Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features

SADAOKI FURUI, MEMBER, IEEE

Abstract—This paper describes results of speaker recognition experiments using statistical features and dynamic features of speech spectra extracted from fixed Japanese word utterances. The speech wave is transformed into a set of time functions of log area ratios and a fundamental frequency. In the case of statistical features, a mean value and a standard deviation for each time function and a correlation matrix between these functions are calculated in the voiced portion of each word, and after a feature selection procedure, they are compared with reference features. In the case of dynamic features, the time functions are brought into time registration with reference functions.

The results of the experiments show that there is only a slight difference between the recognition accuracies for statistical features and dynamic features over the long term. Since the amount of calculation necessary for recognition using statistical features is only about one-tenth of that for recognition using dynamic features, it is more efficient to use statistical features than dynamic features. When training utterances are recorded over ten months for each customer and spectral equalization is applied, 99.5 percent and 96.3 percent verification accuracies can be obtained for input utterances ten months and five years later, respectively, using statistical features extracted from two words. Combination of dynamic features with statistical features can reduce the error rate to half that obtained with either one alone.

I. 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 intersession 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 speaker's 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 of 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 spectral equalization as a method to reduce this effect is examined using a data base which consists of speech utterances recorded over seven years. Combination of statistical and dynamic features was investigated to get high speaker recognition accuracy. Some discussion of and experiments concerning speaker recognition using dynamic features are also included.

Manuscript received April 29, 1980; revised October 21, 1980.

The author is with the Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation, Musashino-shi, Tokyo 180, Japan.

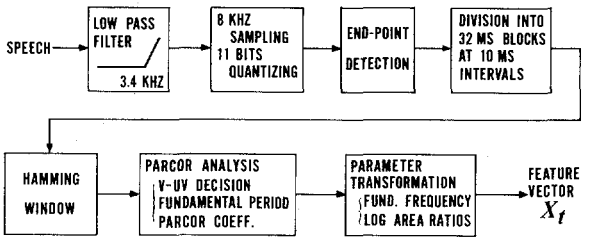

Fig. 1. Block diagram of feature extraction.

Sambur [15] and Markel *et al.* [16] tried speaker recognition systems using statistical features, and Wohlford *et al.* [17] have compared various statistical feature speaker recognition systems, but they have used speech utterances recorded over short periods. Doddington [18] and Rosenberg [19] tried speaker recognition systems using dynamic features based on long-term data bases, but even in these experiments, time intervals between training and test utterances were less than two months.

II. Feature Extraction and Speaker Recognition Method

A. Feature Extraction

Two Japanese words, /namae/ 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 speech every 10 ms. First- to twelfth-order PARCOR coefficients [10] 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

0096-3518/81/0600-0342$00.75

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 of 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 $\boldsymbol{X}_t$, consisting of the fundamental frequency f_{0t} and the 12 log area ratios $\{g_{it}\}_{i=1}^{12}$:

$$\boldsymbol{X}_t = (x_{1t}, x_{2t}, \cdots, x_{13t}) = (g_{1t}, g_{2t}, \cdots, g_{12t}, f_{0t}) \quad (1)$$

B. Recognition Model Using Statistical Features

For the time function of the vector $\boldsymbol{X}_t$ in the voiced portion of each word, mean value $\boldsymbol{\mu}$, standard deviation $\boldsymbol{\sigma}$, covariance matrix $\boldsymbol{\Sigma}$ and correlation matrix $\boldsymbol{\Lambda}$ are measured [7].

$$\boldsymbol{\mu} = \left(\sum_{t=1}^{M} \boldsymbol{X}_t \right) \Big/ M \quad (2)$$

$$\boldsymbol{\Sigma} = (\sigma_{ij}) = \left\{ \sum_{t=1}^{M} (\boldsymbol{X}_t - \boldsymbol{\mu})' (\boldsymbol{X}_t - \boldsymbol{\mu}) \right\} (M-1) \quad (3)$$

$$\boldsymbol{\sigma} = (\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \cdots, \sqrt{\sigma_{13,13}}) \quad (4)$$

$$\boldsymbol{\Lambda} = (\lambda_{ij}), \lambda_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii}\sigma_{jj}} \quad (5)$$

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 $\boldsymbol{X}_t$ is a 13-dimensional vector, the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ have 13 elements each, and the symmetrical matrix $\boldsymbol{\Lambda}$ has 169 elements, of which 78 elements are independent. From this set of the 13 elements of the vector $\boldsymbol{\sigma}$ and the 78 elements of the matrix $\boldsymbol{\Lambda}$, a fixed reduced set which is most effective in separating the populations of customer (registered speaker) and imposter sample utterances is selected. Typically, 20 elements are selected, based on the 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 of the vector $\boldsymbol{\mu}$ and the selected set of $\boldsymbol{\sigma}$ and $\boldsymbol{\Lambda}$ elements are used as the elements of a feature vector $\boldsymbol{T}$ for speaker recognition.

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

$$D(\boldsymbol{T}, \boldsymbol{R}_r) = |V|^{1/N} (\boldsymbol{T} - \boldsymbol{R}_r) \boldsymbol{V}^{-1} (\boldsymbol{T} - \boldsymbol{R}_r)' \quad (6)$$

where $\boldsymbol{T}$ is the feature vector of the test utterance and $\boldsymbol{R}_r$ is the feature vector of the rth speaker's reference template. $\boldsymbol{R}_r$ is constructed by averaging all the training utterances. $\boldsymbol{V}$ is the covariance matrix of the feature vector, which is calculated from training utterance of each customer and averaged over all customers. Since a preliminary experiment showed that the cross correlation between a set of the elements of the feature vector $\boldsymbol{T}$, which is selected from the vector $\boldsymbol{\sigma}$ and the matrix $\boldsymbol{\Lambda}$, is usually very small, a part of the matrix $\boldsymbol{V}$ which is related to these combinations (these cross-correlation coefficients) is set at zero in order to reduce the number of calculations in (6).

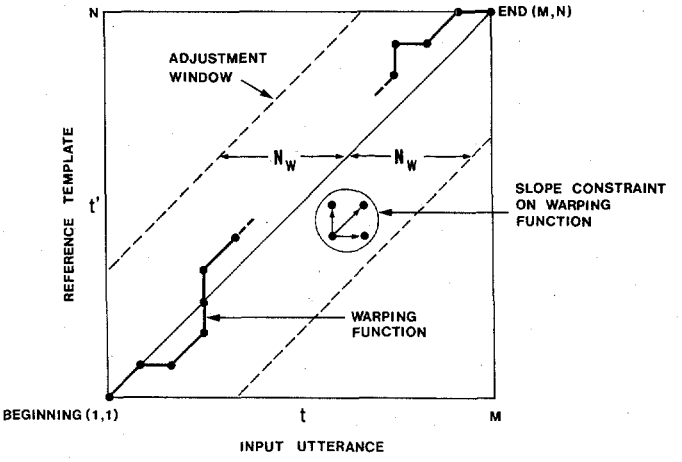

Fig. 2. Warping function constraints for the dynamic time warping procedure.

C. Recognition Model Using Dynamic Features

The time function of the vector $\boldsymbol{X}_t$ 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 nonlinear time warping method using a dynamic programming technique [11].

We donate the input utterance as $\boldsymbol{X}_t$, $1 \leqq t \leqq M$, and the reference template of speaker r as $\boldsymbol{Y}_{t'r}$, $1 \leqq t' \leqq N$. 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 direction of the warping function is restricted to vertical, horizontal or diagonal.

A complete specification of the warping function results from a point-by-point measure of similarity between the input utterance $\boldsymbol{X}_t$ and the reference template $\boldsymbol{Y}_{t'r}$:

$$D(\boldsymbol{X}_t, \boldsymbol{Y}_{t'r}) = |\boldsymbol{W}^{-1}|^{1/N} (\boldsymbol{X}_t - \boldsymbol{Y}_{t'r}) \boldsymbol{W} (\boldsymbol{X}_t - \boldsymbol{Y}_{t'r})' \quad (7)$$

where

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

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

The weighted distance accounts for the long-term variation of feature vector $\boldsymbol{X}_t$ and is similar to the weighted distance used for the statistical features. Given the distance function D, the optimum dynamic path is chosen to minimize the accumulated distance D_T along the path. Recognition decision is based on the overall distance accumulated over the optimum warping function.

D. Spectral Equalization Process

The author *et al.* [8] have investigated the long-term intraspeaker variation of 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 effect of long-term spectral variability on speaker recognition. In the previous experiment by the author *et al.* [6], it was found that, although the intra-speaker variation of the general indication or overall pattern of the long-time averaged spectrum, which can be represented by several lower cepstrum coefficients of the averaged spectrum, is very small over a period of a couple of weeks, it becomes more significant than the variation of 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 the training utterances for each speaker are recorded over a short period and the time difference between training and input utterance is as long as one year. From the viewpoint of the speech production mechanism, the effectiveness of spectral equalization means that vocal tract characteristics are much more stable than overall patterns of the vocal cord spectrum.

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

$$G(Z) = 1 + \gamma Z^{-1} + (\gamma^2/4)Z^{-2} \tag{8}$$

where

$Z = e^{j\omega}$, ω = rad/unit time.

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

$$c_0\gamma^3 - 6c_1\gamma^2 + (4c_2 + 8c_0)\gamma - 8c_1 = 0, \tag{9}$$

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

E. 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 of the speaker's utterances with a reference set of measurements of 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 this paper, 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.

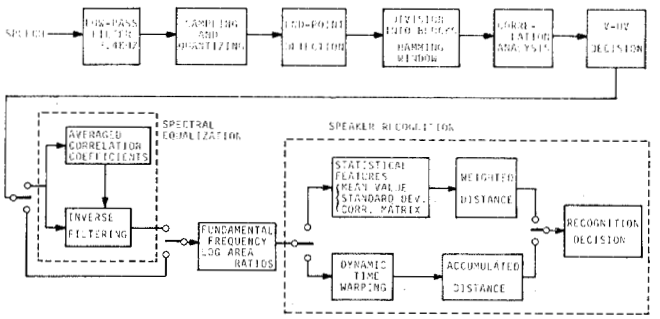

Fig. 3. Block diagram of the total system including feature extraction and recognition.

In this paper, both speaker verification and speaker identification experiments are performed using nine male registered speakers (customers). 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. A block diagram of the total system, which includes feature extraction and recognition, is presented in Fig. 3.

In order to investigate the effect of long-term variability of feature parameters on the two kinds of speaker recognition systems, speaker recognition by statistical features and speaker recognition by dynamic features, two sets of training utterances are 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 min at each session, the number of training utterances is usually 12; when training utterances from only three sessions are used, the number of training utterances is nine. 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.

III. Experimental Results

A. Comparison Between Results Obtained Using Statistical Features and Dynamic Features

Fig. 4 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. This figure provides a comparison between results obtained using statistical features and results obtained using dynamic features. The width of the adjustment window to restrict the warping function was set to 200 ms in the recognition using dynamic features.

These results 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

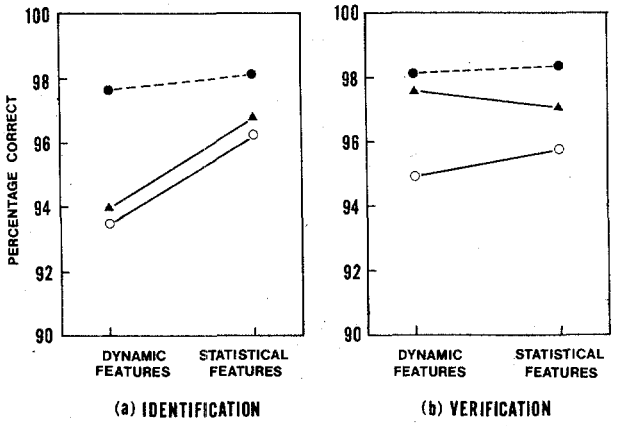

Fig. 4. Comparison between speaker recognition accuracies using statistical features and dynamic features when the long-term training set is used and the time interval between training and input utterances is three months.

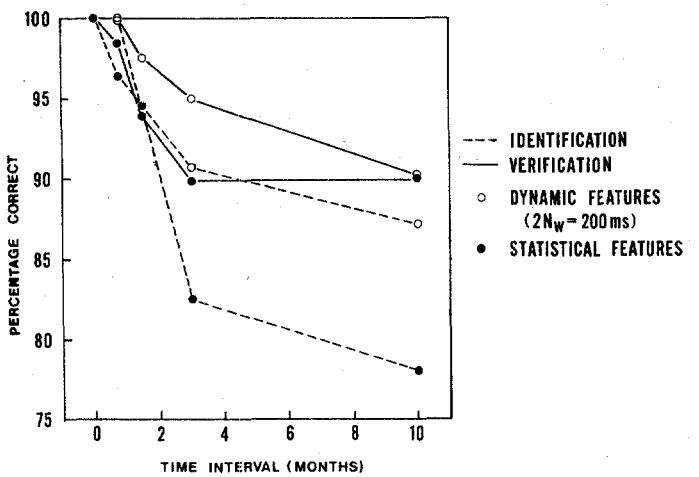

Fig. 5. Recognition accuracy as a function of time interval between training and input utterances for short-term training.

of features for speaker verification. 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.

Fig. 5 shows recognition results obtained using both statistical features and dynamic features when the short-term training set is used and the time interval between training and input utterances ranges from two or three days to ten months. This figure shows recognition accuracy for a single word obtained by averaging the results for each of the two words. The width of the adjustment window is also 200 ms in this experiment.

These results show that both the identification and verification accuracies decrease as the time interval between training and input utterances become long, irrespective of the feature parameter set. Comparison of the results in Fig. 5 for the time interval of three months with the results shown in Fig. 4 obtained by long-term training shows that recognition accuracy associated with statistical features is much more influenced by the length of the training period than accuracy associated with dynamic features. When statistical features are used, the error rate for the time interval of three months with short-term training (Fig. 5) is five times larger than the error rate with long-term

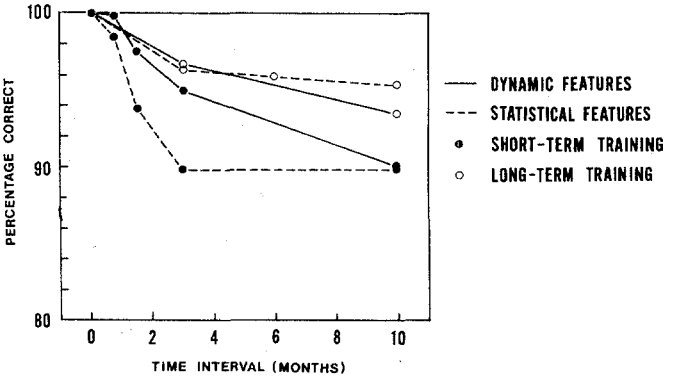

Fig. 6. Verification accuracy as a function of time interval between training and input utterances.

training (Fig. 4) for speaker identification, and three times larger for speaker verification. When dynamic features are used, the difference between error rates for short-term training and long-term training is small. Therefore, recognition accuracy associated with dynamic features is much greater than accuracy associated with statistical features when the short-term training set is used and the time interval between training and input utterance is three months.

Fig. 6 shows the relation between speaker verification accuracy and the length of time between training and input utterances for the four kinds of experimental conditions: the combinations of the two training periods and the two feature parameter sets [12]. Although the accuracy decreases as the time interval becomes long 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. Also, long-term training is necessary to get high accuracy for statistical features, which is the same result as that observed in the comparison between the results of Figs. 4 and 5.

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.

Fig. 7 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. Results for the two kinds of training sets are compared with each other. The decision was made using one word and using two words. Fig. 8 shows the results associated with dynamic features for the same experimental conditions.

Generally speaking, the most reliable performance over the long range can be obtained by statistical features based on long-term training and decision using two words in combination both for speaker identification and verification. This method produces 97.0 percent and 90.2 percent recognition accuracies

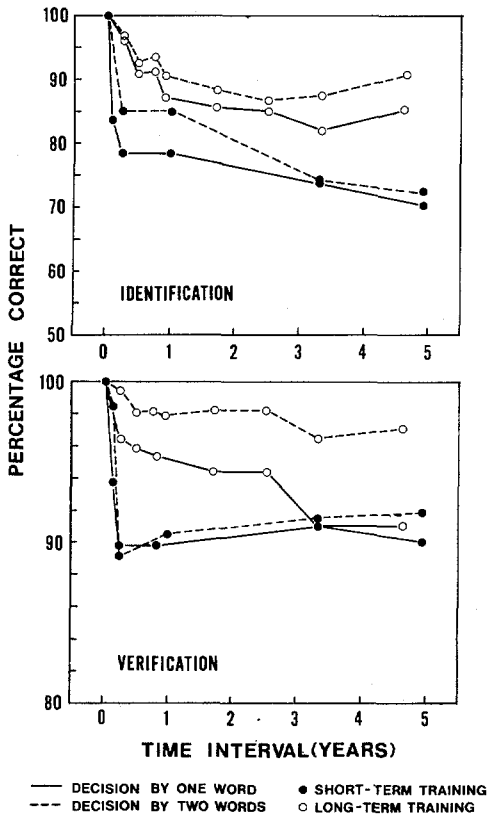

Fig. 7. Recognition accuracy using statistical features as a function of time interval between training and input utterances.

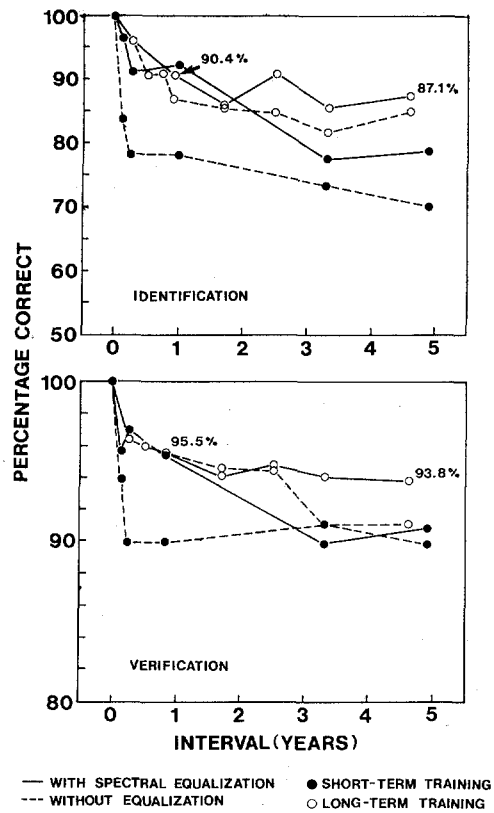

Fig. 9. Effectiveness of the spectral equalization procedure on speaker recognition using statistical features for one-word decisions.

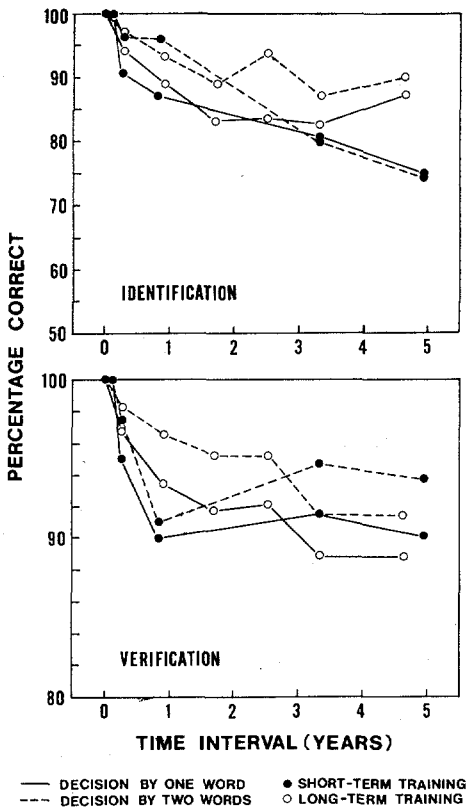

Fig. 8. Recognition accuracy using dynamic features as a function of time interval between training and input utterances.

for verification and identification, respectively, even when the time interval is almost five years.

B. Effectiveness of Spectral Equalization

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. The recognition accuracy based on one-word decisions and two-word decisions are shown in Figs. 9 and 10, respectively.

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 short-term training. The effectiveness is especially large in the case of 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. These results are observed for recognition both for one word and two words. Recognition accuracy for one word with long-term training and spectral equalization is 95.5 percent and 90.4 percent for verification and identification, respectively, when the time interval

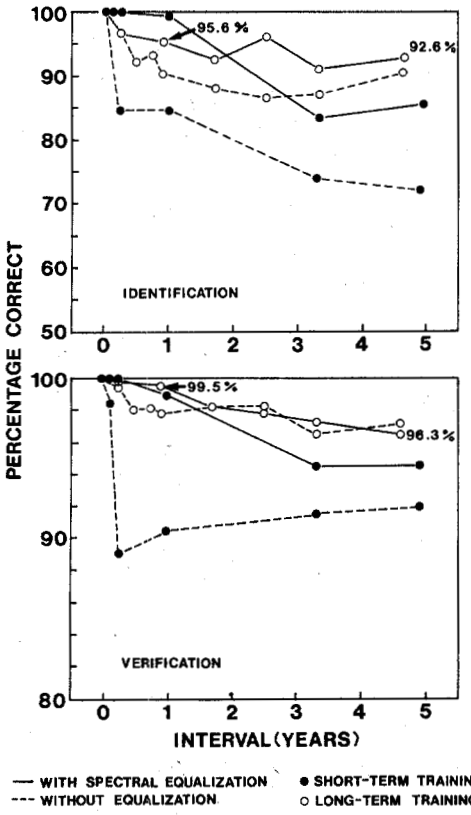

Fig. 10. Effectiveness of the spectral equalization procedure on speaker recognition using statistical features for two-word decisions.

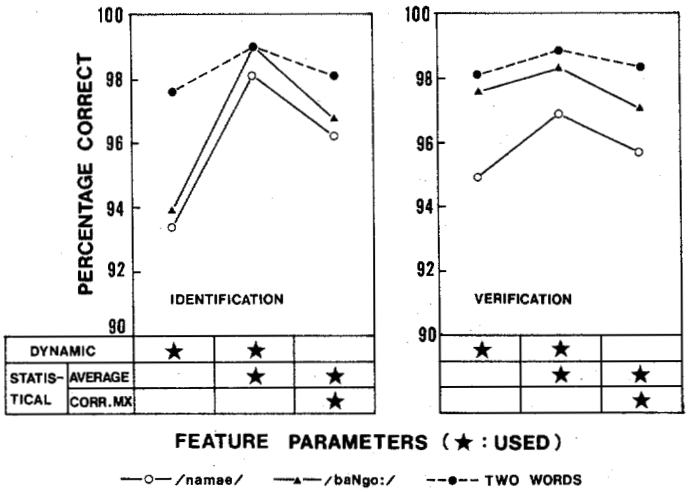

Fig. 11. Effect of combining statistical and dynamic features on speaker recognition when the long-term training set is used and the time interval is set at three months.

is ten months, and 93.8 percent and 87.1 percent, respectively, when the time interval is almost five years. Recognition accuracy for two words with the same conditions 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 effectiveness of spectral equalization on speaker recognition using dynamic features has not been examined in this study, the author [14] has examined the effectiveness of a similar technique on a speaker verification system for telephone speech using cepstrum coefficients and polynomial coefficients at Bell Laboratories.

C. Comparison Between Computation Time for Two Feature Parameter Sets

For speaker verification or for each comparison of an input utterance with a reference template in speaker identification using statistical features, roughly $40 \times M$ (M is the number of voiced frames in the spoken word) multiply-add calculations are necessary to get the statistical features and calculate the weighted distance for decision making after obtaining the time functions of fundamental frequency and LAR (log area ratio) parameters. On the other hand, for speaker recognition using dynamic features, roughly $10 \times M^2$ multiply-add calculations are necessary to do the dynamic time warping and get the overall distance. Hence, when the number of voiced frames is 40, which is a typical number for Japanese spoken words, the number of calculations for statistical features is only about one-tenth of that for dynamic features.

It would seem that statistical features, which give almost the same performance as dynamic features except in some very special conditions and need only one-tenth the calculation time, are better than dynamic features, although there remain some additional comparisons between these two feature sets which should be examined under conditions such as transmission of speech waves over a telephone network or other special conditions.

D. Combination of Statistical and Dynamic Features

Although statistical and dynamic features have been used separately in speaker recognition so far, and these features are originally extracted through the same spectral analysis procedure, they can be considered to have some speaker characterizing information independent of each other. Considering this, these two kinds of 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 long-term training set was used and the time interval between training and input utterances was set at three months. Spectral equalization was not applied to the speech wave, since it has been shown that this technique is not effective for these training and time interval conditions (see Section III-B).

Speaker identification and verification performance results based on the combination of the two kinds of features are shown in Fig. 11, which also shows results for single sets of features. This figure shows that, both in identification and verification, the recognition error rate can be reduced by half by combination of the two sets of features, irrespective of the number of

TABLE I
CONFUSION MATRICES FOR SPEAKER IDENTIFICATION USING COMBINATION OF STATISTICAL FEATURES AND DYNAMIC FEATURES
(A) Speaker Identification Using One Word
(B) Speaker Identification Using Two-Word Combination

Out / In	W	I	F	K	H	S	M	U	T	Recognition Accuracy
W	47								1	97.9%
I		48								100 %
F			48							100 %
K				48						100 %
H					46	2				95.8%
S					1	47				97.9%
M							48			100 %
U				1	1			46		95.8%
T									48	100 %
Mean Recognition Accuracy										98.7%

Out / In	W	I	F	K	H	S	M	U	T	Recognition Accuracy
W	24									100 %
I		24								100 %
F			24							100 %
K				24						100 %
H					22	2				91.7%
S						24				100 %
M							24			100 %
U								24		100 %
T									24	100 %
Mean Recognition Accuracy										99.1%

words for decision making. 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.

Confusion matrices for speaker identification combining the two feature sets are shown in Table I. The distribution of the error rate among customers for speaker verification for the same conditions is shown in Fig. 12. There seems to be no distinct concentration of error rate among the customers.

E. Supplementary Recognition Experiments Using Dynamic Features

1) *Effects of the Adjustment Window Width:* The effect of varying the width of the dynamic time warping adjustment window on the speaker recognition accuracy was examined by an identification experiment. Fig. 13 shows the results of the experiment in which the long-term training set is used and the time interval between training and input utterances is three months. This figure also includes the results of a spoken digit recognition experiment which was carried out by the author [13] to compare these results with the speaker identification results.

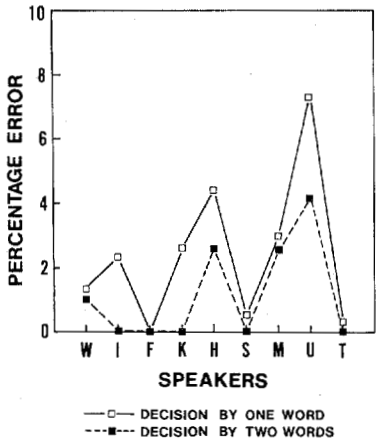

Fig. 12. Error rate as a function of customer in speaker verification using the combination of statistical and dynamic features for long-term training and three months interval.

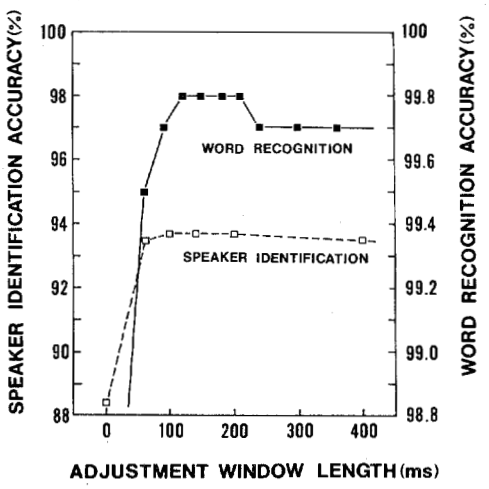

Fig. 13. Speaker identification and word recognition accuracies as functions of adjustment window width. In speaker identification, the long-term training set is used and the time interval between training and input utterances is set at three months.

Both in speaker identification and spoken digit recognition, the optimum width of the adjustment window to maximize recognition accuracy is between 120 ms and 200 ms. When the width is smaller than 120 ms, the accuracy decreases greatly. Even when the width is larger than 200 ms, the accuracy seems to decrease slightly.

Fig. 14 shows the relation between the adjustment window width and the identification accuracy when the short-term training set is used and the time interval between training and input utterances ranges from two or three days to ten months. Although the accuracy decreases as the time interval increases irrespective of window width, the optimum width is greater than or equal to 200 ms irrespective of the length of the time interval.

2) *Effectiveness of Weighting in the Distance Calculation:* A supplementary experiment was performed using a simple Euclidean distance as the distance measure between feature vectors on the dynamic warping plane, in order to compare recognition results with those obtained using the weighted distance which accounts for long-term variability of the feature vector. The

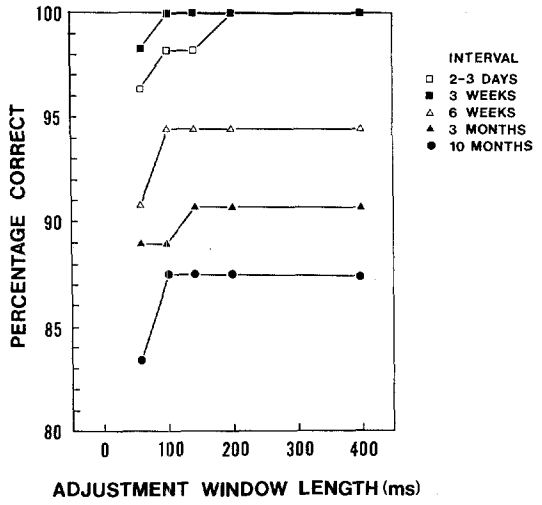

Fig. 14. Speaker identification accuracy as a function of adjustment window width for short-term training and various time intervals.

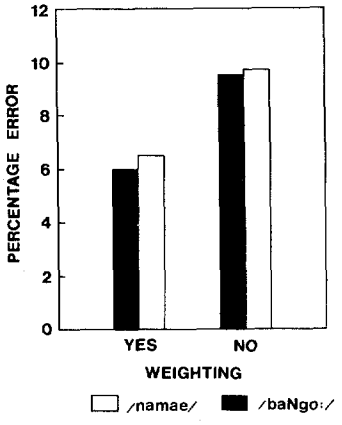

Fig. 15. Effectiveness of weighted distance in dynamic time warping on speaker identification using dynamic features for long-term training and three months interval.

long-term training set was used and the time interval between training and input utterances was set at three months in this experiment.

The average recognition error rate is shown in Fig. 15 compared with the results using the weighted distance. This figure indicates that the error rate associated with the Euclidean distance is 1.5 times larger than that associated with the weighted distance for both words. This result confirms the effectiveness of the weighted distance examined in this paper.

3) *Effect of the Number of Training Utterances:* In a previous study [7], the effect of the number of training utterances on speaker recognition accuracy using statistical features of spoken words was investigated, and it was found that 12 utterances recorded at four sessions were necessary and sufficient to get high accuracy for each customer. When the 12 utterances were recorded at only one session, very poor results were obtained, since they are insufficient for an adequate representation of the speaker's voice.

A supplementary experiment has been performed to study the effect of the number of training utterances on speaker recognition accuracy using dynamic features. The time interval between the last training utterance and input utterances

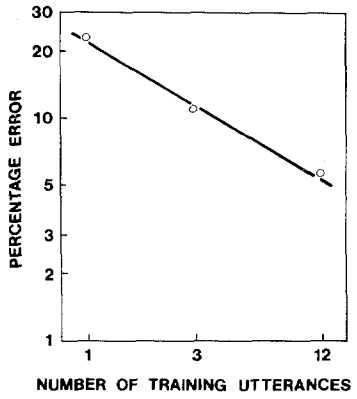

Fig. 16. Error rate as a function of the number of training utterances in speaker identification using dynamic features for three months interval.

was set at three months and two kinds of training sets were used, one consisting of only one utterance, and the other consisting of three utterances recorded at three sessions at intervals of three months.

Fig. 16 shows the results of the experiment including the previous results using 12 training utterances. The weighting factor for the distance measure used in dynamic time warping was calculated from the 12 training utterances and used for all the experimental conditions. These results indicate that the larger the number of training utterances becomes, the smaller the error rate becomes, with a linear relation between them on the log-log plane. An additional experiment for numbers of training utterances over 12 seems necessary in order to determine whether the error rate saturates or perhaps even decreases.

IV. Conclusion

This paper describes results of speaker recognition experiments using statistical features and dynamic features of speech spectra extracted from fixed Japanese word utterances. The effects of the long-term variability of these two sets of feature parameters on speaker recognition accuracy were studied and compared. Spectral equalization was evaluated as a technique for reducing these effects. The effectiveness of combining these two sets of features to produce high speaker recognition accuracy was studied. Some discussion of and experiments concerning speaker recognition using dynamic features were also included.

The experimental results show a difference between the performance of statistical features and dynamic features. Although the recognition accuracy decreases as the time interval between training and input utterances in both cases, the accuracy for dynamic features decreases less than the accuracy for statistical features as a function of the time interval when the time interval is as short as three months or less. Under this condition of the time interval, there is only a slight difference between the results for long-term training (training over ten months) and short-term training (training over ten days) in the case of dynamic features, whereas there is a large difference between the results for long-term training and short-term training in the case of statistical features. However, when the time interval is ten months or longer, dynamic features produce almost the same

results as statistical features, and long-term training is essential for both feature parameter sets.

Since the amount of calculation necessary for recognition using statistical features after obtaining the time functions of fundamental frequency and LAR's is only about one-tenth that for recognition using dynamic features, statistical features are considered to be better than dynamic features except when the training period is restricted to ten days or less and the time interval is about three months or less, in which case dynamic features produce much better results than statistical features.

When two words are used for the speaker recognition decision function, the error rate is nearly half of that obtained using a single word. When two words are used and the training is carried out over a period of ten months for each customer, high performance can be observed even for input utterances spoken five years later.

The speaker recognition experiment using statistical features ascertained the effectiveness of the spectral equalization procedure 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. When the time interval is less than a year and a half and spectral equalization is applied, there is no difference between the recognition accuracies using short-term training and long-term training. When the time interval is longer than a year and a half, it is essential to use long-term training and apply spectral equalization. When this method is adopted and two words are used for decision making, 99.5 percent and 95.6 percent recognition accuracy can be obtained for verification and identification, respectively, after a time interval of ten months, and 96.3 percent and 92.6 percent recognition accuracy can be obtained, respectively, after a time interval of five years.

Combination of the mean vector for the statistical features with dynamic features can reduce the identification and verification error rate to half the error rates obtained for a single feature parameter set of statistical or dynamic features.

In summary, the following procedures are desirable in speaker recognition to cope with long-term variability of feature parameters and obtain high performance after an interval of several years.

i) Use statistical features and do training over a period of about ten months for each customer. The necessary length of the training period depends on the length of the time interval between the last training utterance and the input utterance.

ii) Update reference template for each customer periodically.

iii) Apply a spectral equalization procedure.

iv) Use two or more words for decision making.

v) Combine dynamic features with statistical features.

Although the data base used in these experiments extended over a very long period, the comparison between dynamic and statistical features based on the use of only two words uttered by only nine male speakers may not be sufficient. Further investigations, current or projected, include large-scale evaluation over telephone lines.

ACKNOWLEDGMENT

The author especially wishes to acknowledge the guidance provided by Dr. S. Saito, former Director of the Saito Research Section, Dr. K. Noda, Director of the Research Division, and T. Koike, Chief of the Fourth Research Section, Musashino Electrical Communication Laboratory. The author also wishes to thank Dr. A. E. Rosenberg at Bell Laboratories and ASSP reviewers for the revision of this paper.

REFERENCES

[1] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, 1976.

[2] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 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.*, vol. AU-19, pp. 32–43, 1971.

[4] W. A. Hargreaves and J. A. Starkweather, "Recognition of speaker identity," *Language and Speech*, vol. 6, pp. 63–67, 1963.

[5] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Amer.*, vol. 46, pp. 1026–1031, 1969.

[6] S. Furui, F. Itakura, and S. Saito, "Talker recognition by long-time averaged speech spectrum," *Electron. Commun. Jap.* (Scripta Publishing Co., U.S.A.), vol. 55a, pp. 54–61, 1972.

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

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

[9] S. Saito and S. Furui, "Personal information in dynamic characteristics of speech spectra," in *Proc. 4th Int. Joint Conf. Pattern Recognition*, 1978, pp. 1014–1018.

[10] F. Itakura and S. Saito, "On the optimum quantization of feature parameters in the PARCOR speech synthesizer," in *Conf. Rec., IEEE 1972 Conf. Speech Commun. and Process*, New York, 1972, paper L4, pp. 434–437.

[11] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *1971 Proc. 7th Int. Cong. Acoustics*, paper 20-C-13, 1971.

[12] S. Furui, "Effects of long-term spectral variability on speaker recognition," *J. Acoust. Soc. Amer.*, vol. 64, suppl. 1, paper NNN28, p. S183, 1978.

[13] ——, "Evaluation of time-warping methods in isolated word recognition," in *1976 Spring Joint Meeting Acoustic Soc. Japan* (in Japanese), paper 3-2-16, 1976.

[14] ——, "New techniques for automatic speaker verification using telephone speech," *J. Acoust. Soc. Amer.*, vol. 66, suppl. 1, paper R4, p. S35, 1979.

[15] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 283–289, 1976.

[16] J. D. Markel, B. T. Oshika, and A. H. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 330–337, 1977.

[17] R. E. Wohlford, E. H. Wrench, Jr., and B. P. Landell, "A comparison of four techniques for automatic speaker recognition," in *Conf. Rec. 1980 Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, Denver, CO, 1980, pp. 908–911.

[18] G. R. Doddington, "Speaker verification," RADC-TR-U1-963700-F, 1974.

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

Sadaoki Furui (M'79), for a photograph and biography, see p. 272 of the April 1981 issue of this TRANSACTIONS.