synthesis of recursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 61–65, Feb. 1973.
[5] K. S. Chao and K. S. Lu, "On sequential refinement schemes for recursive digital filter design," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 396–401, July 1973.
[6] F. Brophy and A. C. Salazar, "Recursive digital filter synthesis in the time domain," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 45–55, Feb. 1974.
[7] M. S. Bertrán, "Approximation of digital filters in one and two dimensions," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 438–443, Oct. 1975.
[8] J. A. Cadzow, "Recursive digital filter synthesis via gradient based algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 349–355, Oct. 1976.
[9] R. Hastings-James and S. K. Mehra, "Extensions of the Padé-approximant technique for the design of recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 501–509, Dec. 1977.
[10] R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," *Comput. J.*, vol. 6, pp. 163–168, July 1963.

Takashi Yahagi (M'78) was born in Tokyo, Japan, on February 1, 1943. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1966, 1968, and 1971, respectively.

In 1971, he joined Chiba University, Chiba, Japan, as a Lecturer in the Department of Electronic Engineering. Since 1974, he has been an Associate Professor in the Department of Electronic Engineering. In 1978, he was awarded a Postdoctorate Fellowship from the Royal Norwegian Council for Scientific and Industrial Research, Oslo, Norway. From 1978 to 1979, he was with the Division of Engineering Cybernetics, the Norwegian Institute of Technology, the University of Trondheim, Trondheim, Norway. His current interests are in the areas of digital signal processing, control theory, and estimation theory.

Dr. Yahagi is a member of the Institute of Electronics and Communication Engineers of Japan, the Society of Instrument and Control Engineers, Japan, and the Japan Association of Automatic Control Engineers.

Cepstral Analysis Technique for Automatic Speaker Verification

SADAOKI FURUI, MEMBER, IEEE

Abstract—This paper describes new techniques for automatic speaker verification using telephone speech. The operation of the system is based on a set of functions of time obtained from acoustic analysis of a fixed, sentence-long utterance. Cepstrum coefficients are extracted by means of LPC analysis successively throughout an utterance to form time functions, and frequency response distortions introduced by transmission systems are removed. The time functions are expanded by orthogonal polynomial representations and, after a feature selection procedure, brought into time registration with stored reference functions to calculate the overall distance. This is accomplished by a new time warping method using a dynamic programming technique. A decision is made to accept or reject an identity claim, based on the overall distance. Reference functions and decision thresholds are updated for each customer.

Several sets of experimental utterances were used for the evaluation of the system, which include male and female utterances recorded over a conventional telephone connection. Male utterances processed by ADPCM and LPC coding systems were used together with unprocessed utterances. Results of the experiment indicate that verification error rate of one percent or less can be obtained even if the reference and test utterances are subjected to different transmission conditions.

Manuscript received May 5, 1980; revised September 25, 1980.

The author was with Bell Laboratories, Murray Hill, NJ 07974. He is now with the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation, Tokyo 180, Japan.

I. Introduction

SPEAKER verification is a process to accept or reject the identity claim of a speaker by comparing a set of measurements of the speaker's utterances with a reference set of measurements of the utterance 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 for analysis in previous evaluations have included pitch, intensity, the first three formants, and selected prediction coefficients. The system which uses pitch and intensity contours has been evaluated using telephone speech over a period of five months with a test population of over 100 male and female speakers. The evaluation indicated an error rate of approximately ten percent for new customers and approximately five percent for adapted customers [4]. It has also been shown that the performance of this system is relatively insensitive to transmission systems in which the speech is encoded using adaptive differential pulse code modulation (ADPCM) coding or linear predictive coding (LPC) vocoding [5].

0096-3518/81/0400-0254$00.75

This paper describes new techniques for an automatic speaker verification system for telephone-quality speech. The differences between the present implementation and previous implementations of the system lie in the features selected for analysis and the method of overall distance computation. In addition, new and enlarged samples of speech, including several kinds of transmission systems have been used for evaluation.

II. System Operation

A block diagram indicating the principal operations of the system is shown in Fig. 1. There are two inputs to the system, the identity claim and the 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 recording interval is scanned to find the endpoints of the utterance. The utterance is then analyzed. Linear predictor coefficients are extracted successively and these coefficients are transformed into cepstrum coefficients. The cepstrum coefficients are averaged over the duration of the entire utterance and the average values are subtracted from the cepstrum coefficients of every frame to compensate for frequency-response distortions introduced by the transmission system.

The time functions of the cepstrum coefficients are expanded by an orthogonal polynomial representation over short time segments. Then the utterance is represented by the time functions of coefficients of the orthogonal polynomial representation. A part of the set of these coefficients is selected for speaker verification, based on the statistical analysis of the effectiveness of each coefficient.

A crucial property of the system is automatic time registration of the time functions of the sample utterance to the time functions retrieved as the reference template of the claimed identity. An overall distance between the sample utterance and the reference template is obtained as the result of time registration using a dynamic programming technique. The distance of each element is weighted by intraspeaker variability and summed to produce the overall distance. Finally, the overall distance is compared with a threshold distance value to determine whether the identity claim should be accepted or rejected.

Details concerning the analysis procedures, reference construction, time registration, and distance calculation will be presented in the following sections.

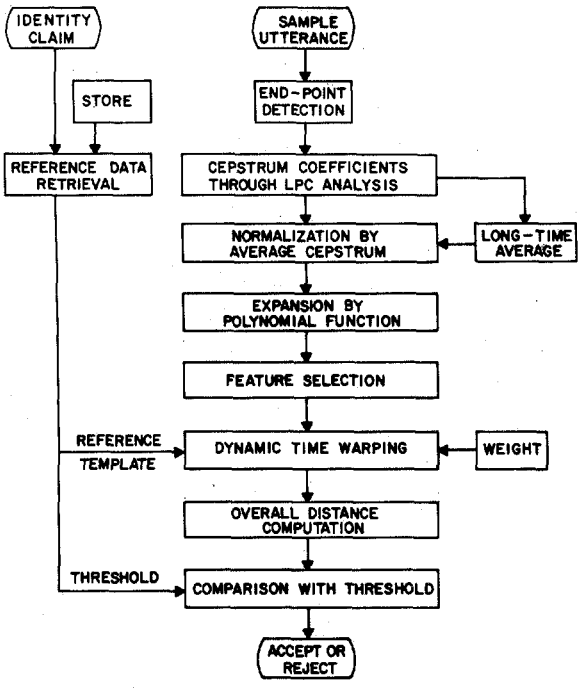


Fig. 1. Block diagram indicating the principal operations of the system.

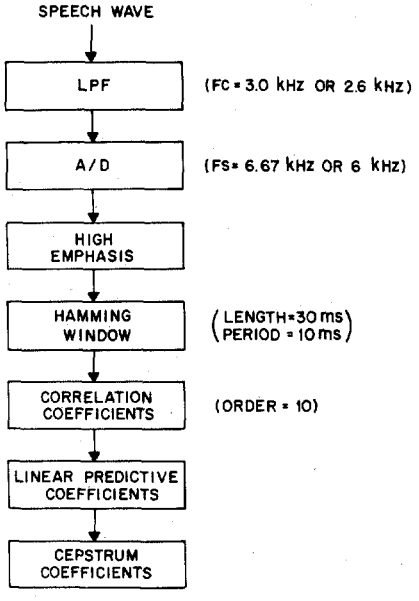


Fig. 2. Block diagram for cepstrum extraction.

A. Normalized Cepstrum Extraction

The speech wave is bandlimited from 100 Hz to 3.0 kHz and sampled at a 6.67 kHz rate, or bandlimited from 100 Hz to 2.6 kHz and sampled at 6 kHz. The digitized speech is then scanned forward from the beginning of the recording interval and backward from the end to determine the beginning and end of the actual sample utterance. The endpoint detection is accomplished by means of an energy calculation. A high emphasis filter $(1 - 0.95Z^{-1})$ is applied to the delimited speech, and a 30 ms Hamming window is applied to the emphasized speech every 10 ms. First to tenth-order linear predictor coefficients are extracted from each frame by the autocorrelation method. The linear predictor coefficients are transformed into cepstrum coefficients, using the following recursive relationships [9]:

$$c_1 = a_1$$

$$c_n = \sum_{k=1}^{n-1} (1 - k/n)\, a_k c_{n-k} + a_n, \qquad 1 < n \leq p \tag{1}$$

where c_i and a_i are the ith-order cepstrum coefficient and linear predictor coefficient, respectively. Fig. 2 shows the block diagram of these processes.

Atal [9] examined several different parametric representations of speech derived from the linear prediction model for their effectiveness for automatic recognition of speakers. Among all the parameters investigated, the cepstrum was found to be the most effective. It was also pointed out that cepstrum coefficients have the additional advantage that one can derive from them a set of parameters which are invariant to any fixed frequency-response distortion introduced by the recording apparatus or the transmission system. The new parameters are obtained simply by subtracting from the cepstrum coefficients a set of values representing their time averages over the duration of the entire utterance. This process can normalize the gross spectral distribution of the utterance, and it is similar to the inverse filtering process which has been used in a spoken word recognition system at Bell Laboratories [6]. The normalization technique introduced by Atal is used in the speaker verification system studied in this paper.

In previous studies by the author [7], [8] it was shown that this normalization process is also effective in reducing long-term intraspeaker spectral variability for maintaining high speaker verification and identification accuracy over a long period.

B. Polynomial Coefficients

Time functions of the normalized cepstrum coefficients are expanded by an orthogonal polynomial representation over 90 ms intervals every 10 ms. The 90 ms interval length seemed adequate for preserving transitional information between phonemes. The first three orthogonal polynomials are used. They are [10]

$$P_{0j} = 1$$
$$P_{1j} = j - 5$$
$$P_{2j} = j^2 - 10j + \frac{55}{3}. \tag{2}$$

Thus, if the control function samples for an utterance within the segment being measured are $x_j (j = 1, 2, \cdots, 9)$, then the first three coefficients of the orthogonal polynomial representation are

$$a = \left(\sum_{j=1}^{9} x_j\right) \Big/ 9$$
$$b = \left(\sum_{j=1}^{9} x_j P_{1j}\right) \Big/ \sum_{j=1}^{9} P_{1j}^2$$
$$c = \left(\sum_{j=1}^{9} x_j P_{2j}\right) \Big/ \sum_{j=1}^{9} P_{2j}^2. \tag{3}$$

These coefficients represent mean value, slope, and curvature of the time function of each cepstrum coefficient in each segment, respectively.

As the original time functions of cepstrum coefficients are considered to be more efficient than the 0th-order polynomial coefficients for speaker verification, the original time functions of cepstrum coefficients are used to replace the 0th-order polynomial coefficients in this implementation. When the 0th-order polynomial coefficients are not used, the block diagram of feature extraction is modified as shown in Fig. 3, since cepstrum normalization does not affect the first- and second-order polynomial coefficients.

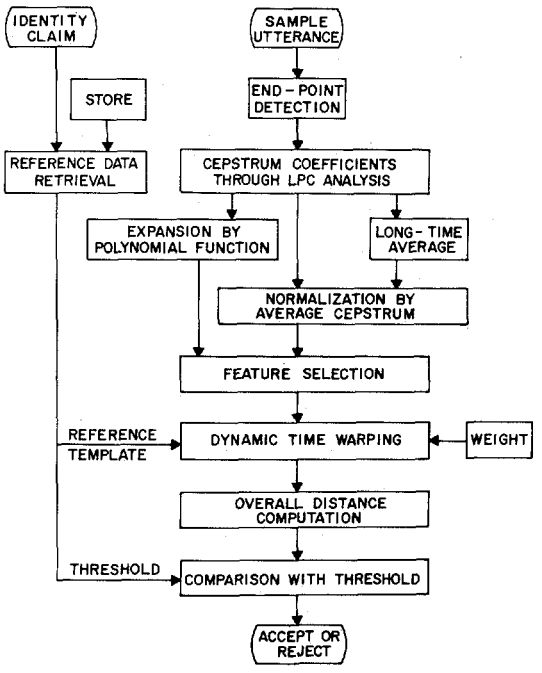


Fig. 3. Block diagram indicating the principal operations of the system (modification of Fig. 1).

Accordingly, the utterance is represented by time functions of the cepstrum coefficients $x_t(i)$, and the first- and second-order polynomial coefficients, $b_t(i)$ and $c_t(i)$, where t is the frame number and i is the index of the cepstrum coefficient $(1 \leqslant i \leqslant p)$. Since p is set to ten in this system, the result is a representation by a time function of a 30-dimensional vector. From these 30 elements, a set of elements, which are most effective in separating the overall distance distribution of customer and impostor sample utterances are selected for speaker verification. The selection is made based on the inter-to-intraspeaker variability ratio for each element:

$$h_i = \bar{\bar{d}}_{Bi} / \bar{\bar{d}}_{Wi}$$
$$\bar{\bar{d}}_{Bi} = \underset{\substack{j,k \\ (j \neq k)}}{E} (\bar{d}_{ijk}), \qquad \bar{\bar{d}}_{Wi} = \underset{j}{E} (\bar{d}_{ijj})$$
$$\bar{d}_{ijk} = \underset{\substack{l,m \\ (l \neq m, \text{ if } j=k)}}{E} d_{ijklm} \tag{4}$$

where $\underset{j}{E}$ means averaging over the index j, and d_{ijklm} is the distance between the time function of ith element derived from lth utterance by speaker j and mth utterance by speaker k after time registration.

C. Time Registration

A sample utterance is brought into time registration with the reference template to calculate the distance between them. This is accomplished by a new time warping method using dynamic programming technique. As there is often some uncertainty in the location of both the initial and final frames

due to breath noise, etc., the unconstrained endpoint technique [11] is applied.

We denote two contours as $R(n)$, $1 \leqslant n \leqslant N$, and $T(m)$, $1 \leqslant m \leqslant M$. We denote $R(n)$ and $T(m)$ as guide contour and slave contour, respectively. The purpose of the time warping algorithm is to provide a mapping between the time indexes n and m such that a time registration between the two utterances is obtained. We denote the mapping w, between n and m as

$$m = w(n). \tag{5}$$

The function w must satisfy a set of boundary conditions at the endpoints of the utterance and some restrictions on the form it assumes. In our case, the following conditions are applied:

$$w(n+1) - w(n) = 0, 1, 2 \quad (w(n) \neq w(n-1)) \tag{6a}$$

$$= 1, 2 \quad (w(n) = w(n-1)) \tag{6b}$$

$$1 \leqslant w(1) \leqslant \delta + 1 \tag{6c}$$

$$M - \delta \leqslant w(N) \leqslant M \tag{6d}$$

$$\max w(n) = M, \quad N - \delta \leqslant n \leqslant N \tag{6e}$$

$$\frac{M}{N} n - m_0 \leqslant w(n) \leqslant \frac{M}{N} n + m_0. \tag{6f}$$

Equations (6a) and (6b) require that $w(n)$ be monotonically increasing, with a maximum slope of two, and a minimum slope of 1/2. The minimum slope constraint is a consequence of the prohibition against two consecutive steps with slope 0. In (6c), (6d), and (6e), δ represents the maximum anticipated range of mismatch (in frames) between boundary points of the two utterances. In our case, a value of δ of 15 (frames) was used, representing a 150 ms region in which the initial and final frames could be mapped.

The warping function can reach the final boundary of the slave contour prior to the last frame, i.e., it is possible that

$$w(n) = M \quad \text{for} \quad n < N \tag{7}$$

in which case it is not physically meaningful to continue the path.

Equation (6f) restricts the warping function within some fixed region along the diagonal line which connects (1, 1) and (N, M) points on the $(n - m)$ plane. In our case m_0 was set to 20 (frames). From these conditions the warping function is constrained to follow a path inside the shaded region of Fig. 4. The vertices of the labeled points A and B are obtained as the intersections of the lines.

$$\left.\begin{aligned} m - M &= \tfrac{1}{2}(n - N + \delta) \\ m - \delta - 1 &= 2(n-1) \end{aligned}\right\} \text{point } A$$

$$\left.\begin{aligned} m &= \tfrac{1}{2} n \\ m - M + \delta &= 2(n - N) \end{aligned}\right\} \text{point } B \tag{8}$$

As can be seen in Fig. 4, the warping function can start from any frame of the slave contour between the first and $\delta + 1$th frame, but it must start from the first frame of the guide contour. If we allow the warping function to also start from any frame of the guide contour, the computation time becomes excessive. If the intraspeaker variation of utterance lengths is not large, we have found that the likelihood is great that the optimum warping function starts from the first frame of the shorter of either the reference template or test utterance. The basis for this result is described in Section VI-F. Based on this assumption we adopted the procedure of using as the guide contour the shorter of either the reference template or test utterance. This means that the longer one is mapped to the slave contour axis which is the ordinate of the warping plane and the shorter one is mapped to the guide contour axis which is the abscissa of the warping plane in Fig. 4.

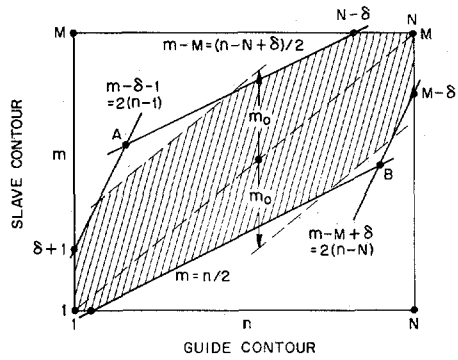


Fig. 4. Allowable region of the dynamic programming warping function path.

A complete specification of the warping function results from a point-by-point measure of similarity between the guide contour $R(n)$ and the slave contour $T(m)$.

D. Distance Measure

A similarity measure or distance function D must be defined for every pair of points (n, m) within the shaded region of Fig. 4. Given the distance function D, the optimum dynamic path w is chosen to minimize the accumulated distance D_T along the path, i.e.,

$$D_T = \min_{\{w(n)\}} \sum_{n=1}^{N} D(R(n), T(w(n))). \tag{9}$$

When the warping function reaches the final boundary of the slave contour prior to the last frame, the accumulated distance D_T is scaled by the factor (N/N_S) where N_S is the frame at which (7) is satisfied, so as to equalize the number of distances which enter into the total distance D_T. The optimum path w can be determined by the method of dynamic programming easily.

Let us denote the feature vector of the nth frame of the guide contour as $R(n) = (r_1(n), r_2(n), \cdots, r_i(n), \cdots, r_K(n))$ and the mth frame of the slave contour as $T(m) = (t_1(m), t_2(m), \cdots, t_i(m), \cdots, t_K(m))$, where K is the number of the elements of the feature vector. In this paper, two kinds of distance measures are used and evaluated.

$$D_1(R(n), T(m)) = \sum_{i=1}^{K} g_i^2 (r_i(n) - t_i(m))^2 \tag{10a}$$

$$D_2(R(n), T(m)) = \left(\sum_{i=1}^{K} g_i \, | r_i(n) - t_i(m) | \right)^2 \qquad (10b)$$

where g_i is the weighting function, which is the reciprocal of the mean value of intraspeaker variability for the ith element, defined as follows:

$$g_i = 1/\bar{\bar{d}}_{Wi}$$

$$\bar{\bar{d}}_{Wi} = \underset{j}{E}(\bar{d}_{ijj}) = \underset{\substack{j \\ k \neq l}}{E} \sum_{n=1}^{N} (t_{ijk}(n) - t_{ijl}(w(n)))^2 \qquad (11)$$

where $t_{ijk}(n)$ is the nth frame of the kth utterance by speaker j.

E. Decision Threshold

The overall distance accumulated over the optimum warping function is compared with a threshold to determine whether to accept or reject an identity claim. In many kinds of speaker verification experiments, the threshold is set *a posteriori* so that the two kinds of error rate (the rate of rejecting utterances which should be accepted and the rate of accepting utterances which should be rejected) are equal. But these experiments are unrealistic, and procedures for setting thresholds in advance in practical situations are not well established.

In this paper, two methods for setting an *a priori* threshold are evaluated. In the first method, the threshold is set to an experimentally decided fixed value, and the same threshold is used for all customers. In the second method, the 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. In the latter case, the threshold is updated at the same time as the reference template updating, based on the distribution of interspeaker distances. The following equation, based on empirical results, is used to set the threshold for each customer:

$$\theta(k) = a(\hat{\mu}_{DB}(k) - \hat{\sigma}_{DB}(k)) + b \qquad (12)$$

where $\theta(k)$ is the threshold for the customer k, $\hat{\mu}_{DB}(k)$ and $\hat{\sigma}_{DB}(k)$ are mean value and standard deviation for the distribution of interspeaker distance, respectively. a and b are constant parameters which are set experimentally, the same values being used for all customers and for all data sets.

Fig. 5 shows an example of typical intraspeaker and interspeaker distance distributions. Equation (12) indicates that as the mean value of the interspeaker distance becomes larger and the standard deviation becomes smaller, the decision threshold becomes larger. The intraspeaker distance distribution is not taken into account in the calculation of the decision threshold. There are two reasons for this. First, the intraspeaker distance distributions are fairly uniform from speaker to speaker. Second, it is difficult to obtain stable estimates of the distribution of intraspeaker distance for small numbers of training utterances, whereas it is easy to estimate interspeaker distance distributions by cross comparison of the training utterances between different customers.

A posteriori equal error decision thresholds were also used to compare the results with those obtained using *a priori* thresholds.

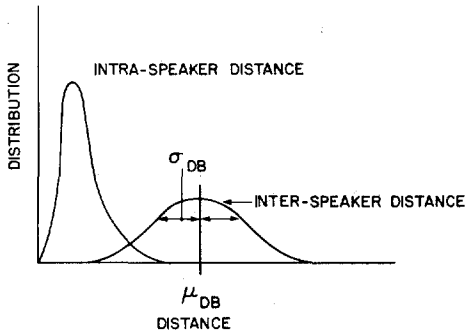


Fig. 5. Example of typical intraspeaker and interspeaker distance distributions.

F. Reference Construction

The establishment and updating of reference information is another important element of the system. For each kind of data set, three or five utterances were used to construct a reference template for each customer. Two methods of reference updating were observed. In the first method, the reference template was updated every seventh access by the customer using his latest utterances (method 1). In the second method, it was updated each time the system was accessed by the customer (method 2). The procedure for establishing the initial reference template is the following. The first training utterance is used as a basic utterance, to which the second is brought into time registration. After registration the time functions of the feature parameters of the first two utterances are averaged and the third is brought into time registration with the averaged function and then averaged into it. When five utterances are used to construct the reference template, the fourth and fifth utterances are also brought into time registration and included in the averaging.

The training utterances are also used for the calculation of the weighting function which is used in the distance measure of (10a) and (10b), the interspeaker to intraspeaker variability ratio of (4) which is used in feature selection, and the interspeaker distance distribution which is used to set the decision threshold using (12).

III. Sample Utterances

Several kinds of utterance sets were used to evaluate this system. Fig. 6 is a block diagram which shows the procedures used to create the utterance sets. The speech was uttered in a sound booth and recorded over conventional dialed up telephone lines or a high-quality microphone. The signal was bandlimited from 100 to 3200 Hz, which is the nominal telephone bandwidth. The telephone speech was processed by the following three transmission systems:

1) clear channel—i.e., no additional processing,

2) adaptive differential pulse code modulation (ADPCM) coding,

3) linear predictive vocoding (LPC).

The ADPCM coder used in this experiment was a simulation of the coder built by Bates [12], based on the work of Cummiskey *et al.* [13]. Fig. 7 shows a block diagram of the

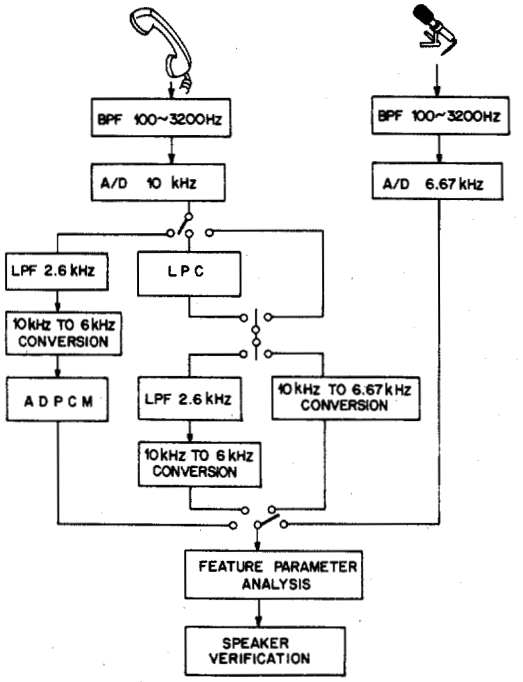


Fig. 6. Block diagram indicating the procedure to make the utterance sets.

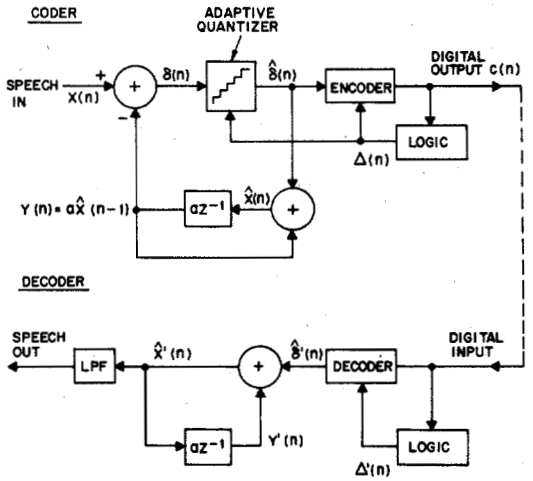


Fig. 7. Block diagram of the ADPCM system.

ADPCM system. Since the required sampling rate for the ADPCM coder was 6 kHz, a sampling rate conversion system was used to convert it from 10 kHz to 6 kHz at the input to the coder [14]. The signal bandwidth was reduced to 2.6 kHz for the ADPCM coder in the sampling rate conversion system. In the coder, a 4-bit adaptive quantizer was used to code the differential signal giving an overall bit rate of 24 kbits/s for the coder [13].

A block diagram of the LPC vocoder is given in Fig. 8. The implementation was based on the autocorrelation method of linear prediction [15], [16]. Pitch detection and voiced-unvoiced decision were performed using the modified autocorrelation pitch detector of Dubnowski *et al.* [17]. A 12 pole LPC analysis ($p = 12$) was performed using a pitch adaptive variable frame size, at a rate of 100 frames per second [18].

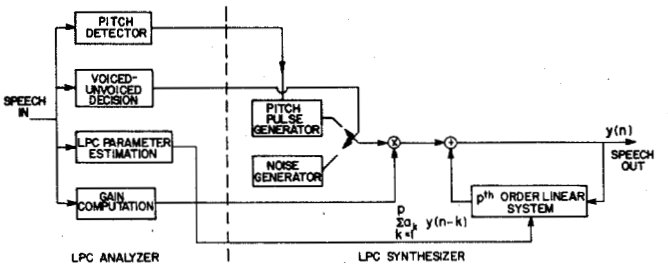


Fig. 8. Block diagram of the LPC system.

TABLE I
UTTERANCE SETS USED IN EXPERIMENTS

No.	Male or Female	Number of Customers & Impostors	Recording	Channel	Sampling Frequency
(1)	Male	10 + 40	Telephone	Clear	6.67 kHz
(2)	"	"	"	"	6
(3)	"	"	"	ADPCM	6
(4)	"	"	"	LPC	6
(5)	"	21 + 55	High Quality Microphone	Clear	6.67
(6)	Female	10 + 40	Telephone	Clear	6.67

No quantization of the LPC parameters was used in this experiment.

The sampling rate of the signal passed through the LPC vocoder or the clear channel was converted from 10 kHz to 6.67 kHz, or if bandlimited to 2.6 kHz, converted to 6 kHz.

The telephone speech utterance set includes the following.

1) 50 recordings made by each of 10 male and 10 female speakers over a period of two months. The first 10 recordings were made once a day; the remaining 40 were made twice a day (morning and afternoon). These speakers were designated "customers."

2) One recording made by each of 40 male and 40 female naive speakers. These speakers were designated "impostors." There was no attempt to mimic the "customers."

The speech recorded over a high-quality microphone was bandlimited from 100 to 3200 Hz and sampled at 6.67 kHz. The high-quality speech utterance set includes the following.

1) 26 recordings made by each of 21 male customers over a period of two months. Each was recorded on a different day.

2) One recording made by each of 55 male impostors with no attempt to mimic the customers.

Two all-voiced sentences were used in the recordings. The males used the sentence, "We were away a year ago" and the females used the sentence, "I know when my lawyer is due." Table I summarizes the six kinds of utterance sets used in this experiment. All the low-pass filters of 3.2 kHz and 2.6 kHz are digital filters, except that the 2.6 kHz low-pass filter applied to ADPCM speech is an analog hardware filter.

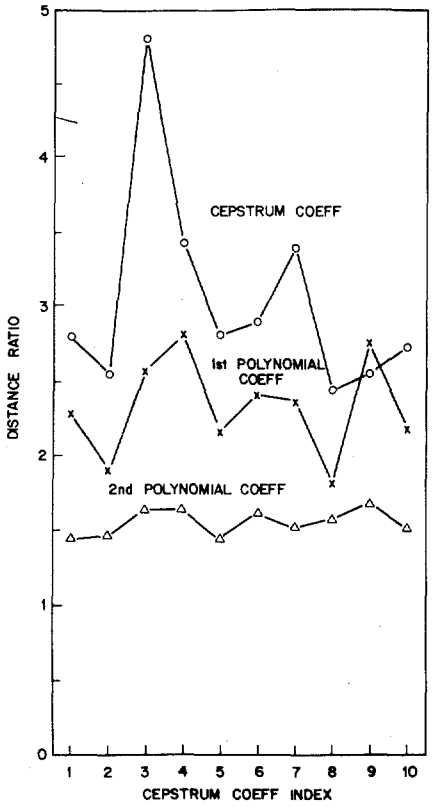


Fig. 9. Interspeaker to intraspeaker distance ratio for the first ten utterances by five speakers each extracted from utterance set (1).

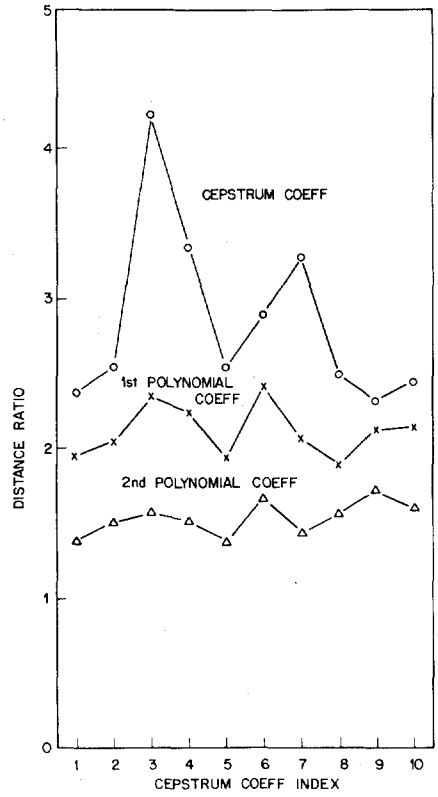


Fig. 10. Interspeaker to intraspeaker distance ratio for the middle ten utterances by five speakers each extracted from utterance set (1).

IV. Experimental Results

A. Results for Utterance Set (1)

The first experiment was performed using utterance set (1), which is a set of utterances by ten male customers and 40 male impostors recorded over a conventional telephone connection, transferred through a clear channel and sampled at 6.67 kHz. The distance measure D_1 is used in the experiments in Sections IV-A-D. The cepstrum normalization technique using the averaged value of the cepstrum is not applied to the experiments in these sections.

1) Distance ratio: In order to evaluate the feature parameters from the viewpoint of their effectiveness for speaker verification, the ratio of the average value of interspeaker distance to the average value of intraspeaker distance defined by (4), was calculated for each parameter.

Fig. 9 shows the results for an utterance set which uses the first 10 utterances by five customers each. Fig. 10 shows the results for an utterance set which comprises the middle 10 utterances by five customers each. It can be seen from these figures that all of these parameters have distance ratios greater than one, which means that all of them are useful to distinguish speakers. The original cepstrum coefficients are generally most effective and the higher the order of the polynomial coefficients becomes, the less effective they become, irrespective of the order of the cepstrum.

A preliminary experiment indicated that using utterances from ten customers in the feature selection process produced no improvement in speaker verification accuracy over the procedure described here using five customers.

TABLE II
Average Error Rates. Utterance Set: No. (1). FR: False Rejection (False Alarm). FA: False Acceptance (Miss Rate).

Threshold		FR	FA	$\frac{FR + FA}{2}$
A Priori	Estimated	0.29%	0.08%	0.19%
	Fixed	0.29%	0.31%	0.30%
A Posteriori		0%		

Based on these results, 18 parameters which have a relatively large distance ratio were selected. These are all ten cepstrum coefficients and all the first-order polynomial coefficients except coefficient index numbers 5 and 8. None of the second-order polynomial coefficients were included in this selected parameter set. The choice of 18 for the number of selected parameters was decided arbitrarily.

2) Speaker verification: Table II shows the mean-error rate of speaker verification when five utterances were used to establish an initial reference template which was updated every seventh utterance by each customer. The mean interval between training and test utterances is nearly six days. Three types of decision thresholds were applied. The error rates were averaged over ten customers and presented in this table. When the threshold is set *a posteriori* the error rate is completely zero. When the threshold is set *a priori* the mean-error rate of false acceptance and false rejection can be made as small as 0.19 percent using the optimum threshold estimation tech-

TABLE III
AVERAGE ERROR RATES. UTTERANCE SET: NO. (6). FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		FR	FA	$\frac{FR+FA}{2}$
A Priori	Estimated	0.29%	0.43%	0.36%
	Fixed	0.29%	0.54%	0.42%
A Posteriori		0.06%		

nique presented in Section II-E. When the threshold is fixed to a value which is common to all customers, the mean-error rate increases to 0.30 percent. These results show that the speaker verification techniques proposed in this paper are very powerful for telephone speech.

3) Effects of time interval between training and test utterances: Intersession variability for a given speaker is one of the most important problems in speaker verification [7], [8]. In order to check the effect of the time interval between training and test utterances on speaker verification accuracy this interval was varied from six days to six weeks comparing test utterances with reference templates constructed at times corresponding to the specified intervals. In this experiment 14 utterances were used as test inputs by ten customers each, and five utterances were used to construct each reference template. The experimental results indicated that verification accuracy is not affected by time intervals between training and input utterances of at least six weeks.

B. Results for Utterance Set (6)

The second experiment was performed using utterance set (6) which comprises the utterances by ten female customers and 40 female impostors recorded over a conventional telephone connection.

The ratio of interspeaker distance to intraspeaker distance for each parameter was calculated using the first ten utterances by five customers each. The result was similar to the result for male speakers shown in Fig. 9. The original cepstrum coefficients are most effective and the second-order polynomial coefficients are less efficient than the first-order ones. This result was used for the selection of 18 parameters, which include all ten cepstrum coefficients and all the first-order polynomial coefficients except coefficients index numbers 4 and 9.

Table III presents the speaker verification results under the same conditions observed for the male speaker set described in the previous section; a reference file was constructed using five utterances and updated every seventh access by each customer. Although the error rates for the female utterance set are slightly larger than those for the male utterance set, they are still very small.

Results for a speaker verification experiment in which a reference template was constructed using five utterances and the time interval between training and test utterances was varied up to six weeks were quite similar to that of the male speaker set. There was no significant increase in error rate when the interval is extended to six weeks.

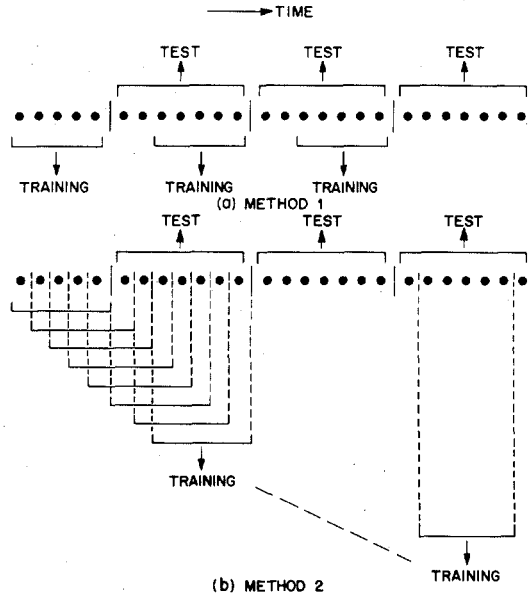


Fig. 11. Two reference updating procedures used for utterance set (5). In Method 1, a reference template is updated every seventh access by the customer. In Method 2, a reference template is updated each time the system is accessed by the customer. In both methods, the latest five utterances are used to update the reference template.

C. Results for Utterance Set (5)

In the third experiment, utterance set (5) which comprises 26 utterances by 21 male customers each and a single utterance by 55 impostors recorded over a high-quality microphone was used to test the speaker verification system. In this case the 18 selected parameters include the first nine cepstrum coefficients and the first nine first-order polynomial coefficients.

Fig. 11(a) shows the time relation between training and test utterances in speaker verification experiments for the condition that five utterances were used to construct a reference template for each customer and that it was updated every seventh access by the customer. Table IV shows error rates averaged over 21 customers. Results of the first, middle, and last seven input utterances are averaged separately. False rejection error is very large for the first seven input utterances. Initial variability like this was also shown in the previous experiment by Rosenberg [4].

In order to improve the results for the first seven input utterances, the second method for reference updating was introduced. The reference template was updated at each time of the customer's access using the latest five utterances as shown in Fig. 11(b). Table V shows the results of the verification experiment using this method. In this case, the error rate for the first seven input utterances is not significantly larger than those for the middle and last seven utterances. Compared with Table IV, it can be also concluded that frequent updating of the reference template is quite efficient for several initial input utterances but it is not necessary to do it for the remaining utterances. If we apply the second reference updating method to the first seven input utterances and the first reference updating method to the remaining utterances, we can achieve verification error rate of less than one percent using the *a priori* estimated threshold.

TABLE IV
AVERAGE ERROR RATES. UTTERANCE SET: NO. (5). REFERENCE UPDATING: METHOD 1. FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		Error	Utterances		
			First	Middle	Last
A Priori	Estimated	FR	4.76%	0.68%	0.68%
		FA	0.68%	0.73%	0.98%
		$\frac{FR+FA}{2}$	2.72%	0.71%	0.83%
	Fixed	FR	9.52%	0.68%	2.04%
		FA	0.66%	0.66%	0.88%
		$\frac{FR+FA}{2}$	5.09%	0.67%	1.46%
A Posteriori		Equal	1.17%	0.73%	0.24%

TABLE V
AVERAGE ERROR RATES. UTTERANCE SET: NO. (5). REFERENCE UPDATING: METHOD 2. FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		Error	Utterances		
			First	Middle	Last
A Priori	Estimated	FR	0.68%	0.68%	0%
		FA	0.86%	0.60%	0.60%
		$\frac{FR+FA}{2}$	0.77%	0.64%	0.30%
	Fixed	FR	1.36%	0.68%	0.68%
		FA	0.56%	0.74%	0.57%
		$\frac{FR+FA}{2}$	0.96%	0.71%	0.63%
A Posteriori		Equal	0.94%	0.17%	0.34%

D. Results for Utterance Set (3)

In the fourth experiment, utterance set (3) which comprises 50 utterances by ten male customers each and a single utterance by 40 impostors, recorded over a conventional telephone connection and transformed by a 24 kbit/s ADPCM system, was used to evaluate the speaker verification system. In this case, the 18 selected parameters include all ten cepstrum coefficients and all the first-order polynomial coefficients except coefficients with index numbers 1 and 2.

Table VI shows the results of speaker verification experiments when an initial reference template for each customer was constructed using five utterances and updated every seventh access. Although the error rates are slightly larger than those obtained for clean speech presented in Table II, both false rejection and false acceptance are still less than one percent even when the decision threshold is set *a priori* by (12). Speaker verification results showing the effect of extending the interval between training and test utterances up to six weeks indicated that the error rate was slightly greater than that obtained for clean speech.

TABLE VI
AVERAGE ERROR RATES. UTTERANCE SET: NO. (3). FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		FR	FA	$\frac{FR+FA}{2}$
A Priori	Estimated	0.29%	0.83%	0.56%
	Fixed	1.43%	1.46%	1.45%
A Posteriori		0.04%		

TABLE VII
AVERAGE ERROR RATES BY ESTIMATED *a priori* THRESHOLD.

Utterance set	No. (1)	No. (6)	No. (5)	No. (3)
Customers	10 Male	10 Female	21 Male	10 Male
Impostors	40 Male	40 Female	55 Male	40 Male
Transmission	Telephone	Telephone	Microphone	Telephone 24 kb/s ADPCM
False Rejection (False Alarm)	0.29%	0.29%	0.68%	0.29%
False Acceptance (Miss Rate)	0.08%	0.43%	0.86%	0.83%
Average	0.19%	0.36%	0.77%	0.56%
Number of Trials	5,500	5,500	12,726	5,500

Table VII shows the summary of the results of speaker verification experiments for utterance sets (1), (3), (5), and (6), using the *a priori* threshold specified by (12). For utterance set (5), reference templates were updated following each access for the first seven test utterances using the latest five utterances. After the seventh utterance, updating was carried out every seventh access. For all other utterance sets updating was carried out only after each seventh access. For all utterance sets except (5) there were 35 customer test utterances and 515 impostor test utterances per customer for a total of 350 customer and 5150 impostor trials, respectively. For utterance set (5) there were 21 customer test utterances and 585 impostor test utterances per customer for a total of 441 customer and 12 285 imposter trials, respectively.

Although this table indicates a higher error rate for microphone speech than for telephone speech, the difference is statistically insignificant since the number of utterances which caused the verification error is very small.

V. Experiments Using Mixed Transmission Systems

A. Experimental Design

In order to investigate the effects of several transmission systems on the speaker verification system more thoroughly, the reference and test utterances were subjected to different transmission systems and evaluated using the same techniques used in previous experiments with homogeneous transmission conditions. One difference between the techniques used in previous experiments and this experiment is that the transmission characteristics normalization method, subtracting the time averages from cepstrum coefficients, described in Section II-A, was applied to all utterances in this experiment.

Utterance sets (2), (3), and (4), each of which comprises 50 utterances by ten male customers each and a single utterance by 40 male impostors were used in the experiment. They were recorded over a conventional telephone connection transmitted over clear, ADPCM and LPC vocoder channel, respectively. All of these utterances were sampled at 6 kHz.

B. Result of Preliminary Experiments

1) Distance ratio: Ten utterances by five customers each were used to calculate the ratio of averaged interspeaker distance to averaged intraspeaker distance for each feature parameter. In the ten utterances, five utterances were transmitted over clear channel and the remaining five utterances were transmitted over the ADPCM system for each speaker. The results which were similar to that obtained for the utterance set which comprises only ADPCM speech indicated that the feature parameters, especially normalized cepstrum coefficients and the first-order polynomial coefficients, have a great amount of individual information which is not affected by the difference between clear and ADPCM channels.

2) Comparison between two distance measures: Before starting the speaker verification experiment which uses different combinations of the utterance sets, a preliminary experiment was carried out to compare speaker verification performance using the two distance measures D_1 and D_2 described in Section II-D.

In this experiment the reference template for each customer was constructed by the training utterances transmitted over the ADPCM system, and test utterances were transmitted over the LPC vocoder system. The results are given in Table VIII showing error rates for the two distance measures. The error rates are quite similar although the error rates obtained using D_2 are generally somewhat smaller than those obtained using D_1. The correlation coefficient between the two sets of distances is 0.992. The calculation time for D_2 is much smaller than D_1. Based on these results, D_2 is used hereafter in the transmission systems experiments.

C. Result of Speaker Verification Experiments

Table IX shows the summary of the results of speaker verification experiments for nine combinations of transmission systems. When the reference and test utterances are subjected to different transmission systems, the error rate is slightly larger than the error rate which is obtained when all the utterances are subject to the same transmission system. But even in the worst case, which is the combination of ADPCM and LPC vocoded speech, the average error rate by the estimated *a priori* threshold is only one percent. It means that the speaker verification method investigated in this paper has little degradation even when the reference and test utterances are subjected to different transmission systems.

Fig. 12 shows plots of false rejection and flase acceptance for each transmission system combination as a function of individual customers. Part (a) shows false rejection rates and

TABLE VIII
Average Error Rates. Training Utterances: ADPCM. Test Utterances: LPC. FR: False Rejection (False Alarm). FA: False Acceptance (Miss Rate).

Distance Measure	Threshold		FR	FA	$\frac{FR+FA}{2}$
D_1 (square)	A Priori	Estimated	1.71%	0.70%	1.21%
		Fixed	2.00%	2.04%	2.02%
	A Posteriori		0.16%		
D_2 (absolute)	A Priori	Estimated	1.14%	1.13%	1.14%
		Fixed	1.71%	2.08%	1.90%
	A Posteriori		0.12%		

TABLE IX
Average Error Rates. FR: False Rejection (False Alarm). FA: False Acceptance (Miss Rate).

Training	Threshold	Error	Test		
			Clear	ADPCM	LPC
Clear	A Priori (Estimated)	FR	0.29%	1.43%	0.86%
		FA	0.62	0.64	0.74
		$\frac{FR+FA}{2}$	0.46	1.04	0.80
	A Posteriori	Equal	0.02	0.08	0.02
ADPCM	A Priori (Estimated)	FR	1.43	0.86	1.14
		FA	0.66	0.95	1.13
		$\frac{FR+FA}{2}$	1.05	0.91	1.14
	A Posteriori	Equal	0.08	0.06	0.12
LPC	A Priori (Estimated)	FR	0.57	1.14	0.29
		FA	0.97	0.80	1.38
		$\frac{FR+FA}{2}$	0.77	0.97	0.84
	A Posteriori	Equal	0.04	0.19	0.02

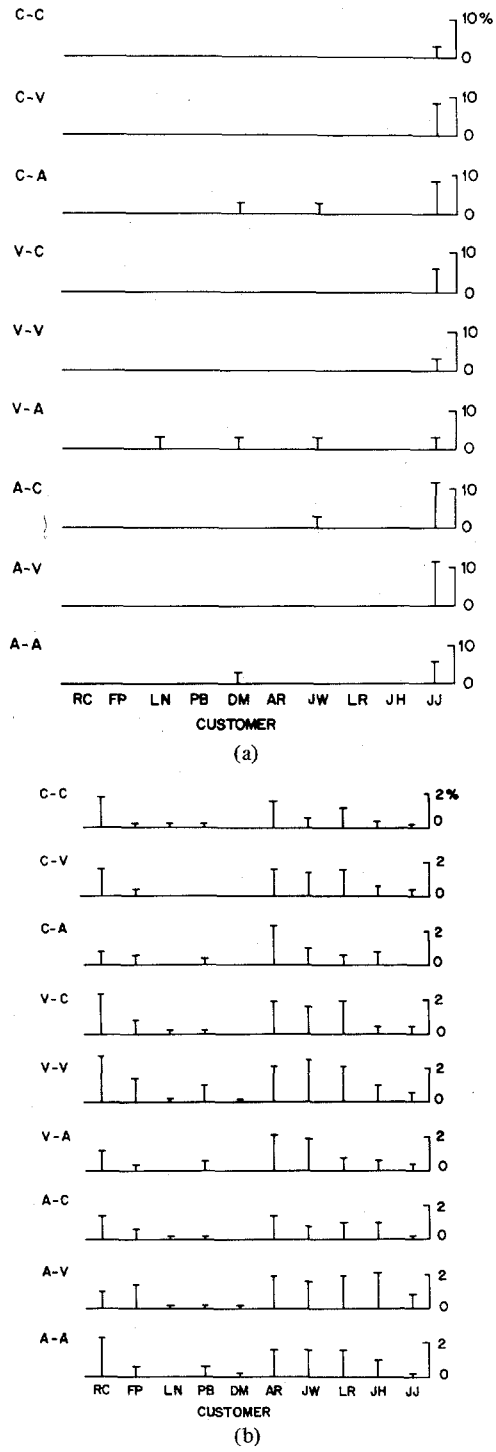


Fig. 12. False rejection (a) and false acceptance (b) as a function of the training system testing system pair and customer. C—clear channel, V—LPC vocoder system, A—ADPCM system.

part (b) shows false acceptance rates. The reader should note that the scales of the two figures are different. A high degree of variability in scores exists among customers for each pair of transmission systems. The variability between scores for pairs of transmission system is almost negligible compared with the variability of scores within a pair of transmission systems.

Table X shows error rates when cepstrum normalization is omitted for the combination of LPC vocoder and ADPCM

TABLE X
AVERAGE ERROR RATES WITH NO CEPSTRUM NORMALIZATION. TRAINING UTTERANCES: ADPCM. TEST UTTERANCES: LPC.

Threshold	Error	Error Rate
A Priori (Estimated)	False Rejection (False Alarm)	1.14%
	False Acceptance (Miss Rate)	1.30%
	Average	1.22%
A Posteriori	Equal	0.52%

channel transmissions for test and training utterances, respectively. The distances between reference and test utterances are generally much greater than those obtained when cepstrum normalization is applied. Accordingly, the parameter b in the threshold estimation equation is changed to a value which is appropriate to make the two kinds of error rates almost same. The larger error rates obtained when cepstrum normalization is omitted confirms the effectiveness of the cepstrum normalization technique.

The error rates for the homogeneous conditions in Table IX are slightly different from the previous results described in Sections IV-A and IV-D, since the sampling frequency of "clear" speech is different between the previous and present experiments, and the cepstrum normalization technique was not applied in the previous experiments. From these comparisons, it is apparent that when the difference of the transmission characteristics between reference and test utterances is small, cepstrum normalization slightly increases the verification error rate by removing the long-term speaker-related information.

In the next section, the effectiveness of the cepstrum normalization will be investigated using an utterance set which has very large differences between the transmission characteristics of reference and test utterances.

D. Experiments with Artificial Transmission Variation

The utterance set by ten male customers and 40 male impostors recorded over a conventional telephone connection was used in a speaker verification experiment. All utterances were passed through a 3 kHz low-pass filter and sampled at 6.67 kHz. Training utterances were processed with preemphasis, whereas preemphasis was omitted for test utterances. This results in a simple but large difference in frequency characteristics between training and test utterances. Two experiments were performed to study the effect of cepstrum normalization; verification using normalized cepstrum and verification using unnormalized cepstrum. The results are shown in Table XI. There are very large differences between the results for normalized cepstrum and unnormalized cepstrum. It is evident that cepstrum normalization is very powerful, and small error rates can be obtained even when there are large frequency characteristic differences between the training and test utterances.

TABLE XI
AVERAGE ERROR RATES. TRAINING UTTERANCES: PROCESSED BY PREEMPHASIS. TEST UTTERANCES: UNPROCESSED BY PREEMPHASIS. FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Cepstrum Normalization	Threshold		FR	FA	$\frac{FR+FA}{2}$
YES	A Priori	Estimated	0.86%	0.64%	0.75%
		Fixed	0.86%	0.78%	0.82%
	A Posteriori		0.14%		
NO	A Priori	Estimated	10.57%	13.15%	11.86%
		Fixed	17.14%	16.49%	16.82%
	A Posterior		9.66%		

VI. DISCUSSION

A. Comparison Between Cepstrum and Log Area Ratio

In order to check the advantage of the cepstrum coefficients derived through LPC analysis (LPC-cepstrum), which have been used in this paper, log area ratio parameters, which are arctanh transformation of PARCOR coefficients, were extracted from the utterance set (1) and studied. Log area ratios were found to be very good parameters for speaker verification in previous experiments by the author [19]. Fig. 13 shows distance ratios for each time function of log area ratios and polynomial coefficients derived from them. The results for cepstrum coefficients which were extracted from the same utterance set was shown in Fig. 10. Comparing Figs. 10 and 13, it can be seen that cepstrum coefficients are more efficient than log area ratios for speaker verification.

Table XII shows the results of a speaker verification experiment using log area ratios and polynomial coefficients derived from them compared to the results using cepstrum coefficients. In this experiment, 10 utterances by 10 customers each and a single utterance by 40 impostors were used. The first three utterances were used to construct a reference template for each customer, and the remaining seven customer utterances and impostor utterances were used as test utterances. A constrained endpoint dynamic time warping technique was used in this experiment. The error rate results show that cepstrum coefficients have an advantage over log area ratios.

Fig. 14 shows examples of spectral envelopes derived from 10 cepstrum coefficients or 10 log area ratios for a spoken sentence "We were away a year ago." Log area ratios are transformed into linear predictor coefficients and the spectral envelope is computed using the correlation function of the coefficients. Time sequences of the envelope for the first 100 frames are shown in these figures. The frame interval is 10 ms. Spectral envelopes derived from cepstrum coefficients are much smoother than those derived from log area ratios along both the frequency axis and time axis. In other words, the spectral envelope sequence by cepstrum coefficients is more stable than that obtained using log area ratios.

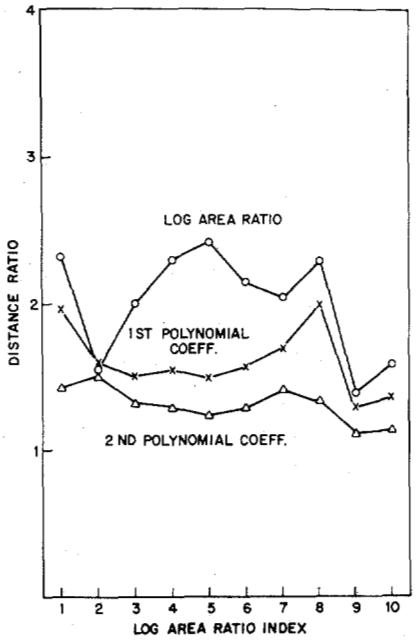


Fig. 13. Interspeaker to intraspeaker distance ratio for each time function of log area ratios and polynomial coefficients derived from them. Ten utterances by five male speakers each were used for the analysis.

TABLE XII
AVERAGE ERROR RATES. UTTERANCE SET: SUBSET OF THE UTTERANCE SET (1). THRESHOLD: *A posteriori* EQUAL ERROR THRESHOLD. TIME REGISTRATION: CONSTRAINED ENDPOINT DYNAMIC TIME WARPING. NUMBER OF TRAINING UTTERANCES: 3.

Feature Parameters	Error Rate
Cepstrum	0.80%
Log Area Ratio	1.59%

Fig. 15 shows comparisons of four kinds of spectra; short time speech spectrum, spectral envelope derived from log area ratio, spectral envelope derived from LPC-cepstrum, and spectral envelope derived from conventional cepstrum coefficients which are extracted through Fourier transformation of the log power spectrum (FFT-cepstrum). Results for two frames in the sentence are presented in the figure. It can be seen that spectral envelopes derived from LPC-cepstrum and FFT-cepstrum are quite similar and are much smoother than those derived directly from LPC parameters, which is very sensitive to spectral peaks.

B. Comparison Between LPC-Cepstrum and FFT-Cepstrum

As indicated in Fig. 15, a spectral envelope derived from the LPC-cepstrum is slightly different from a spectral envelope derived from the FET-cepstrum. In order to study the effect of this difference on speaker verification, several experiments were performed using utterance set (6) which consists of female utterances. The size of the time window was set to 256 samples (38.4 ms) to extract the FFT-cepstrum, while the window size used to extract LPC-cepstrum was 30 ms. The FFT-

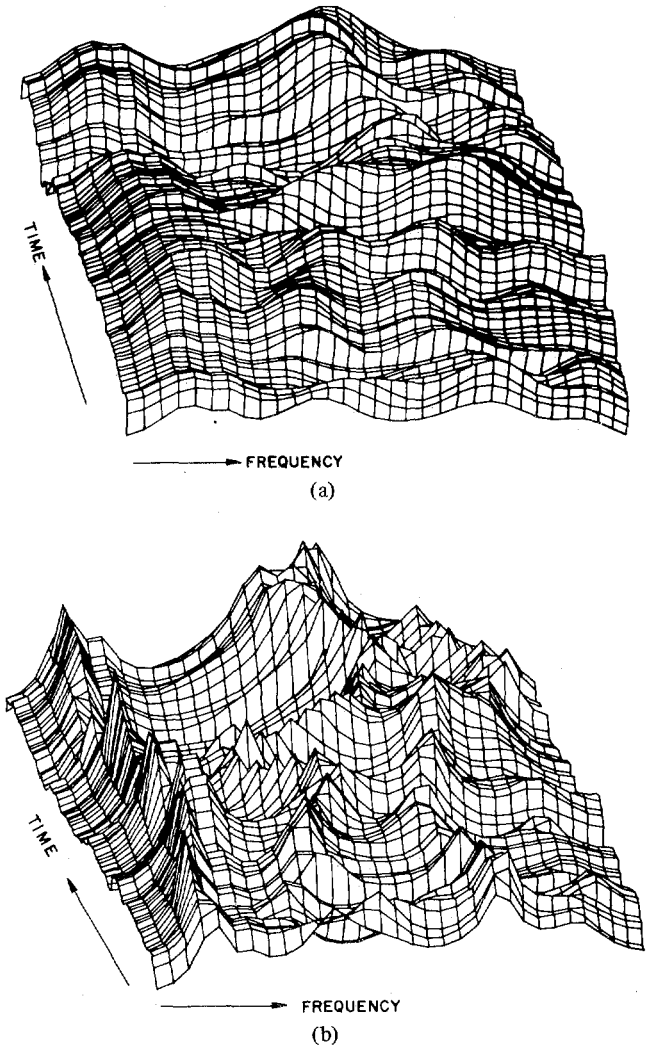


Fig. 14. Spectral envelopes derived from ten cepstrum coefficients (a) or ten log area ratios (b) for a spoken sentence, "We were away a year ago." Time sequences of the envelope for the first 100 frames (1 s long) are shown.

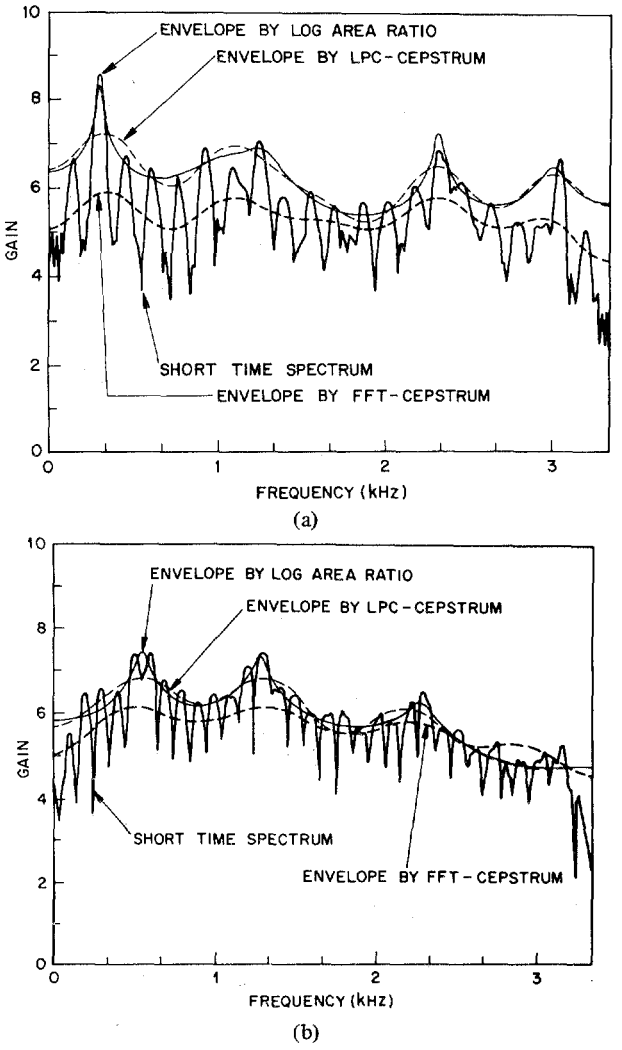


Fig. 15. Comparison of four kinds of spectra; short time speech spectrum, spectral envelope derived from log area ratio, spectral envelope derived from LPC-cepstrum, and spectral envelope derived from FFT-cepstrum. (a) For the sound /i/ in "We . . . ." (b) For the sound /o/ in ". . . ago."

cepstrum computation time which includes two Fourier transformations is almost twice that of the LPC-cepstrum.

The distance ratio for each parameter derived from the FFT-cepstrum was calculated. The result was similar to that for the LPC-cepstrum except that the speaker dependent information in the FFT-cepstrum tends to concentrate in the first-order cepstrum. Overall average distance ratios for cepstrum coefficients, the first-order polynomial coefficients and the second-order polynomial coefficients are 2.15, 1.89, and 1.45, respectively, for LPC-cepstrum, and 2.07, 1.83, and 1.37, respectively, for FFT-cepstrum. LPC-cepstrum has slightly larger values than FFT-cepstrum, but the difference is very small.

Table XIII shows the results of a speaker verification experiment using FFT-cepstrum under the same condition as that using LPC-cepstrum whose results were presented in Table III. The difference in error rates between these two experiments is very small. Speaker verification results using FFT-cepstrum when the interval between training and test utterances was long was similar to the results obtained using LPC-cepstrum.

It is apparent that the LPC-cepstrum produces almost the same results in speaker verification as the conventional FFT-cepstrum, while it takes only half the time to calculate the LPC-cepstrum compared with the FFT-cepstrum.

TABLE XIII
AVERAGE ERROR RATES. UTTERANCE SET: NO. (6) (FFT-CEPSTRUM). FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		FR	FA	$\frac{FR+FA}{2}$
A Priori	Estimated	0.29%	0.33%	0.31%
	Fixed	0.86%	0.74%	0.80%
A Posteriori		0.02%		

C. *Effectiveness of Polynomial Coefficients*

In order to study the effectiveness of the use of polynomial coefficients on speaker verification, an additional experiment was performed in which the polynomial coefficients were

TABLE XIV
AVERAGE ERROR RATES. TRAINING UTTERANCES: PROCESSED BY PREEMPHASIS. TEST UTTERANCES: UNPROCESSED BY PREEMPHASIS. FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Feature Parameters	Threshold		FR	FA	$\frac{FR+FA}{2}$
Cepstrum Coefficients	A Priori	Estimated	2.29%	1.69%	1.99%
		Fixed	3.14%	2.31%	2.73%
	A Posteriori		0.64%		
Cepstrum and Polynomial Coefficients	A Priori	Estimated	0.86%	0.64%	0.75%
		Fixed	0.86%	0.78%	0.82%
	A Posteriori		0.14%		

omitted using only the time functions of the first to tenth cepstrum coefficients. The training and test utterance recording conditions are the same as the experiment described in Section V-D, where preemphasis is applied to the training utterances but omitted for the test utterances. Cepstrum normalization was applied to all utterances. Table XIV shows the verification error rates including previous results which were obtained when polynomial coefficients are included. It can be seen that error rates are increased by a factor of three or more when polynomial coefficients are omitted.

D. Optimum Length of Speech Segment for Polynomial Expansion

In the experiments described so far, the length of the speech segment for which time functions of cepstrum coefficients are expanded by an orthogonal polynomial representation has been set to 90 ms. This value was determined to be adequate for preserving transitional information between phonemes. In order to check the appropriateness of this value of length, additional speaker verification experiments were performed varying the length between 50 ms and 210 ms. Training utterances were processed with preemphasis but test utterances were not. Cepstrum normalization was applied to all utterances. The condition for the experiment in the previous section corresponds to the condition of zero length segment. The speaker verification error rates with *a priori* or *a posteriori* threshold are plotted in Fig. 16 as a function of segment length, including the results of the previous section. For the *a priori* threshold condition, the averaged values of false acceptance and false rejection rates are plotted. The verification error rate is a minimum for 170 ms and the error rate increases both for shorter and longer lengths.

Although 90 ms is not the optimum value, the difference between the error rates for 90 ms and the optimum value of 170 ms is small. The number of computation increases in proportion to the segment length. Based on these considerations,

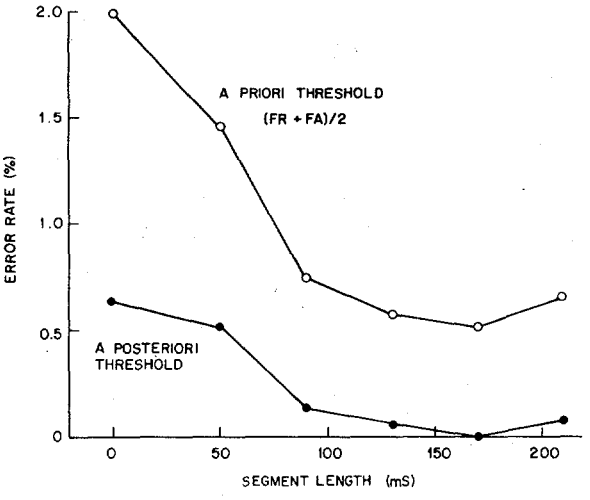


Fig. 16. Error rate versus the length of the speech segment for orthogonal polynomial expansion. Training utterances were processed with preemphasis, whereas test utterances were not.

TABLE XV
AVERAGE ERROR RATES. UTTERANCE SET: NO. (1). TIME REGISTRATION: CONSTRAINED END-POINT DYNAMIC TIME WARPING. FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		FR	FA	$\frac{FR+FA}{2}$
A Priori	Estimated	0.40%	0.43%	0.42%
	Fixed	0.57%	0.58%	0.58%
A Posteriori		0.04%		

it can be concluded that 90 ms is a reasonable value for polynomial expansion in this speaker verification system.

E. Comparison Between Unconstrained Endpoint and Constrained Endpoint Dynamic Time Warping Methods

Table XV shows speaker verification results when constrained endpoint dynamic time warping is used. Other conditions are the same as the experiment whose results were shown in Table II. The error rate using the constrained endpoint method is almost twice as that using the unconstrained endpoint method. This result shows the advantage of the unconstrained endpoint dynamic time warping method, which can cope with the uncertainty in the location of both the initial and final frames due to breath noise, etc., over the constrained endpoint method.

F. Effectiveness of Dynamic Time Warping Guided by Shorter One of Either Input or Reference

In Section II-C it was stated that optimum matches are obtained by using as guide the shorter of either the input or reference contours. This warping procedure was used in all the speaker verification experiments described in this paper. To show the effect of not observing this procedure an additional experiment was performed. Utterance set (5) was used in a speaker verification experiment in which the input utterance exclusively was used as the guide. This is referred to as

the UEGI (unconstrained endpoint guided by input) method, in contrast to the UEGS method [unconstrained endpoint guided by (the) shorter (of reference and input)] adopted in all other experiments.

A reference template of each customer was constructed using five utterances and updated at every access by the customer for the first seven test utterances (method 2) and updated every seventh access by the customer for the remaining test utterances (method 1). Decision thresholds were set *a priori* based on (12).

Mean error rates for this utterance set using the UEGI procedure are plotted in Fig. 17 along with the results obtained earlier using the UEGS procedure. It can be seen that, although false acceptance error rates are comparable for the two techniques, false rejection rates for the first and middle seven input utterances are much greater for the UEGI procedure. This outcome may be attributed to the fact that until stable references are established by updating, the lengths of reference and input utterances are quite variable. Therefore, large discrepancies can be expected between the UEGI and UEGS procedures. However, with stable references associated with the last seven input utterances the lengths of input and reference utterances are more consistent and little or no discrepancy is expected between the two techniques.

When warping is guided by the input utterance, the first frame of the input may be warped to the first through $(\delta + 1)$th frame of the reference where δ specifies the width of the allowable range. Similarly, when warping is guided by the reference, the first frame of the reference may be warped to the first through $\delta + 1$th frame of the input.

An experiment was carried out using 26 utterances from each of the 21 male customers in utterance set (5). Each customer's utterances were paired with the customer's reference and matched two ways, using the input utterance as the guide and the reference utterance as the guide. For each such pair, the slave frame matched to the first guide frame for the better of the two matches (the match resulting in the lower overall distance) was tabulated. This tabulation is presented in the histograms of Fig. 18.

Along the abscissa is plotted the slave frame minus one (matched to guide frame number one) with the input as slave plotted along the positive axis and the reference as slave plotted along the negative axis. Equivalently, the positive axis represents matches in which the reference is guide while the negative axis represents matches in which the input is guide. Each histogram point represents the number of optimum matches corresponding to the indicated slave frame. The region enclosed by the shaded vertical bars represents optimum matches which can be obtained by using either the reference as guide or input as guide. For example, optimum matches within the shaded region to the left of the y-axis, which are actually obtained using the input as guide, are substantially the same when guided by the reference, matching the first reference frame to the first input frame. Thus, from Fig. 18(a) all but 9.5 percent of the optimum matches are obtained by using the reference as guide, while all but 5.7 percent are obtained by using the input as guide.

Fig. 18(b) and (c) decompose the matches of Fig. 18(a) into two categories. In Fig. 18(b) all the optimum matches are shown for which the input length is less than or equal to the reference length, while in Fig. 18(c) are shown all the optimum matches for which the reference length is less than the input length. It can be seen immediately that the greatest number of optimum matches is associated with using as guide the shorter of the input and reference. That is, in Fig. 18(b) all but 2.6 percent of the optimum matches are obtained by using the input as guide, while in Fig. 18(c) all but 3.7 percent of the optimum matches are obtained by using the reference as guide.

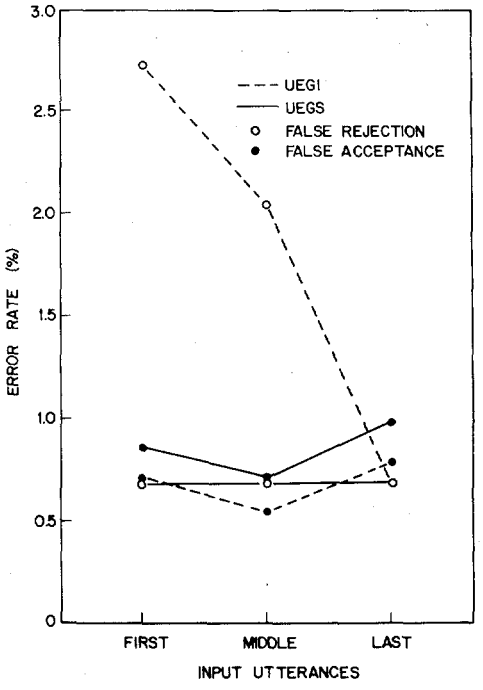


Fig. 17. Comparison of error rates for two dynamic time warping techniques; UEGI and UEGS. Results for utterance set (5).

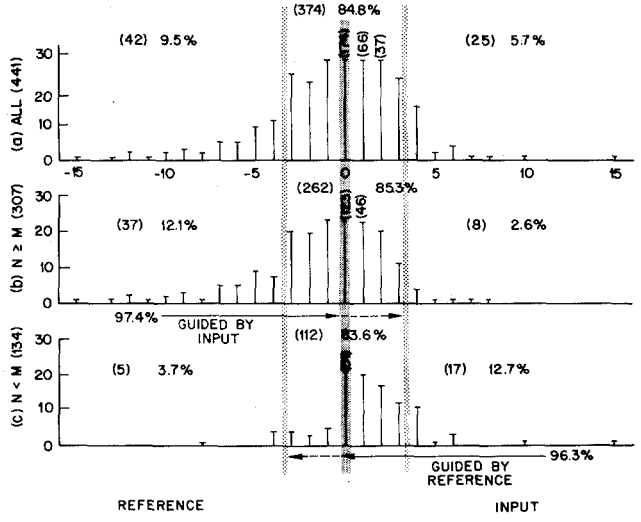


Fig. 18. Histograms for the starting frame of the optimum warping function, where a positive number means that warping function starts from the input axis and a negative number means that it starts from the reference axis. N: number of frames of reference function, M: number of frames of input function.

TABLE XVI
AVERAGE ERROR RATES. UTTERANCE SET: NO. (1). TIME REGISTRATION CONSTRAINED END-POINT DYNAMIC TIME WARPING. NUMBER OF TRAINING UTTERANCES: 3. FR: FALSE REJECTION (FALSE ALARM). FA: FALSE ACCEPTANCE (MISS RATE).

Threshold		FR	FA	$\frac{FR + FA}{2}$
A Priori	Estimated	0.69%	0.87%	0.78%
	Fixed	1.14%	1.15%	1.15%
A Posterior		0.14%		

G. Effect of the Number of Training Utterances

Table XVI shows the results of a speaker verification experiment in which three utterances were used to construct a reference template. Other conditions are the same as the experiment whose result was shown in Table XV (Section XI-E), which means that constrained endpoint dynamic time warping method was used. Comparing Table XV and XVI, it can be seen that using three utterances to make a reference template is not adequate. The error rate becomes almost twice that obtained when five training utterances are used.

Next, the number of training utterances was increased to ten, and a speaker verification experiment was performed. However, it produced no improvement compared with the results using five training samples to construct a reference template. It can be concluded that five utterances are necessary and sufficient to make a reference template.

H. Threshold Estimation

In this paper, (12) is used to set an *a priori* decision threshold for each customer. This equation and two parameters in it were determined experimentally. Fig. 19 shows the relation between $\hat{\mu}_{DB}(k) - \hat{\sigma}_{DB}(k)$ and equal error threshold $\theta_{eq}(k)$ which produces the equal error of false acceptance and false rejection. This is the result of the speaker verification experiment using utterance set (3) which produced the error rates shown in Table VI. The correlation coefficient between $\hat{\mu}_{DB}(k) - \hat{\sigma}_{DB}(k)$ and $\theta_{eq}(k)$ calculated from these values is 0.753. This result indicates the appropriateness of using (12) to estimate the optimum decision threshold.

To determine the effect of varying the parameter b in (12) on the error rate, all the customer utterances and impostor utterances which were tested in the speaker verification experiment using the mixed transmission system (Section V) were scanned by varying the parameter b. The parameter a was set to 0.6, which was determined experimentally. The number of errors was tabulated at each step by comparing the actual overall distances with the estimated threshold using the varied parameter value b. False acceptance rate and false rejection rate were averaged and plotted in Fig. 20. Results for nine conditions, which are nine combinations of three training systems and three testing systems, are shown. As the result of increase in false acceptance rate for large threshold values and increase in false rejection rate for small threshold values, the averaged error rate has a concave slope as a function

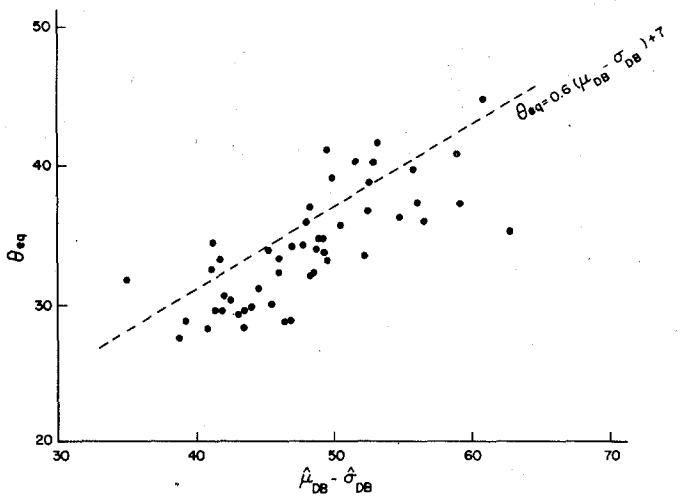


Fig. 19. Relation between $\hat{\mu}_{DB}(k) - \hat{\sigma}_{DB}(k)$ and equal error threshold θ_{eq}. Results of the speaker verification experiment using utterance set (3).

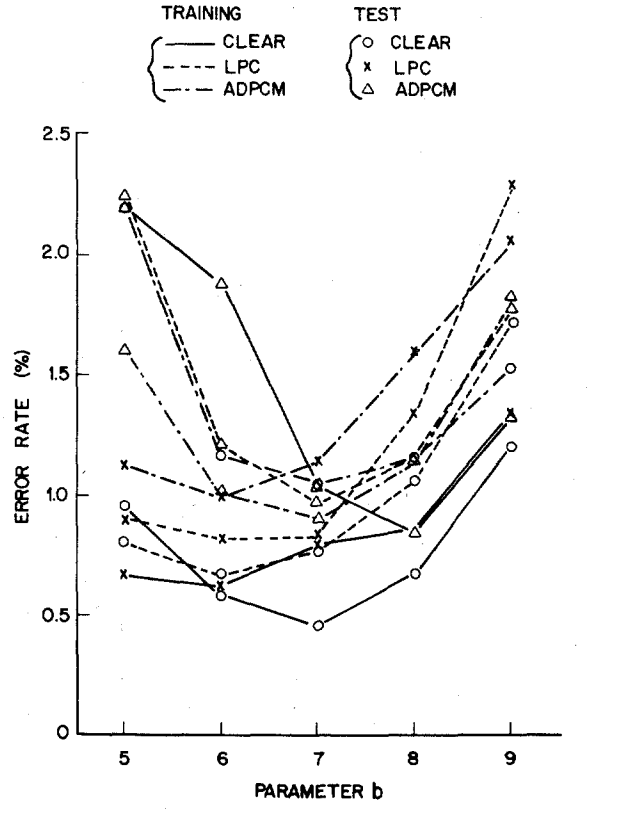


Fig. 20. Error rate versus b parameter in the threshold estimation equation. The effects of parameter variation on the average of false rejection and false acceptance error rates are shown using the estimated decision threshold. Results for nine training system testing system pairs are plotted.

of b. The optimum value of the parameter b, which produces the minimum average error rate, is seven almost irrespective of the experimental condition. This is the value consistently used in this paper to estimate an optimum threshold for each customer. As there is some tradeoff between the two kinds of error rates, if it is desirable to keep the false acceptance rate at a much lower value, the parameter b should be set at a value smaller than seven, even though it increases the false rejection

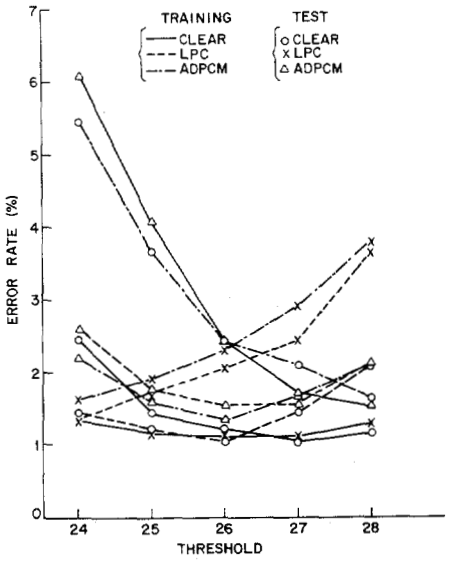


Fig. 21. Error rate versus decision threshold. The effects of threshold variation on the average of false rejection and false acceptance error rates are shown using the threshold. Results for nine training system-testing system pairs are plotted.

rate. Conversely, values of b larger than seven produce a smaller false rejection rate and a larger false acceptance rate. The dashed line in Fig. 19 indicates the relation

$$\theta_{eq} = 0.6\,(\hat{\mu}_{DB} - \hat{\sigma}_{DB}) + 7. \tag{13}$$

Fig. 21 shows verification error rate, which is the mean value of false acceptance rate and false rejection rate, as a function of decision threshold for the same experimental conditions. Results for the nine conditions are plotted. The reader should note that the scale of this figure is different from the previous one. This result shows that the optimum value of the threshold varies considerably depending on the utterance set. Thus, it is very difficult to set the threshold in advance independently of the utterance set.

These results indicate the effectiveness of the threshold estimating method using (12).

I. Withholding Decision

Another tabulation was carried out to assess the effect on error rate of withholding decision (sequential decision) on trials for which $|D_T - \theta| \leqslant \Delta$, where D_T and θ are the overall distance and threshold, respectively. When the decision is withheld on a given utterance, a new distance is calculated which is the mean of the distances of the withheld utterances and the succeeding utterance. Utterance sets (1), (3), (5), and (6) were used in this experiment. As these utterance sets include only one utterance for each impostor, impostor utterances were not used in this experiment.

Fig. 22 shows error rates as a function of the withholding threshold Δ. Part (a) shows the results when reference templates were updated every seventh trial for each customer, and part (b) shows averaged error rates over several conditions when the interval between training and test utterances were varied from six days to six weeks. Fig. 23 shows the percentage of withheld trials, which is the percentage of addi-

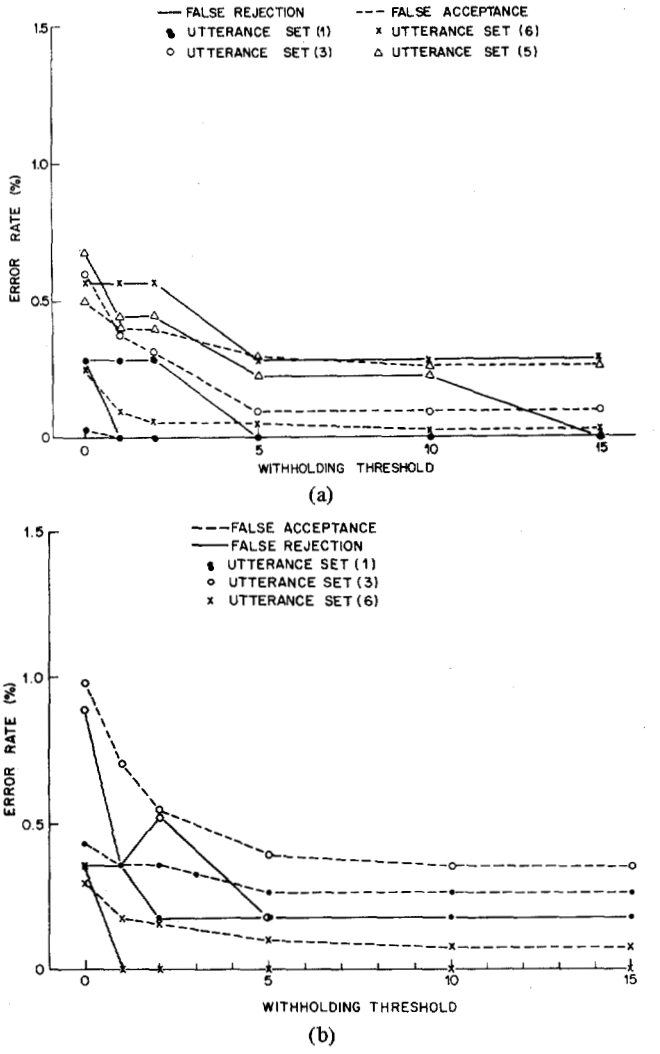


Fig. 22. Error rate versus withholding threshold. (a) False rejection and false acceptance rates on the condition that a reference template is updated every seventh access by each customer. Mean time interval between training and test utterances is six days. (b) False rejection and false acceptance rates averaged over several conditions when the time interval between reference and test utterances is varied from six days to six weeks.

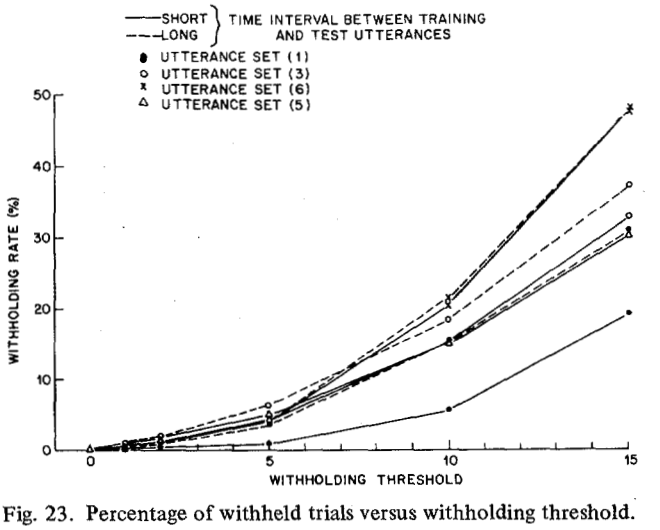
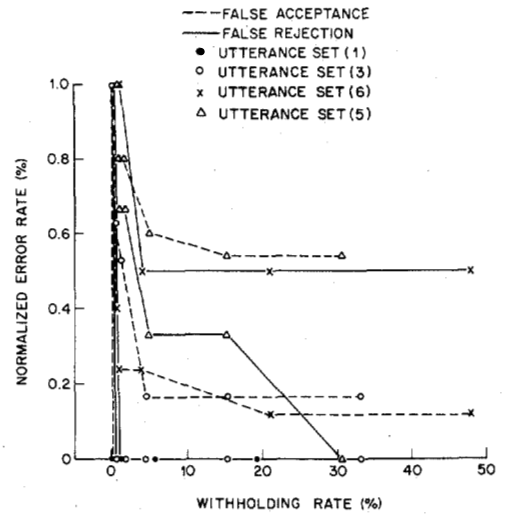
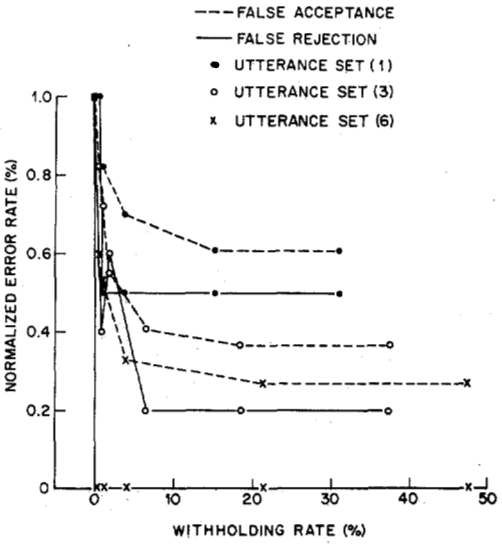


Fig. 23. Percentage of withheld trials versus withholding threshold.



(a)



(b)

Fig. 24. Normalized error rates versus percentage of withheld trials. (a) and (b) correspond to (a) and (b) in Fig. 22, respectively.

tional trials, as a function of the withholding threshold Δ. Fig. 24 shows the relation between the percentage of withheld trials and the error rate normalized by the error rate obtained without withholding. Part (a) and part (b) correspond to part (a) and part (b) in Fig. 22, respectively. Fig. 24 indicates that at least 30 percent improvement in error rate can be obtained with decisions withheld on five percent of the trials and an average 73 percent improvement can be obtained with decisions withheld on ten percent of the trials.

J. Combination with Pitch and Intensity Contours

Speaker verification systems based on pitch and intensity contours have been evaluated in Bell Laboratories using the same utterance sets used in this paper [1]-[5].

As the information conveyed by pitch and intensity contours is considered to be almost independent of that conveyed by the cepstrum, the combination of these two kinds of information will improve the performance.

In order to test the independence of these two kinds of information, the distribution of speaker verification error rates by cepstrum and that by pitch and intensity contours were compared with each other and a correlation coefficient between them was calculated. When all the error rates plotted in Fig. 12 and the error rates obtained with the same conditions using pitch and intensity contours are used, the correlation coefficient is 0.22 and -0.36 for false rejection and false acceptance, respectively. It can be concluded that the two kinds of information are fairly independent. Based on these results, an improvement in performance can be expected by combining these two kinds of information.

VII. Summary

A new system for automatic speaker verification has been implemented on a 16-bit laboratory computer and evaluated. A fixed, sentence-long utterance is analyzed by cepstrum coefficients by means of LPC analysis. Frequency-response distortions introduced by transmission systems are removed automatically. Time functions of cepstrum coefficients are expanded by orthogonal polynomial representations and compared with stored reference functions. After dynamic time warping, a decision is made to accept or reject an identity claim. Reference functions and decision thresholds are updated for each customer. The total processing time is approximately 40 times real time in this computer simulation.

In the first part of the experiment, three sets of utterances were used for the evaluation of the system. The first and second sets each comprises 50 utterances by ten customers each and a single utterance by 40 impostors recorded over a conventional telephone connection. The third set comprises 26 utterances by 21 customers each and a single utterance by 55 impostors recorded over a high-quality microphone. The first and third sets were uttered by male speakers, whereas the second set was uttered by female speakers. The evaluation indicated mean error rates of 0.19 percent, 0.36 percent, and 0.77 percent for each utterance set, respectively.

Second, the first utterance set was processed by an ADPCM coding system and an LPC coding system. These utterance sets were used for a speaker verification experiment together with an unprocessed utterance set. Experimental results indicate that the transmission system affects the verification accuracy only slightly even if the reference and test utterances are subjected to different transmission conditions.

Third, the time interval between reference and test utterances was changed from six days to six weeks. Results of the experiment indicate no significant increase of verification error with the increase of time interval.

These results verify the robustness of the new speaker verification system presented in this paper. Some discussions on new techniques used in this system are also included in this paper.

Further investigations, current or projected, include a large-scale and long-term evaluation over telephone lines permitting direct customer access and on-line response, and specialized hardware processing to improve response time.

ACKNOWLEDGMENT

The author wishes to thank J. L. Flanagan and A. E. Rosenberg for their guidance and stimulating discussions, and to acknowledge the assistance of C. A. McGonegal in providing the data base.

REFERENCES

[1] G. R. Doddington, "A computer method of speaker verification," Ph.D. dissertation, Dep. Elec. Eng., University of Wisconsin, 1970.

[2] R. C. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 80–89, Apr. 1973.

[3] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 169–176, Apr. 1975.

[4] A. E. Rosenberg, "Evaluation of an automatic speaker-verification system over telephone lines," *Bell Syst. Tech. J.*, vol. 55, pp. 723–744, July–Aug. 1976.

[5] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "The effects of ADPCM coding and LPC vocoding on an automatic speaker verification system," *J. Acoust. Soc. Amer.*, vol. 65, supplement 1, YY7, Spring 1979.

[6] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.

[7] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," *Electron. Commun.*, vol. 57-A, pp. 34–42, 1974.

[8] —, "Effects of long-term spectral variability on speaker recognition," *J. Acoust. Soc. Amer.*, vol. 64, supplement 1, NNN 28, Fall 1978.

[9] 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, June, 1974.

[10] N. L. Johnson and F. C. Leone, *Statistics and Experimental Design in Engineering and the Applied Sciences*, 2nd ed., vol. 1. New York: Wiley, 1977, pp. 471–481.

[11] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 575–582, Dec. 1978.

[12] S. L. Bates, "A hardware realization of a PCM-ADPCM code converter," M.S. thesis, Dep. Elec. Eng. and Comp. Sci., Massachusetts Institute of Technology, Cambridge, MA, Jan. 1976.

[13] P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive quantization in differential PCM coding of speech," *Bell Syst. Tech. J.*, vol. 52, pp. 1105–1118, Sept. 1973.

[14] R. E. Crochiere and L. R. Rabiner, "Optimum FIR digital filter implementations for decimation, interpolation, and narrow-band filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 444–456, Oct. 1975.

[15] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," *Reports 6th Int. Cong. Acoust.*, Y. Kohashi, Ed., vol. C-5-5, C17–20, Tokyo, Japan, 1968.

[16] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer, 1976.

[17] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 2–8, Feb. 1976.

[18] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 24–33, Feb. 1977.

[19] S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds," *Electron. Commun.*, vol. 56-A, pp. 62–71, 1973.

Sadaoki Furui (M'79) was born in Tokyo, Japan, on September 9, 1945. He received the B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan, in 1968, 1970, and 1978, respectively.

After joining the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation, Tokyo, in 1970, he studied the analysis of speaker characterizing information in the speech wave, and its application to speaker recognition and interspeaker normalization in speech recognition. He is currently a Staff Engineer at Musashino Electrical Communication Laboratory. From December 1978 to December 1979 he joined the staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, NJ, as an exchange visitor working on speaker verification.

Dr. Furui is a member of the Acoustical Society of Japan and the Institute of Electronics and Communication Engineers of Japan.