

論文 / 著書情報
Article / Book Information

Title	Speaker Recognition by Statistical Features and Dynamic Features
著者	古井貞熙
Author	SADAOKI FURUI
掲載誌/書名	, Vol. 30, No. 3, pp. 467-482
Journal/Book name	Review of the Electrical Communication Laboratories, Vol. 30, No. 3, pp. 467-482
発行日 / Issue date	1982,

Speaker Recognition by Statistical Features and Dynamic Features

Sadaoki FURUI

Performance of four text-dependent speaker recognition methods using statistical or dynamic features of speech spectrum extracted from spoken words or a short sentence were compared with each other. It has been clarified that almost the same recognition accuracies can be obtained by either statistical or dynamic features, and that the statistical features have advantages of a small amount of calculation and memory size for feature extraction and recognition. The combination effect of these two kinds of features and effectiveness of a spectral equalization technique have been ascertained. Very high speaker recognition performance can be obtained by a new method using statistical features.

1 Introduction

Automatic speaker recognition has recently received a great deal of attention among speech researchers. One of the most difficult problems in speaker recognition is that inter-session variability (variability over time) for a given speaker has a significant effect on recognition accuracy⁽¹⁾⁻⁽⁵⁾. In the speaker recognition experiment reported by Luck⁽⁵⁾, it was found necessary to include speech samples taken over a 5-week period in the reference data for an adequate representation of the speakers' voice. Furui et al.⁽⁶⁾⁻⁽⁹⁾ have investigated the long-term variability of both statistical and dynamic speaker dependent features. Statistical features are extracted from longtime averaged spectrum for a sentence-long utterance⁽⁶⁾ and time-averaged characteristics of log area ratios and fundamental frequency derived from the voiced portion of spoken words⁽⁷⁾⁽⁸⁾. Dynamic characteristics have been analyzed by the use of time functions of log area ratios and fundamental frequency and by warping functions calculated in the course of time registration⁽⁹⁾.

In this paper, the effects of long-term variability of statistical and dynamic features on speaker recognition are compared, and the effectiveness of four kinds of text-dependent speaker recognition methods, which use statistical or dynamic

features, are evaluated from this point of view. These experiments use several kinds of data bases, which consist of word utterances or sentence utterances recorded over long periods. The effectiveness of spectral equalization as a method to reduce this effect is examined using a data base which consists of word utterances recorded over seven years. A combination of statistical and dynamic features was investigated to get high speaker recognition accuracy.

An abstract of the four speaker recognition methods is shown in Table 1. Methods 1 and 2 use statistical features, whereas Methods 3 and 4 use dynamic features. In Methods 1 and 3, a spectral envelope is represented by log area ratios, whereas in Methods 2 and 4, it is represented by LPC-cepstrum coefficients. In Methods 1, 2 and 3, fundamental frequency is also used as a feature parameter. Methods 2 and 4, which have been recently investigated by the author, are improvements upon Methods 1 and 2, respectively.

Sambur⁽¹⁰⁾ and Markel et al.⁽¹¹⁾ tried speaker recognition systems using statistical features, and Wohlford et al.⁽¹²⁾ compared various statistical feature speaker recognition systems, but they have used speech utterances recorded over short periods. Doddington⁽¹³⁾ and Rosenberg⁽¹⁴⁾ tried speaker recognition systems using dynamic features based on long-term data bases. However, even in these

Table 1 ABSTRACT OF THE FOUR SPEAKER RECOGNITION METHODS REPORTED IN THIS PAPER.

	Statistical or Dynamic Features	Utterances	Recognition Method and Measure
Method 1	Statistical	Words	Weighted distance for statistical features derived from LAR's and fundamental frequency
Method 2	Statistical	Words	Weighted distance for statistical features including fourier coefficients derived from LPC-cepstrum and fundamental frequency
Method 3	Dynamic	Words	DP-matching for time functions of LAR's and fundamental frequency
Method 4	Dynamic	Short sentence	DP-matching for time functions of LPC-cepstrum and polynomial coefficients

LAR: Log Area Ratio

experiments, time intervals between training and test utterances were less than two months.

2 Comparison between Methods 1 and 3

2.1 Speech Analysis

A blockdiagram of the experimental system for comparison between Methods 1 and 3, which includes feature extraction and recognition, is presented in Fig. 1. Two Japanese words, /name/ and /baNgo:/, which mean "name" and "number"

respectively, were uttered in isolation repeatedly over a period of seven years by nine male speakers. They were uttered naturally and there was no attempt to mimic any other speakers. These utterances have been used for speaker recognition experiments.

As shown in Fig. 1, the speech wave is converted into a discrete sequence and 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. A 32 ms Hamming window is applied to the delimited

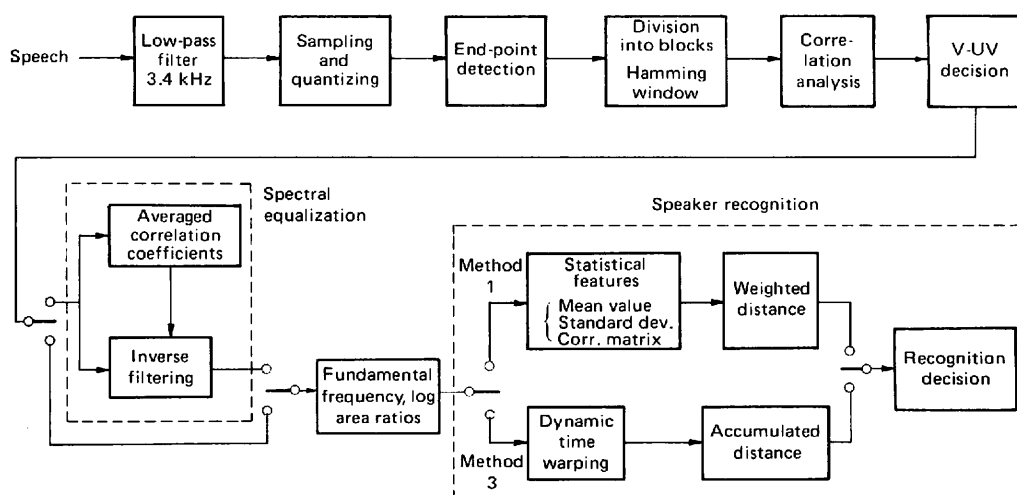


Fig. 1—Feature extraction and recognition blockdiagram for Method 1 and Method 3.

speech every 10 ms. First- to twelfth-order PARCOR coefficients⁽¹⁵⁾ and the fundamental period are extracted from each frame. The fundamental period is extracted from the peak position of the correlation function of the prediction residual, and the voiced-unvoiced decision is made based on the peak value. Since the voiced speech spectrum is much more stable than the unvoiced one, only the voiced portion of each word is used for speaker recognition. The PARCOR coefficients and fundamental period are transformed into log area ratios and fundamental frequency, respectively, to make the frequency distribution for each parameter approach a normal distribution. Log area ratios are defined as arctanh transformation of PARCOR coefficients.

Thus, the speech wave at time t can be represented by a 13-dimensional vector X_t , consisting of fundamental frequency f_{0t} and 12 log area ratios $\{g_{it}\}_{i=1}^{12}$:

$$\begin{aligned} X_t &= (x_{1t}, x_{2t}, \dots, x_{13t}) \\ &= (g_{1t}, g_{2t}, \dots, g_{12t}, f_{0t}) \end{aligned} \quad (1)$$

2.2 Spectral Equalization Process

The author et al.⁽⁸⁾ have investigated the long-term intra-speaker variation in speech spectra, its effect on speaker recognition, and techniques to remove this effect in a speaker recognition system which uses statistical features extracted from spoken words. The results of experiments have made clear that a first- or second-order critical damping inverse filter, which represents the overall pattern of the long-time averaged spectrum, is quite useful for reducing the long-term spectral variability effect on speaker recognition. In a previous experiment by the author et al.⁽⁶⁾, it was found that, although the intra-speaker variation in the general indication or overall pattern for the long-time averaged spectrum, which can be represented by several lower cepstrum coefficients for the long-time averaged spectrum, is very small over a period of a couple of weeks, it becomes more significant than the variation in its fine structure for a longer period.

When inverse filtering (or spectral equalization) is applied to input utterances, high recognition accuracy can be obtained, even if training utterances for each speaker are recorded over a short period and the time difference between training

and input utterances is as long as one year. From the speech production mechanism viewpoint, the spectral equalization effectiveness means that vocal tract characteristics are much more stable than overall patterns for the vocal cord spectrum.

The above mentioned experiment was performed by letting the time difference between training and input utterances range up to two years. This paper presents speaker recognition experiments tried on spectral equalization effectiveness, in which the time difference ranges up to five years. A second-order critical damping inverse filter is used in these experiments. The inverse filter transmission characteristic is

$$G(Z) = 1 + \gamma Z^{-1} + (\gamma^2/4) Z^{-2} \quad (2)$$

where

$$Z = e^{j\omega}, \quad \omega = \text{rad/unit time.}$$

Based on the least squares method, parameter γ is given as the real root which satisfies $|\gamma| < 2$ for the cubic algebraical equation

$$c_0 \gamma^3 - 6c_1 \gamma^2 + (4c_2 + 8c_0) \gamma - 8c_1 = 0, \quad (3)$$

where c_i is the i th order correlation function averaged over all the voiced frames for each input utterance. The actual inverse filtering processing can easily be done by convolution of the correlation function for the input utterance.

2.3 Feature Extraction and Recognition Procedure for Method I

To determine the time function of vector X_t in the voiced portion of each word, mean value μ , standard deviation σ , covariance matrix Σ and correlation matrix A are measured⁽⁷⁾.

$$\mu = \left(\sum_{t=1}^M X_t \right) / M \quad (4)$$

$$\Sigma = (\sigma_{ij}) = \left\{ \sum_{t=1}^M (X_t - \mu)(X_t - \mu)' \right\} / (M-1) \quad (5)$$

$$\sigma = (\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \dots, \sqrt{\sigma_{13,13}}) \quad (6)$$

$$A = (\lambda_{ij}), \quad \lambda_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii} \sigma_{jj}} \quad (7)$$

where M is the number of voiced frames in the

word utterance. The arctanh transformation is also applied to each λ_{ij} to make its frequency distribution approach the normal distribution.

Since X_t is a 13-dimensional vector, vectors μ and σ have 13 elements each, and symmetrical matrix A has 169 elements, of which 78 elements are independent. From this set of 13 elements for vector σ and the 78 elements for matrix A , a fixed reduced set is selected, which is most effective in separating the populations of customer (registered speaker) and impostor sample utterances. Typically, 20 elements are selected, based on earlier experimental results⁽⁷⁾. The criterion for selection is based on the inter-to-intraspeaker variability ratio for each element, calculated over a population of training utterances. All elements for vector μ and the selected set of σ and A elements are used as the elements of a feature vector T for speaker recognition.

The recognition decision is based on a weighted distance between the feature vector of a test utterance and a reference template:

$$D(T, R_r) = |V|^{1/2} (T - R_r)' V^{-1} (T - R_r) \quad (8)$$

where T is the feature vector of the test utterance and R_r is the feature vector of the r th speaker's reference template. R_r is constructed by averaging all the training utterances. V is the covariance matrix of the feature vector, which is calculated from training utterances by each customer and averaged over all customers. Since a preliminary experiment showed that the cross correlation between a set of the elements for feature vector T , which is selected from vector σ and matrix A , is usually very small, a part of the matrix V , which is related to these combinations (these cross-correlation coefficients), is set at zero in order to reduce the number of calculations in Eq. (8).

2.4 Recognition Procedure for Method 3

The time function for vector X_t in the voiced portion is used directly for speaker recognition. An input utterance is brought into time registration with a reference template to calculate the distance between them. This is accomplished by a non-linear time warping method using a dynamic programming technique⁽¹⁰⁾.

The input utterance is indicated as X_t , $1 \leq t \leq M$, and the reference template for speaker r is indicated as $Y_{t'}$, $1 \leq t' \leq N$. The purpose of the time

warping algorithm is to provide a mapping between the two indices t and t' , such that a time registration between the two utterances is obtained. As shown in Fig. 2, the warping function is restricted to a fixed region, the adjustment window, which has width $2 \times N_w$ along the diagonal line connecting points (1,1) and (M, N) on the t - t' plane. The warping function direction is restricted to vertical, horizontal or diagonal.

A complete specification for the warping function results from a point-by-point measure of similarity between input utterance X_t and reference template $Y_{t'}$:

$$D(X_t, Y_{t'}) = |W|^{-1/2} (X_t - Y_{t'})' W (X_t - Y_{t'}) \quad (9)$$

where

$$W = \text{diag}(w_1, w_2, \dots, w_{13})$$

$$w_i = 1/v_{ii}.$$

The weighted distance accounts for the long-term variation in feature vector X_t and is similar to the weighted distance used for the statistical features. Given distance function D , the optimum dynamic path is chosen to minimize accumulated distance D_T along the path. Recognition decision is based on the overall distance accumulated over the optimum warping function.

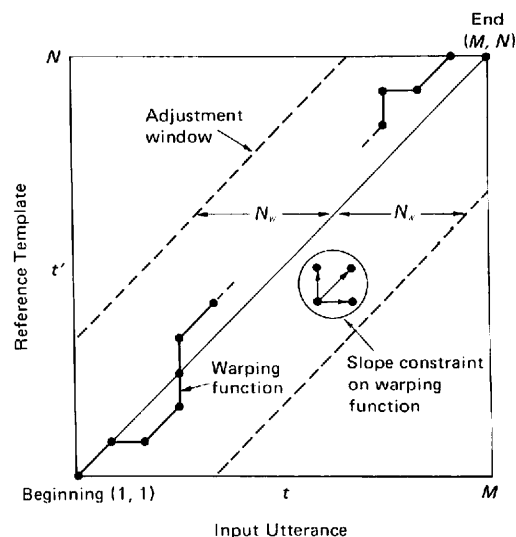


Fig. 2—Slope constraint and adjustment window for the dynamic programming warping function.

2.5 Experimental Setup for Speaker Identification and Verification

Speaker recognition can be divided into speaker verification and speaker identification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker by comparing a set of measurements on the speaker's utterances with a reference set of measurements made on the utterance of the person whose identity is claimed. Speaker identification is the process of determining from which of the registered speakers the given utterance came. In the case of speaker verification, input utterances, whose distances to the reference template are smaller than the threshold, are accepted as utterances by the registered speaker, while input utterances whose distances are larger than the threshold are rejected. In the experiment for comparing Methods 1 and 3, the threshold is set a posteriori for each individual speaker in order to make the two kinds of error rates, false rejection and false acceptance, equal. In the case of speaker identification, the registered speaker, whose reference template is nearest to the input utterance among all the registered speakers, is selected as the speaker of the input utterance.

In this experiments, both speaker verification and speaker identification are performed using nine male registered speakers (customers). There were eight impostors, whose utterances should be rejected, in the verification experiment. The verification and identification decisions are performed using either or both of the two words. In the latter case, the average distance for the two words is used for the decision.

In order to investigate the feature parameters long-term variability effect on the two kinds of speaker recognition systems, two sets of training utterances were prepared for each customer. The short-term training set comprises utterances recorded over a period of ten days in three or four sessions at intervals of two or three days. The long-term training set comprises utterances recorded over a ten month period in four sessions at intervals of three months. The time interval between the last training utterance and input utterance ranged from two or three days to five years. As each speaker utters each word three times at intervals of 1 minute at each session, there are usually 12 training utterances. When training utterances from only three sessions are used, there are nine training utterances. The

latter is the case only when the training is done over a short period and the time interval between training and input utterances is two or three days.

2.6 Experimental Results

2.6.1 Condition I: Long-Term Training and Three Months Interval

Figure 3 shows speaker identification and verification accuracy when the time interval between training and input utterances is three months and the long-term training set is used⁽¹⁷⁾. Spectral equalization was not applied to the speech wave in this experiment. This figure provides a comparison between results obtained by Method 1 and 3. The width of the adjustment window to restrict the warping function was set to 200 ms in Method 3. This figure also provides experimental results when two kinds of features, statistical features and dynamic features, were combined for decision making to study the possibility of improving recognition performance over the performance obtained using either one by itself. The weighted distance associated with the mean vector for the statistical features and the overall distance for the dynamic features were summed to provide a total distance to be used as a decision function. The summation was performed after appropriate weighting of the distances to balance the magnitude of these two values.

The results in Fig. 3 show that statistical features are more effective than dynamic features for speaker identification, whereas there is almost no difference between the results using these two kinds of features for speaker verification. Since the amount of calculation necessary for recognition using statistical features (Method 1) after obtaining the time functions of fundamental frequency and log area ratios and memory size are only about one-tenth those for recognition using dynamic features (Method 3), statistical features are considered to be better than dynamic features.

When the two words are used in combination for decision making, the error rate becomes almost half the average error rate obtained for the single word decision, irrespective of the feature parameter set.

This figure also shows that, both in identification and verification, the recognition error rate can be reduced by half by a combination of the two

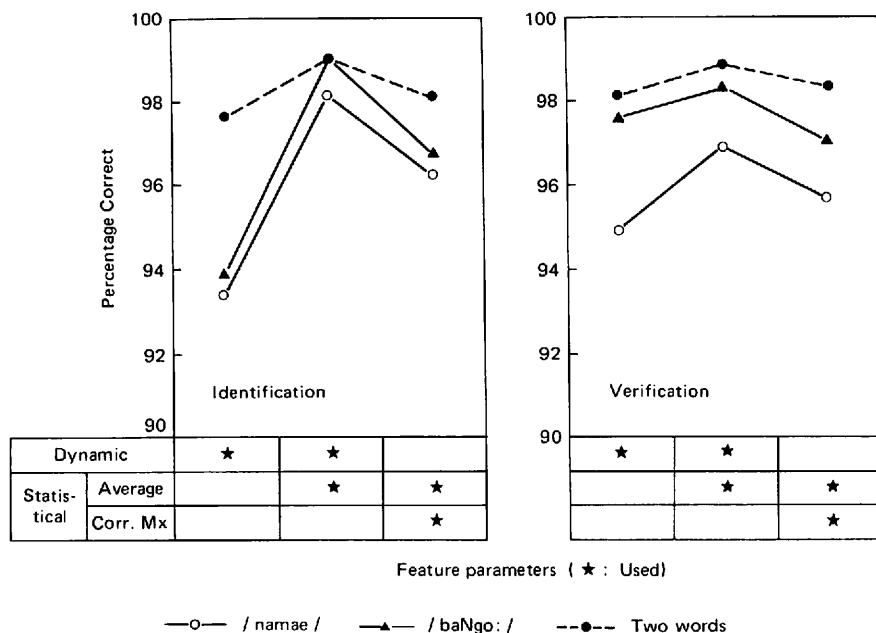


Fig. 3—Comparison between speaker recognition accuracies, using statistical features (Method 1) and dynamic features (Method 3). Effects are shown of combining these features on speaker recognition when long-term training set is used and the time interval is set at three months.

sets of features, irrespective of the number of words for decision making. Although statistical and dynamic features are originally extracted through the same spectral analysis procedure, they can be considered to have some speaker characterizing information independent of each other. The average recognition accuracies for one word are 98.7 percent and 97.6 percent for identification and verification, respectively, and for two words, 99.1 percent and 98.9 percent, respectively.

2.6.2 Condition II: Long-Term or Short-Term Training, Time Interval of up to Ten Months

Figure 4 shows the relation between speaker verification accuracy and the length of time between training and input utterances for four kinds of experimental conditions: the combinations of the two training periods and the two feature parameter sets⁽¹⁸⁾. The time interval between training and input utterances ranges from two or three days to ten months. Recognition accuracy

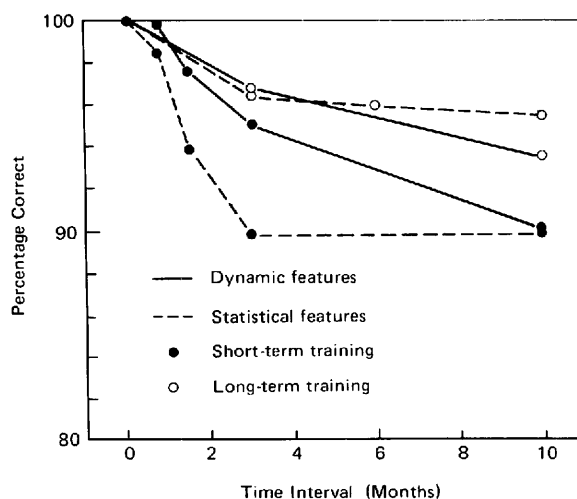


Fig. 4—Speaker verification accuracy as a function of time interval between training and input utterances.

for a single word obtained by averaging the results for each of the two words is shown in this figure. Spectral equalization was not applied to the speech wave in this experiment, either. The width of the adjustment window is also 200 ms in this experiment.

These results show that the verification accuracy decreases as the time interval between training and input utterances becomes long, irrespective of the feature parameter set. Although the accuracy decreases, even if the long-term training set is used, the amount of decrease for long-term training is smaller than that for short term training. When the time interval is three months or so, verification accuracy using statistical features is more affected by the length of the training period than is the accuracy using dynamic features. Long-term training is necessary to get high accuracy for statistical features, when the time interval is three months or so.

However, when the time interval becomes as long as ten months, the accuracy associated with dynamic features is also affected by the length of the training period, and the results are almost the same as those obtained for statistical features. In other words, although the accuracy associated with dynamic features is affected by long-term spectral variability more slowly than the accuracy associated with statistical features, the amount of the influence on the accuracy is almost the same for the two kinds of feature parameter sets over the long term.

2.6.3 Condition III: Long-Term or Short-Term Training, Time Interval of up to Five Years

The effectiveness of the spectral equalization process was examined by means of speaker recognition experiments using statistical features. The results with and without spectral equalization were compared with each other for both short-term training and long-term training.

Figure 5 shows the relation between the time interval as it increases to five years and the recognition accuracy for speaker identification and verification using statistical features based on two-word decisions. These results show that spectral equalization is effective in reducing the errors in recognition as a function of the time interval, both for long-term training and for short-term training. Effectiveness is especially large in the case of

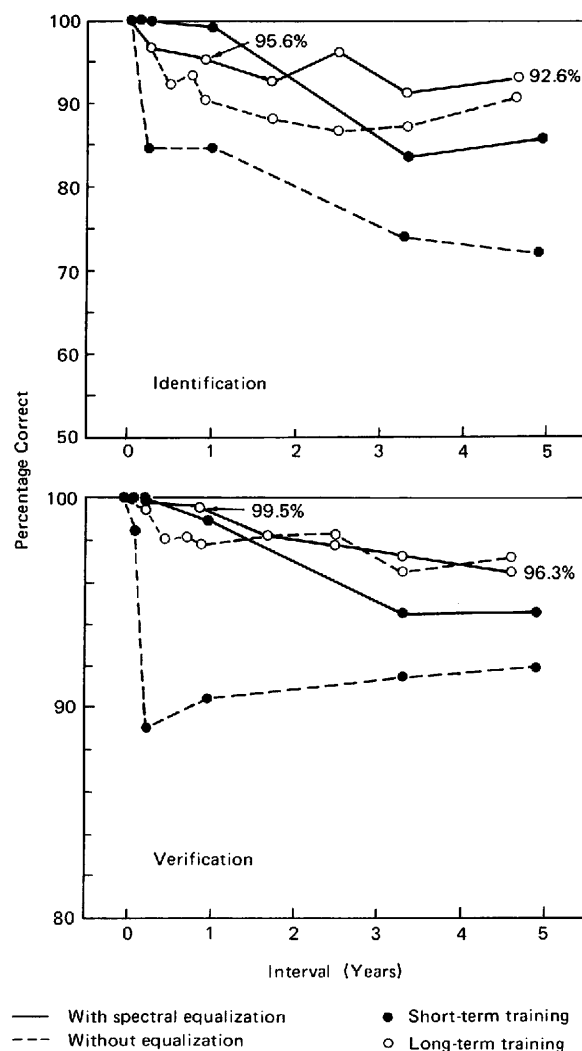


Fig. 5—Recognition accuracy using statistical features as a function of time interval between training and input utterances.

short-term training. For example, when the time interval is less than a year and a half, spectral equalization raises the recognition accuracy for short-term training to almost the same value as that obtained for long-term training, although it produces no improvement in the accuracy for long-term training.

When the time interval is longer than a year and a half, it is desirable to use long-term training and apply spectral equalization. It is impossible to get the same accuracy for short-term training, even if spectral equalization is used. Recognition accuracy for two words with long-term training

and spectral equalization is 99.5 percent and 95.6 percent for verification and identification, respectively, when the time interval is ten months, and 96.3 percent and 92.6 percent, respectively, when the time interval is almost five years.

Although the spectral equalization effectiveness on speaker recognition, using dynamic features, has not been examined in this study, the effectiveness of a similar technique on a speaker verification system was examined in the study on Method 4.

2.7 Experimental Results for One- to Four-Word Decisions in Method 1

The relation between the number of words for the decision on speaker verification or identification and the recognition accuracy in Method 1 was investigated, using the utterances of four kinds of words. These words consist of the two words, which have been used in the previous experiments, and two Japanese words, /ko:geN/ and /bakuoN/, which mean "plateau" and "buzzing", respectively. The number of elements selected from vector σ and matrix A to be used as elements in feature vector T for recognition was also varied between 0 and 20. The training period was long and the time interval was three months. There were nine customers and 111 impostors. All of them were male speakers. The number of test utterances

for all customers was 216.

Figure 6 shows the relation between the total number of elements of the feature vectors and the recognition error rate on logarithm scales. The product of the error rate in percentage and the number of features is approximately constant. When 20 elements are selected ($m = 20$) from vector σ and matrix A for each word and four words are used for the decision, 99.3 percent and 99.1 percent accuracies are obtained for speaker verification and identification, respectively.

3 Speaker Verification by Method 4

3.1 Feature Extraction and Recognition Procedure

As an improvement upon Method 3, which uses dynamic features of speech spectrum, Method 4 has been investigated⁽¹⁹⁾⁻⁽²¹⁾. In this method, not only are the time functions of spectral parameters, which represent the short-time spectral envelope, brought into time registration with stored reference functions, but also a set of new dynamic features are extracted and used for the decision. This method was designed especially for telephone speech and evaluated by speaker verification experiments.

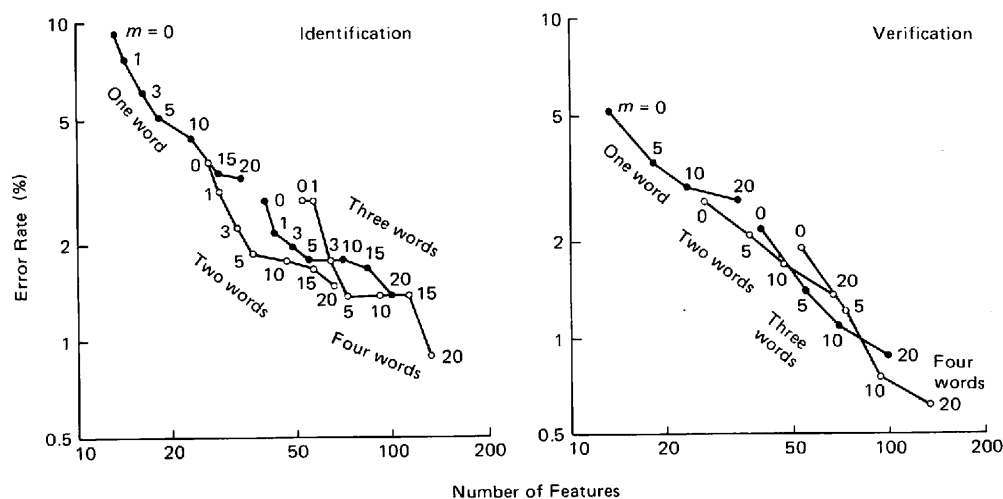


Fig. 6—Relation between total number of features used for speaker recognition and recognition error rate.

3.1.1 Speech Analysis and Feature Extraction

A blockdiagram indicating the principal operations of the verification system is shown in Fig. 7. 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 from a memory. The second input is activated by a request to speak the sample utterance.

The speech wave is bandlimited from 100 Hz to 3.0 kHz and sampled at a 6.67 kHz rate. 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 end-point 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 auto-correlation method and the linear predictor coefficients are transformed into cepstrum coefficients.

The cepstrum coefficients are averaged over the duration of the entire utterance and the averaged values are subtracted from the cepstrum coefficients of every frame to compensate for frequency-response distortions introduced by the transmission system⁽²²⁾ and to reduce long-term intraspeaker spectral variability. This procedure is similar to the spectral equalization investigated and reported in the previous section.

Time functions for the normalized cepstrum coefficients are expanded by an orthogonal poly-

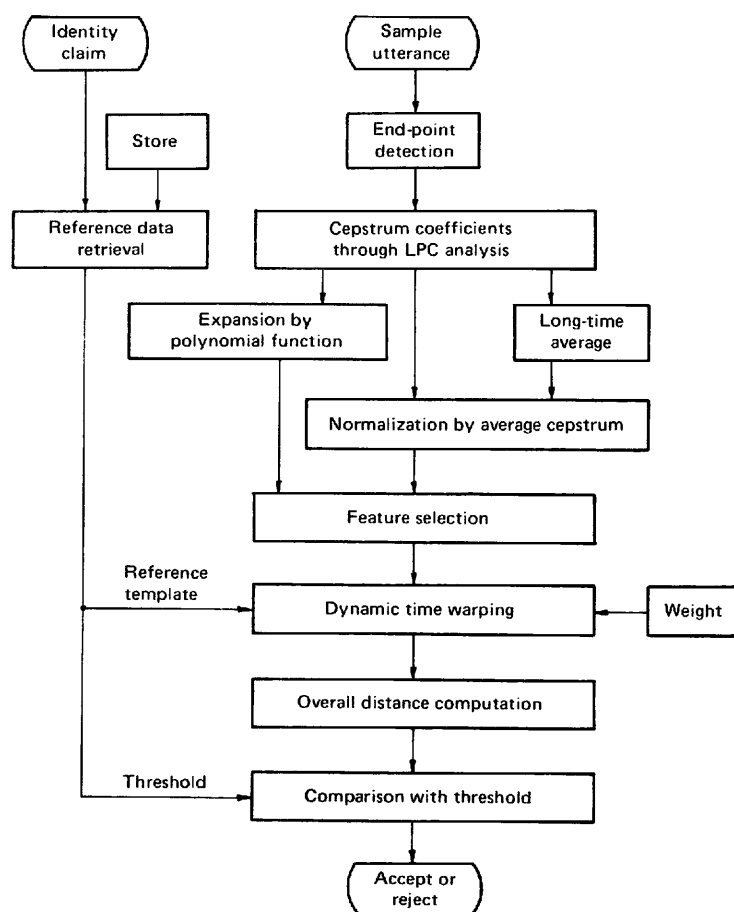


Fig. 7—Blockdiagram indicating the principal operation of Method 4.

nomial representation⁽²³⁾ 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. The first three coefficients of the orthogonal polynomial representation correspond mean value, slope and curvature of the time function for 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.

Accordingly, the utterance is represented by time functions of the cepstrum coefficients and the first- and second-order polynomial coefficients. Since there are ten cepstrum coefficients, the result is a representation by a time function of a 30-dimensional vector. From these 30 elements, a set of elements (typically 18), 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.

3.1.2 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 a 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⁽²⁴⁾ is applied, and the shorter of either the reference template or test utterance is used as the guide contour. 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. Figure 8 shows the warping plane and the allowable region of the dynamic warping function, where $m_0 = 20$ and $\delta = 15$ (frames). The warping function is required to monotonically increase with a maximum slope of two and a minimum slope of $1/2$ ⁽²⁵⁾.

The distance for each element is weighted by

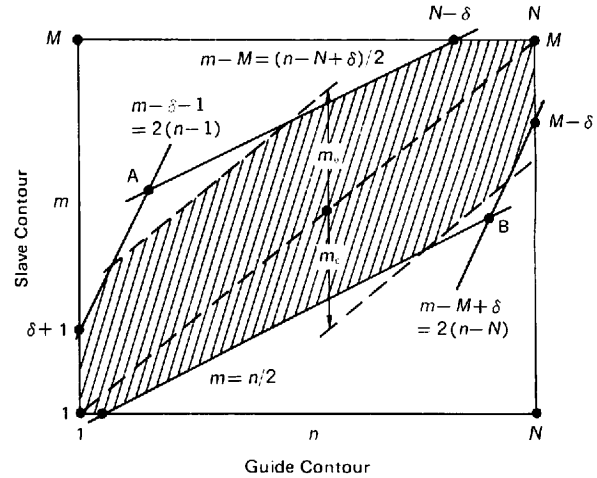


Fig. 8—Allowable region of the dynamic programming warping function path.

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.

3.1.3 Decision Threshold and Reference Templates

In this experiment, 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 by 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:

$$\theta(k) = a(\bar{\mu}_{DB}(k) - \bar{\sigma}_{DB}(k)) + b \quad (10)$$

where $\theta(k)$ is the threshold for the customer k , $\bar{\mu}_{DB}(k)$ and $\bar{\sigma}_{DB}(k)$ are mean value and standard deviation for the distribution of interspeaker distance, 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 threshold over a set of training data. Figure 9 shows an example of typical intraspeaker and interspeaker distance distribution.

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

The establishment and updating of reference information is another important system element. 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, the inter-to-intraspeaker variability ratio, which is used in the feature selection, and the interspeaker distance distribution, which is used to set the decision threshold.

The initial reference template and threshold 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 section, reference templates were updated every seventh access of the system by each customer using the latest five utterances, as shown in Fig. 10. 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.

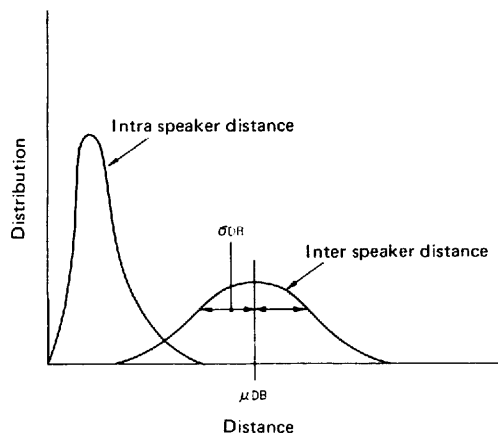


Fig. 9—Example of typical intraspeaker and interspeaker distance distributions.

3.2 Speaker Verification Experiments

3.2.1 Utterance Sets and Preliminary Experiments

Three kinds of utterance sets by native Americans were used to evaluate this system. The first and second sets each comprises 50 utterances by each of 10 speakers designated as customers and a single utterance by each of 40 speakers designated as 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 as customers and a single utterance

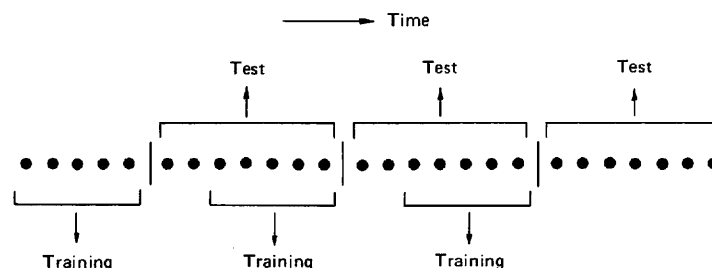


Fig. 10—Reference updating procedure. Reference template is updated every seventh access by the customer, using the latest five utterances.

from each of 55 male speakers designated as 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 customer recordings were made over periods of two months, 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. 11. Shown plotted are the ratios of the average values of interspeaker distances to the average values of the intraspeaker distances for each of the 30 original parameters. (Note once again that the original cepstrum coefficients are used in place of the 0th order polynomial coefficients.) With almost total consistency, the higher the polynomial coefficient order is, the smaller the distance ratios are and, hence, the less effective the parameters are. Based on these results, 18 parameters were selected as features for the speaker verification tests. These included all the original cepstrum coefficients and none of the second order coefficients. Similar results were obtained for the other two utterance sets.

3.2.2 Speaker Verification Experiments Results

The speaker verification experiments results for the three utterance sets are summarized in Table 2. The a priori error rates shown are the results of averaging the false acceptance 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 (10). 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 estimate 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 utterance sets.

In order to check the effect of the time interval between training and test utterances on speaker verification accuracy, the 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

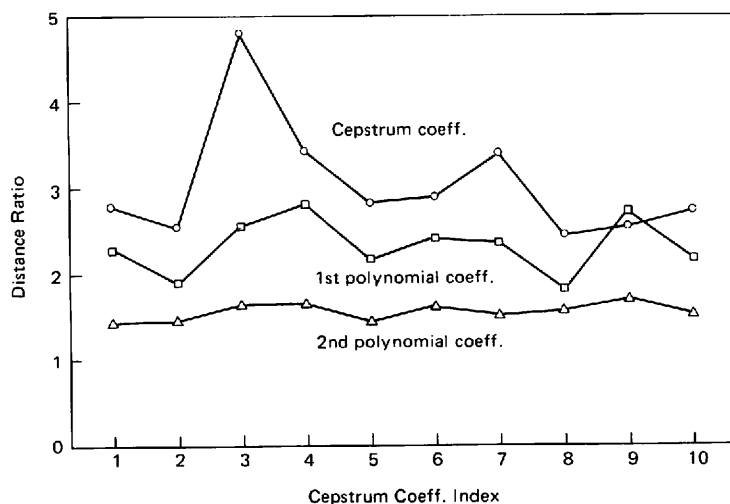


Fig. 11—Interspeaker to intraspeaker distance ratio for the first ten utterances by five speakers, each extracted from utterance set (1).

Table 2 AVERAGE VERIFICATION ERROR RATES FOR THREE UTTERANCE SETS UNDER THREE DECISION THRESHOLD CONDITIONS.

Utterance Set		1	2	3
Customers and Impostors		10 male and 40 male	10 female and 40 female	21 male and 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%

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. It is a future problem to test this system on the condition that the time interval is longer than six weeks.

Some other experiments were also performed to evaluate this system. Results of an experiment indicate that a verification error rate of one percent or less can be obtained, even if the reference and test utterances are subjected to different transmission conditions⁽²¹⁾. An on-line experiment using telephone speech uttered by 60 male speakers and 60 female speakers, which was performed over a period of six months, also indicates the effectiveness of this system. Details of this on-line experiment will be presented in other papers.

4 Speaker Recognition by Method 2

4.1 Feature Extraction and Recognition Procedure

As an improvement upon Method 1, which uses statistical features of speech spectrum, Method 2 has been investigated. A blockdiagram indicating the principal operations of the system is shown in Fig. 12. The difference between Methods 1 and 2 lies in the spectral analysis and the statistical parameters.

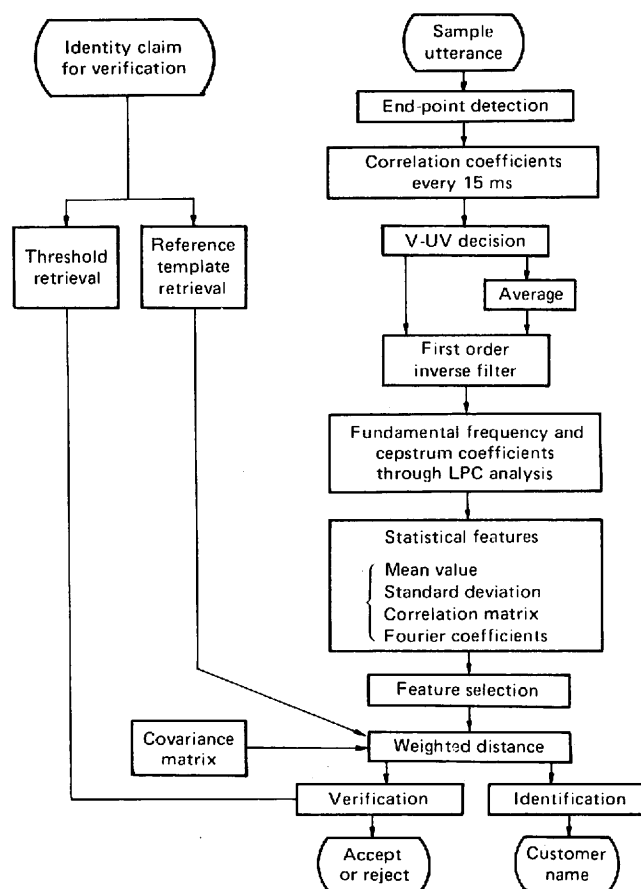


Fig. 12—Blockdiagram indicating the principal operation of Method 2.

In Method 2, the speech wave is sampled at 6.67 kHz rate, and a 30 ms Hamming window is applied to the digitized speech every 10 ms. Fundamental frequency and first- to tenth-order cepstrum coefficients are extracted from each frame through LPC analysis technique after the spectral equalization by the 1st order inverse filter. For the time function of these parameters in the voiced portion of each word, a set of statistical features are measured. The statistical features consist of the features used in Method 1, which are vector μ , vector σ and matrix A , and Fourier cosine expansion coefficients;

$$F_{it} = \frac{1}{N} \sum_{l=1}^N x_{il} \cos \frac{\pi(t-1)l}{N} \quad (11)$$

where x_{it} is the i th spectral parameter at time t ($1 \leq t \leq N$), which is a cepstrum coefficient or fundamental frequency. N is the number of all the voiced frames. Since 0th order Fourier coefficients are equal to the elements of the vector μ , first- to sixth-order Fourier coefficients are calculated ($1 \leq l \leq 6$).

Similar to Method 1, a fixed reduced set, which is most effective in separating the population of customer and impostor sample utterances, is selected based on the inter-to-intraspeaker variability ratio for each element calculated over a population of training utterances. In this method, 60 elements were selected from the elements of vector σ , matrix A and Fourier coefficients, and used as the elements in a feature vector for speaker recognition, together with all elements for vector μ .

Both a priori threshold and a posteriori optimum threshold were used for the speaker verification decision. In the former case, the same threshold is used for all customers.

4.2 Speaker Recognition Experiments Results

A speaker recognition experiment was performed under the same conditions as listed in Section 2.7. Training utterances were collected over a ten months interval, and the time interval between training and input utterances was three months. There were 111 impostors and four words were used for the decision on identification and verification. A supplementary experiment was performed for the case where log area ratios were used for the spectral analysis instead of LPC cepstrum

Table 3 AVERAGE RECOGNITION ERROR RATES WHEN THE LONG-TERM TRAINING SET (12 UTTERANCES IN FOUR SESSIONS) IS USED AND THE TIME INTERVAL IS SET AT THREE MONTHS.

LPC-CEP.	Identification		0%
	Verification	A priori fixed threshold	0.10%
		A posteriori threshold	0.02%
LAR	Identification		0.46%
	Verification	A priori fixed threshold	1.21%
		A posteriori threshold	0.51%

coefficients. Average error rate for each condition is shown in Table 3. These results indicate that LPC cepstrum coefficients are much more efficient than log area ratios, and more than 99.9 percent recognition accuracy can be obtained by Method 2, which uses statistical features including Fourier coefficients extracted from the time functions of LPC cepstrum coefficients and fundamental frequency.

4.3 Recognition Experiments for Telephone Speech

Method 2 was evaluated by telephone speech recorded through a carbon transmitter and 0.4 mm ϕ - 7 dB artificial subscriber line. Customers and impostors were nine male speakers and 11 male speakers, respectively. A set of training utterances was collected in three sessions at intervals of six months for each customer. Nine training utterances were used to construct each reference template. The time interval between training and input utterances was six months. Each impostor uttered each word three times. Fifty four input utterances by all the customers were used in the experiment. A supplementary experiment was performed using speech samples recorded through a high quality microphone at the same time as the telephone speech, in order to compare these results.

The results of experiments are shown in Table 4. Error rates in Table 4 are relatively higher than those in Table 3, since the experimental condition

Table 4 COMPARISON OF AVERAGE RECOGNITION ERROR RATES FOR TELEPHONE SPEECH AND MICROPHONE SPEECH, WHEN THE LONG-TERM TRAINING SET (9 UTTERANCES IN THREE SESSIONS) IS USED AND THE TIME INTERVAL IS SET AT SIX MONTHS.

TEL.	Identification		0%
	Verification	A priori fixed threshold	2.44%
		A posteriori threshold	0.49%
MIC.	Identification		0%
	Verification	A priori fixed threshold	2.44%
		A posteriori threshold	1.36%

associated with the former table is severer than for the latter one. Results in Table 4 indicate that Method 2 provides almost the same performance for telephone speech and microphone speech.

Method 2 was also evaluated by an on-line experiment, which is similar to that performed for Method 4. The on-line experiment also ascertained Method 2 effectiveness. Details of this experiment will be reported in other papers.

5 Conclusion

Four kinds of text-dependent speaker recognition methods, which use statistical or dynamic features of speech spectra, were evaluated. First, performance of two basic methods, Method 1 and Method 3, which use statistical features and dynamic features, respectively, were compared with each other from the viewpoint of long-term variability in these features. In this experiment, spectral equalization was evaluated as a technique for reducing the long-term variability effects.

Experiment results show that, although Methods 1 and 3 provide almost the same performance, the former method is better than the latter one, since the amount of calculation and memory size necessary for the former method is only about one-tenth that for the latter one. The experiment by Method 1, using statistical features, ascertained the spectral equalization procedure effectiveness in reducing the amount of decrease in recognition accuracy with increasing time between training

and input utterances. Although this procedure is useful in both short-term training and long-term training, the effectiveness of this procedure is especially evident in the case of short-term training. A high recognition accuracy can be obtained by Method 1, even for input utterances spoken five years later, when the spectral equalization is applied.

It was also clarified that the recognition error rate can be reduced by half by a combination of the statistical and dynamic features, and that it can also be reduced by a combination of more than one word for decision making, both in Methods 1 and 3.

Next, Methods 2 and 4, which are improved upon Methods 1 and 3, respectively, were evaluated. The results of experiments indicate that these methods can produce much higher recognition accuracy than the two basic methods, and both of them are effective for telephone speech. Although these two methods have not been evaluated by common speech utterance sets, it seems that they provide almost the same performance. Since Method 2 needs only about one-tenth the amount of calculation and memory size, Method 2 is better than Method 4. There remain some additional comparisons between these two methods which should be examined by speech samples uttered by many speakers over a long period and by speech utterances transmitted over a long distance telephone network and degraded by noise.

Acknowledgment

The author especially wishes to acknowledge the guidance provided by Dr. N. Kuroyanagi, Director of Research Division, Dr. S. Hashimoto, Deputy Director of Research Division, and Dr. F. Itakura, Chief of the Fourth Research Section. The author also wishes to thank Dr. J. L. Flanagan and Dr. A. E. Rosenberg at Bell Laboratories for their guidance and stimulating discussions for the research reported in Section 3.

References

- (1) B. S. Atal: Automatic Recognition of Speakers from their Voices, *Proc. IEEE*, **64**, pp. 460–475, 1976.
- (2) A. E. Rosenberg: Automatic Speaker Verification: A Review, *Proc. IEEE*, **64**, pp. 475–487, 1976.

- (3) S. K. Das and W. S. Mohn: A Scheme for Speech Processing in Automatic Speaker Verification, *IEEE Trans Audio Electroacoust.*, **AU-19**, pp. 32–43, 1971.
- (4) W. A. Hargreaves and J. A. Starkweather: Recognition of Speaker Identity, *Language and Speech*, **6**, pp. 63–67, 1963.
- (5) J. E. Luck: Automatic Speaker Verification Using Cepstral Measurements, *J. Acoust. Soc. Amer.*, **46**, pp. 1026–1031, 1969.
- (6) S. Furui, F. Itakura and S. Saito: Talker Recognition by Longtime Averaged Speech Spectrum, *Electro. Commun. Jpn.* (Scripta Publishing Co., U.S.A.), **55A**, pp. 54–61, 1972.
- (7) S. Furui and F. Itakura: Talker Recognition by Statistical Features of Speech Sounds, *Electron. Commun. Jpn.*, **56A**, pp. 62–71, 1973.
- (8) S. Furui: Analysis of Long-Term Variation of Feature Parameters of Speech and Its Application to Talker Recognition, *Electron. Commun. Jpn.*, **57A**, pp. 34–42, 1974.
- (9) S. Saito and S. Furui: Personal Information in Dynamic Characteristics of Speech Spectra, *Proc. 4th Int. Joint Conf. Pattern Recognition*, pp. 1014–1018, 1978.
- (10) M. R. Sambur: Speaker Recognition Using Orthogonal Linear Prediction, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-24**, pp. 283–289, 1976.
- (11) J. D. Markel, B. T. Oshika and A. H. Gray, Jr.: Long-Term Feature Averaging for Speaker Recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-25**, pp. 330–337, 1977.
- (12) R. E. Wohlford, E. H. Wrench, Jr. and B. P. Landell: A Comparison of Four Techniques for Automatic Speaker Recognition, *Conf. Rec. 1980 Int. Conf. Acoust., Speech, Signal Processing*, **3**, Denver, CO, 1980, pp. 908–911.
- (13) G. R. Doddington: Speaker Verification, *RADC-TR-U1-963700-F*, 1974.
- (14) A. E. Rosenberg: Evaluation of an Automatic Speaker Verification System over Telephone Lines, *B.S.T.J.*, **55**, pp. 723–744, 1976.
- (15) F. Itakura and S. Saito: On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer, *Conf. Rec., IEEE 1972 Conf. Speech Commun. and Process*, New York, 1972, paper L4, pp. 434–437.
- (16) H. Sakoe and S. Chiba: A Dynamic Programming Approach to Continuous Speech Recognition, 1971 Proc. 7th Int. Cong. Acoustics, paper 20-C-13, 1971.
- (17) S. Furui: Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-29**, pp. 342–350, 1981.
- (18) S. Furui: Effects of Long-Term Spectral Variability on Speaker Recognition, *J. Acoust. Soc. Amer.*, **64**, suppl. 1, paper NNN28, p. S183, 1978.
- (19) S. Furui: New Techniques for Automatic Speaker Verification Using Telephone Speech, *J. Acoust. Soc. Amer.*, **66**, suppl. 1, paper R4, p. S35, 1979.
- (20) S. Furui and A. E. Rosenberg: Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech, *Conf. Rec. 1980 Int. Conf. Acoust., Speech, Signal Processing*, **3**, Denver, CO, 1980 pp. 1060–1062.
- (21) S. Furui: Cepstral Analysis Technique for Automatic Speaker Verification, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-29**, pp. 254–272, 1981.
- (22) B. S. Atal: Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, *J. Acoust. Soc. Amer.*, **55**, pp. 1304–1312, 1974.
- (23) N. L. Johnson and F. C. Leone: Statistics and Experimental Design in Engineering and the Applied Sciences, 2nd ed., **1**, New York: Wiley, 1977, pp. 471–481.
- (24) 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*, **ASSP-26**, pp. 575–582, 1978.
- (25) F. Itakura: Minimum Prediction Residual Principle Applied to Speech Recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-23**, pp. 67–72, 1975.