], T(w[n])) \qquad (1)$$

Dynamic programming is an efficient technique for obtaining the optimum mapping and associated overall distance making use of a recursion formula for partial accumulated distances together with boundary and local continuity conditions. Since there is often some uncertainty in the location of both the initial and final frames due to breath noises, clicks, etc., the boundary conditions are relaxed by using an unconstrained endpoint technique [6].

Denoting the feature vector of the n-th frame of the reference contour as $R[n] = (r_1[n], r_2[n], \cdots r_k[n])$ and the m-th frame of the test contour as $T[m] = (t_1[m], t_2[m] \cdots t_k[m])$, when k is the (reduced) number of elements in the feature set, the local distance is given by

$$d(R[n], T([m]) = \left[\sum_{i=1}^{k} g_i | r_i[n] - t_i[m] | \right]^2 \qquad (2)$$

where the weighting function g_i is the reciprocal of the mean value of intra-speaker variability for each element i. The intra-speaker

* Permanent Address: Electrical Communication Laboratories, Nippon Telegraph & Telephone Public Corp., Musashino, Tokyo, Japan.

CH1559-4/80/0000-1060$00.75 © 1980 IEEE

variability measurement is calculated over a training set of utterances for all the speakers.

The overall distance accumulated over the optimum warping function is compared with a threshold distance to determine whether to accept or reject the identity claim. There are two possible types of error, false rejection, rejection of the speaker designated as customer, and false acceptance, acceptance of an impostor who is any speaker from the speaker set other than the designated customer. Generally, thresholds are set to minimize the probability of some combination of these errors. If the threshold is set tightly the probability of false rejection will be high but will be offset by a low probability of false acceptance. Conversely, the opposite condition holds for a relaxed threshold. In this experiment we attempt to set the threshold at a value which equates estimates of the two kinds of error, the so-called equal error threshold. Two methods are used. In the first method, the threshold is set to an experimentally decided fixed value which is observed for all customers. In the second method, an optimum threshold is estimated based on the distribution of overall distances between each customer's reference template and a set of utterances of other speakers. This threshold, based on the distribution of inter-speaker distances is updated at the same time as reference template updating. The following equation, based on empirical results, is used to set this threshold for each customer:

$$\Theta(k) = a(\hat{\mu}_{DB}[k] - \hat{\sigma}_{DB}[k]) + b \qquad (3)$$

where $\Theta[k]$ is the threshold for customer k, $\hat{\mu}_{DB}[k]$ and $\hat{\sigma}_{DB}[k]$ are the mean value and standard deviation for the inter-speaker distribution, respectively. a and b are constants fixed for all customers which are set to maximize the correlation between $\Theta(k)$ and experimentally determined equal-error thresholds over a set of training data.

The establishment and updating of reference information is another important element of the system. The initial reference template is constructed from five training utterances as follows. The first training utterance is taken to be the first trial reference to which the second training utterance is brought into time registration. These are then averaged together to produce the second trial reference. The third training utterance is then brought into time registration with this reference and averaged together to produce the third trial reference. This procedure continues through the fifth trial reference which is taken as the initial reference template. The training utterances are also used for the calculation of the weighting function used in the distance computation of equation 2, the inter-speaker to intra-speaker variability ratio which is used in the feature selection and the inter-speaker distance distribution which is used to set the decision threshold specified in equation 3.

The initial reference template and information are updated periodically as the system is accessed in the test mode. This procedure is important to deal with variability and trends over time that can occur in customer utterances. In the evaluation described in this paper, reference templates were updated every seventh access of the system by each customer using the latest five utterances. Variability is often especially high in the test utterances immediately following the establishment of the initial reference template. For these first few accesses it is desirable to update the reference template following each access.

EVALUATION

Experimental results obtained from three utterance sets are reported in this paper. The first and second sets each compromises 50 utterances by each of 10 speakers designated customers and a single utterance by each of 40 speakers designated impostors. The recordings were made over conventional telephone lines, the first by male speakers, the second by female speakers. The third set consists of 26 utterances by 21 male speakers designated customers and a single utterance from each of 55 male speakers designated impostors all recorded over a high quality microphone. The male speakers recorded the sentence "We were away a year ago," while the female speakers recorded "I know when my lawyer is due." All recordings were bandlimited from 300 to 3200 Hz. All customer recordings were made over periods of several weeks, each in separate sessions with no more than two sessions per day.

Preliminary experiments were carried out to obtain selections of features effective in separating customer utterances from impostor utterances for use in the speaker verification test phase of the experiment. Feature selection results for the first 10 utterances by 5 speakers from the male speaker telephone transmission utterance set are shown in Fig. 2. Shown plotted are the ratios of the average values of inter-speaker distances to the average values of the intra-speaker distances for each of the 30 original parameters. (Note once again that the original cepstrum coefficients are used in place of the zeroth order polynomial coefficients.) With almost total consistency, the higher is the polynomial coefficient order, the smaller are the distance ratios and, hence, the less effective are the parameters. Based on these results, 18 parameters were selected as features for the speaker verification tests. These included all the original cepstrum coefficients, all but two of the first order coefficients and none of the second order coefficients. Similar results were obtained for the other two utterance sets.

The results of speaker verification experiments for the three utterance sets are summarized in Table 1. The a priori error rates shown are the results of averaging the false acceptances and false rejection rates over all the customers in the utterance set. Results are shown both for the fixed threshold estimate applied uniformly to all customers and for the customer specific threshold estimate specified in equation 3. For comparison, a posteriori equal error rates are also shown. These are obtained by averaging the overlap areas between customer and impostor distance distributions for each customer in the utterance set. It is seen that average error rates using the a priori threshold estimator are all 1% or less and that the actual overlap between customer and impostor utterances as provided by the a posteriori error rates are even smaller. Because the error rates are so small it is not possible to reliably distinguish performances among these three utterance sets.

CONCLUSION

In a previous experiment, using a similar speaker verification system implementation, an average error rate of 1.5% was obtained when the features were selected from sets of pitch, intensity, and LPC contours over sentence-long utterances [3]. With pitch and intensity features alone the average error rate obtained was 3%. We conclude, therefore, that the use as features of cepstrum coefficients, expanded as polynomials over segments of sentence-long utterances, provides very high performances with average error rates less than 1%.

References

[1] G. R. Doddington, "A Computer Method of Speaker Verification," Ph.D. Dissertation, Department of Electrical Engineering, University of Wisconsin, 1970.

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

[3] A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-23, (April 1975), pp. 169-176.

[4] A. E. Rosenberg, "Evaluation of an Automatic Speaker-Verification System Over Telephone Lines," Bell Syst. Tech. J., 55 (July-August 1976), pp. 723-744.

[5] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Am., 55 (June 1974), pp. 1304-1312.

[6] L. R. Rabiner, A. E. Rosenberg and S. E. Levinson, "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition," IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-26, (December 1978), pp. 575-582.

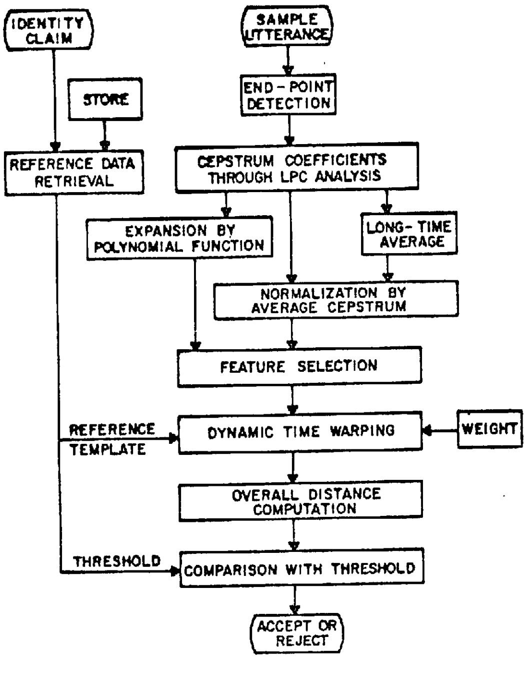

Fig. 1. Block diagram indicating the principal operation of the system.

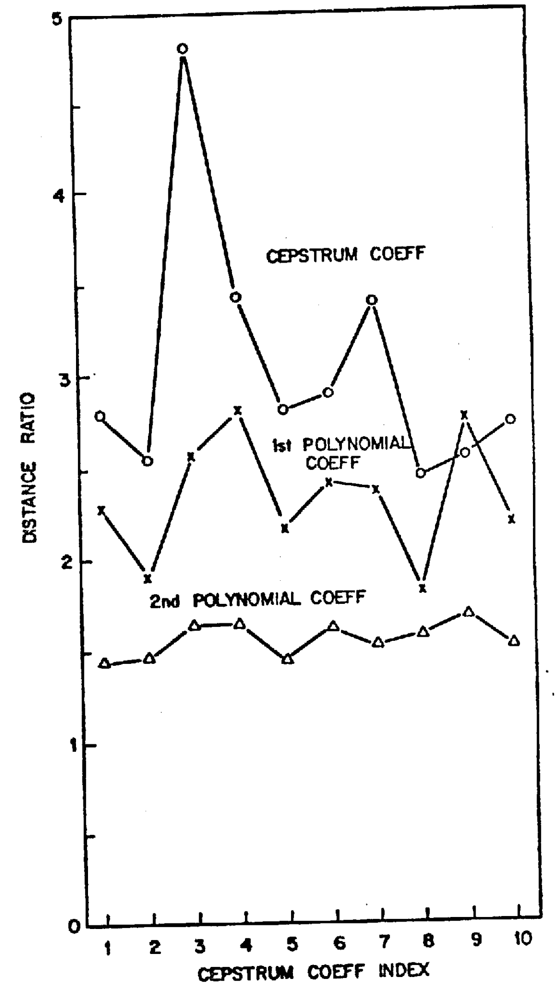

Fig. 2. Inter-speaker to intra-speaker distances for the first 10 utterances by 5 speakers from utterance set 1.

AVERAGE ERROR RATES

UTTERANCE SET		1	2	3
CUSTOMERS & IMPOSTORS		10 MALE & 40 MALE	10 FEMALE & 40 FEMALE	21 MALE & 55 MALE
RECORDING CONDITION		TELEPHONE	TELEPHONE	HIGH QUALITY MICROPHONE
AVERAGE ERROR RATES	A PRIORI FIXED	0.30 %	0.42 %	1.03 %
	A PRIORI ESTIMATED	0.19 %	0.36 %	0.77 %
	A POSTERIORI	0 %	0.06 %	0.64 %

Table I. Overall speaker verification results for 3 utterance sets.