

論文 / 著書情報
Article / Book Information

Title	Analysis of Talker Differences in Statistical Properties of Speech Spectra
著者	古井 貞熙
Author	Sadaoki Furui, Fumitada Itakura
掲載誌/書名	, Vol. 24, No. 5-6, pp. 418-728
Journal/Book name	Review of the Electrical Communication Laboratories, Vol. 24, No. 5-6, pp. 418-728
発行日 / Issue date	1976, 5

Analysis of Talker Differences in Statistical Properties of Speech Spectra

Sadaoki FURUI and Fumitada ITAKURA

Properties and measurement of intertalker and intratalker variability among statistical parameters derived from speech spectra are described. The speech wave is analyzed, using partial auto-correlation coefficients (PARCOR coefficients) and fundamental frequency. Statistical measures, such as averaged value, standard deviation and correlation coefficients between parameters in voiced portion, are derived. Talker recognition experiments are made using one to four words in combination. Results of these experiments reveal that the amount of difference is less for intratalker data than for intertalker comparisons. Characteristics of the long-term variation of these parameters are made clear, and the effectiveness of the combination of words is demonstrated.

1 Introduction

One of the most difficult problems in the extraction of personal information from a speech wave by some physical measurements lies in the fact that the measured characteristics of a talker may change after a long interval. Conventional methods of extracting the personal information from speech wave can be divided into text-independent and text-dependent categories. The first method uses static parameters, which are independent of the phonetic sequence comprising the speech. For instance, the long-time averaged speech spectrum⁽¹⁾⁽²⁾ and correlation matrix of the outputs of bandpass filters for long-time speech⁽³⁾ are investigated. Stable extraction of these parameters is possible and long-term variations therein can be normalized directly, but the personal information peculiar to each phoneme cannot be extracted. In the second method, sequences of dynamical parameters of a preselected speech material are subjected to time-normalization. For instance, the dynamical patterns of the formant, fundamental frequency and power are used.⁽⁴⁾ The dynamical parameters can be more informative for discriminating talkers than the text-independent static parameters, but it is more difficult to normalize long-term variations in dynamical parameters, because of their higher

dimensionality.

In order to use the complementary merits of the two methods, the time-averaged statistical parameters, derived from preselected sample words separately, are examined. The long-term variation of these parameters is taken into consideration and several words are used in combination in decision making. The effects of the word combination on the extraction of the personal information are examined quantitatively and the long-term variation of the parameters is studied in detail.

2 Method of Analysis

2.1 PARCOR Analysis

As is shown in Fig. 1, speech recorded in an anechoic chamber is passed through a 3.4 kHz low-pass filter and is converted into a discrete sequence by 8 kHz sampling and 11-bit quantization. From this sequence, 32 ms long blocks are extracted at 10 ms intervals and multiplied by the Hamming window. Then, the speech wave is analyzed into partial auto-correlation coefficients (PARCOR coefficients)⁽⁵⁾ K_{it} and fundamental period τ_{oi} . The fundamental period is extracted from the position of the peak of the correlation function of the prediction error, and

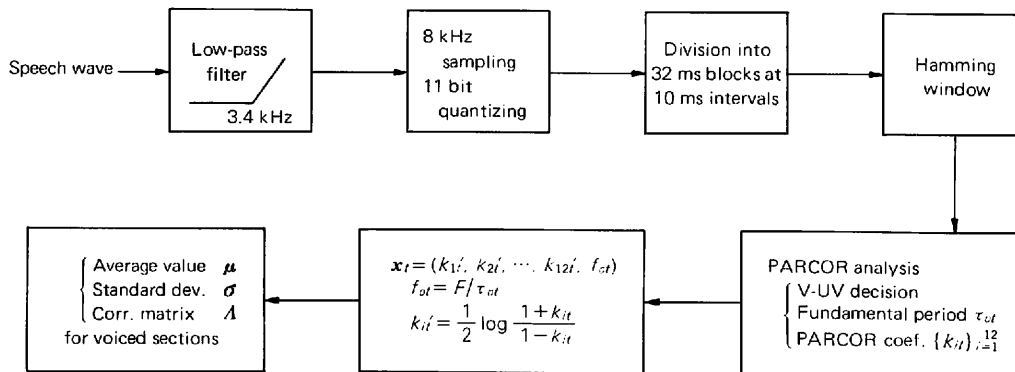
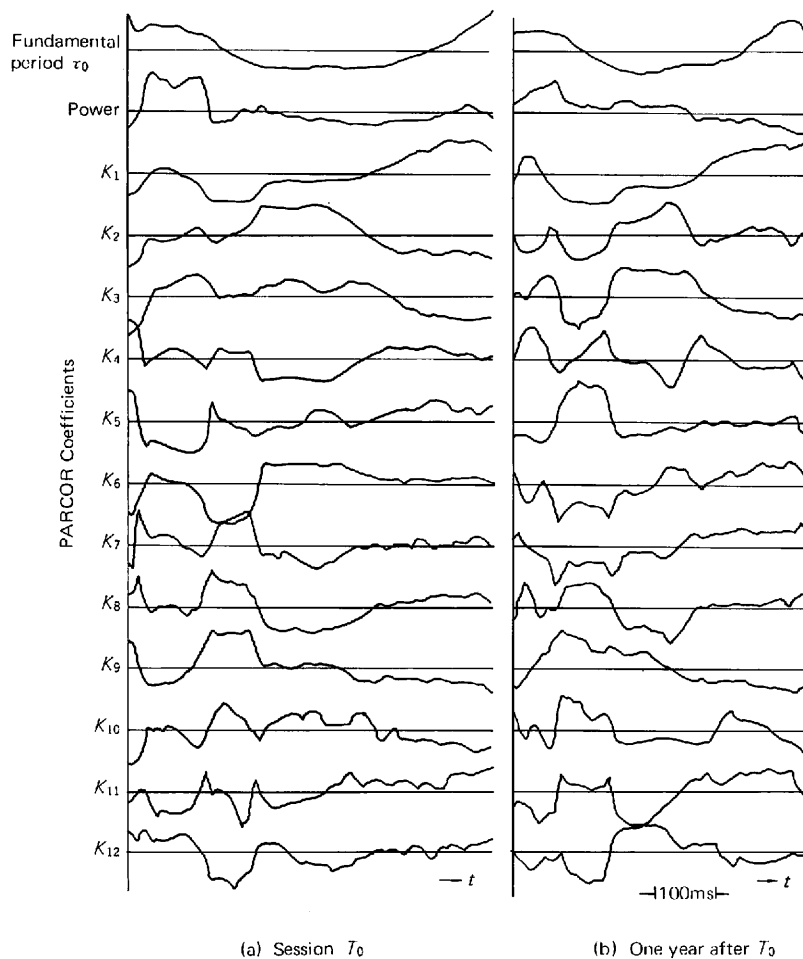


Fig.1—Analyzing method blockdiagram.

the voiced-unvoiced decision is made based on the peak value.

Figure 2 shows examples of the extracted parameters for voiced sections of the word

/namae/ uttered by talker *F* at two different sessions. Average value and dynamic range of each parameter are normalized in this figure. The correlation between parameters can be seen.

Fig.2—Examples of extracted parameters (word /namae/ uttered by talker *F*).

These correlation characteristics differ among talkers. The dynamic range of each parameter is also peculiar to an individual talker. As the PARCOR parameters can be approximately related to vocal-tract area functions, the above-mentioned features can be regarded to represent the dynamic characteristics of the vocal system.

2.2 Statistical Parameters

Since the PARCOR coefficient K_{it} never exceeds one in absolute value, the arctanh transformation⁽⁶⁾ is next applied to make the frequency distribution approach the normal distribution, as shown in Fig. 1. The fundamental period is transformed to the fundamental frequency, based on the same reasoning. Then, the acoustic feature at time t is represented by a 13-dimensional vector X_t , consisting of the fundamental frequency and the transformed 12 PARCOR coefficients:

$$\begin{aligned} X_t &= (x_{1t}, x_{2t}, \dots, x_{13t}) \\ &= (k_{1t}', k_{2t}', \dots, k_{12t}', f_{ot}). \end{aligned} \quad (1)$$

$$\begin{cases} f_{ot} = F / \tau_{ot} \\ k_{it}' = \frac{1}{2} \log \{ (1 + k_{it}) / (1 - k_{it}) \} \end{cases} \quad (2)$$

where F is the sampling frequency (8 kHz), τ_{ot} is the fundamental frequency, and k_{it} is the i -th PARCOR coefficient.

From the voiced sections in a word, average μ , standard deviation σ , covariance matrix Σ , and correlation matrix A for X_t are measured:

$$\mu = \frac{1}{N} \sum_{t=1}^N X_t \quad (3)$$

$$\Sigma = (\sigma_{ij}) = \frac{1}{N-1} \sum_{t=1}^N (X_t - \mu)' (X_t - \mu) \quad (4)$$

$$\sigma = (\sigma_{11}, \sigma_{22}, \dots, \sigma_{13,13}) \quad (5)$$

$$A = (\lambda_{ij}) \quad \lambda_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}}, \quad (6)$$

where N is the number of voiced sections. The arctanh transformation is also applied for λ_{ij} .

Four kinds of words, such as /bango:/, /namae/, /bakuon/ and /ko:gen/ are used in the experiment. These words are selected arbitrarily.

3 Analysis of Variance

Analyses of variance are made for each statistical parameter, in order to test the talker difference and to evaluate the amount of personal information in each parameter. Speech samples of four words, uttered by nine male talkers over a period of about three years at intervals of about three months, are used.

The amounts of talker effect are normalized by the F value of the significant level of 1% for each parameter and are averaged out for vectors μ and σ and for matrix A , respectively. The result is shown in Fig. 3. The talker effect in vector μ is largest, but the effects of vector σ and matrix A are also very large.

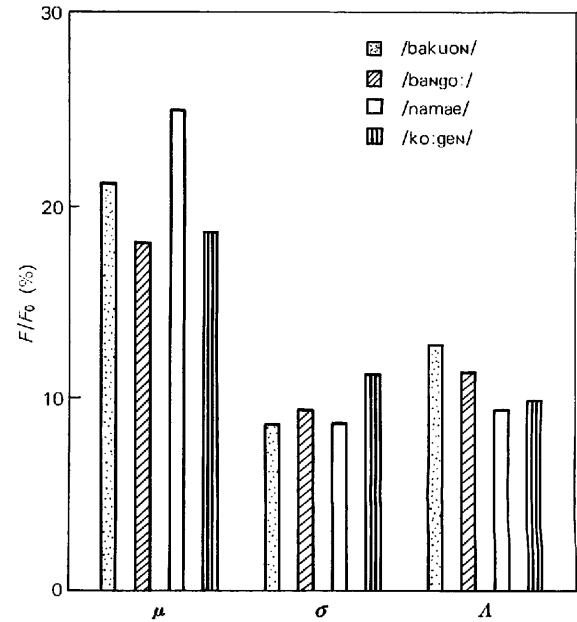


Fig. 3—Amount of talker effect for each parameter.

4 Parameters Time-varying Characteristics

4.1 Principal Component Analysis

In order to examine the short-term and long-term variation of parameter μ directly, principal component analyses are made for short-term samples and long-term samples of each word uttered by each talker. Every sample is projected onto the space spanned by the first, second and

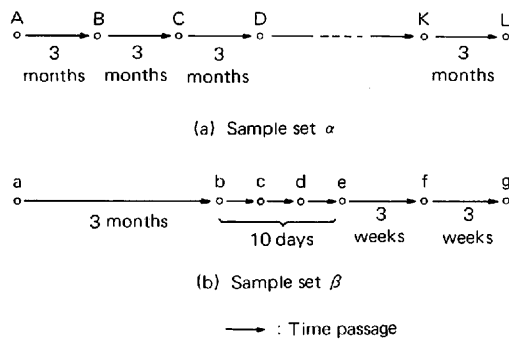


Fig. 4—Speech samples time allocation.

third principal components. A set of samples, $\alpha = \{A, B, C, \dots, L\}$ was collected over a period of about three years at intervals of about three months, as shown in Fig. 4(a). This data is used as long-term samples. Another set of samples, $\beta = \{a, b, c, \dots, g\}$, collected over a period of about five months at intervals of two or three

days to three months, as shown in Fig. 4(b), is used as short-term samples. Analyses are made after the covariance matrices are transformed into correlation matrices.

Figure 5 shows a part of the results. Speech samples of the word /bango:/ uttered by talker M, are used. Cumulative contribution ratios (the ratio of the variance of principal components to the total variance) of the first, second and third principal components are 71.2%, both for short-term and long-term samples. As the analyses are based on the correlation matrices, the absolute amounts of variation in the short-term and long-term cases cannot be compared directly, but some characteristics of the variation can be seen. In the result of the short-term samples, points b, c, d and e, which are collected over a period of about 10 days, are very close in comparison with other points. This reveals that the parameters are stable for a short period, but variations become larger for a long period. Long-term samples vary somewhat randomly at intervals of three months.

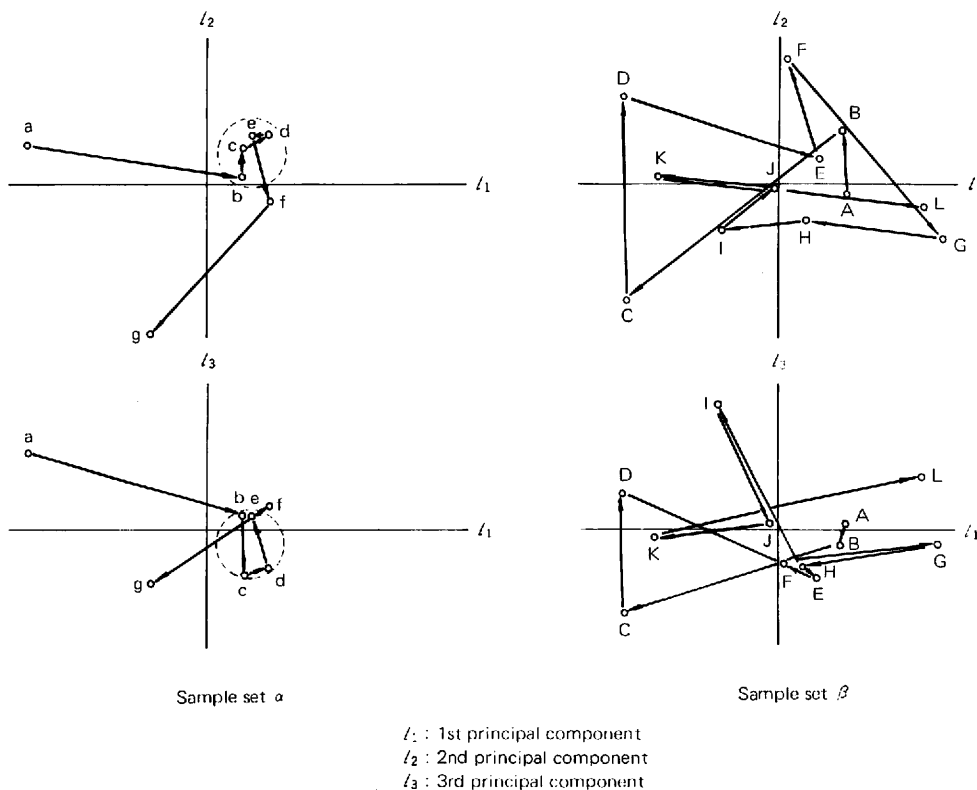


Fig. 5—Sample points projected to the space which consists of three principal components. (word /bango:/ uttered by talker M)

4.2 Relation between Time Interval and Variation Amount

The variation amounts of vector μ between various time intervals are measured by the Euclidian (squared) distance:

$$d(\tau) = E[\|\mu_t - \mu_{t+\tau}\|^2]. \quad (7)$$

The solid line curve in Fig. 6 shows the averaged results for four words and nine talkers. This result indicates that, when the time interval is less than six months, the amount of variation increases linearly as the interval is made longer. However, when the interval is more than six months, the amount of variation is almost uniform, regardless of interval length.

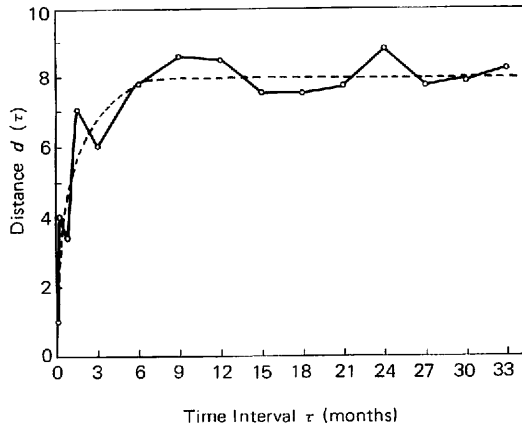


Fig. 6—Vector μ variation amount vs. time passage.

Euclidian (squared) distance can be related approximately to the auto-correlation coefficient of the parameter variation. The broken line curve in Fig. 6 corresponds to the distance variation curve derived from the auto-correlation coefficient of a random process, which is the summation of the output from an RC filter with a time constant of about two months and white noise. The former component corresponds to the time-varying characteristics of the parameters, the latter corresponds to variation independent of time passage. Similar analysis for each of 13 elements of vector μ reveals that every element has similar characteristics with nearly equal time constant. The analyses for the elements of σ and

Λ indicate that these elements have similar characteristics, though they are not so marked as μ elements are.

5 Talker Recognition Method

Talker identification and verification experiments are performed based on the statistical parameters measured by the method mentioned in Chap. 2. Talker identification means the process to determine which of M registered talkers a given speech comes from (M classes). Talker verification means the process to determine whether or not a given speech comes from a claimed talker A (two classes).

When a single word is used, recognition is made based on the following method.⁽⁷⁾ First, analysis of variance about talker-effect is made for each σ and Λ element, based on reference samples of nine talkers. The first m elements, whose talker effects are larger than those of other elements, are selected and combined to form an m -dimensional vector y :

$$y = (y_1, y_2, \dots, y_m). \quad (8)$$

A $(13+m)$ -dimensional vector z is constructed by vector y and μ :

$$z = (z_1, z_2, \dots, z_{13+m})$$

$$z_i = \begin{cases} \mu_i & : (1 \leq i \leq 13) \\ y_{i-13} & : (14 \leq i \leq 13+m). \end{cases} \quad (9)$$

The following weighted distances between an unknown sample and the average values of reference samples of individual talkers are used for talker identification and verification:

$$D(z_u, \theta_k) = |V|^{1/(13+m)} (z_u - \theta_k)' V^{-1} (z_u - \theta_k), \quad (10)$$

where z_u is the feature vector of unknown sample, θ_k is the registered pattern of k -th talker (average of reference samples), and $V = (V_{ij})$ is the covariance matrix averaged over nine talkers.

This distance corresponds to the recognition process, using the maximum likelihood under the assumption that the characteristic vector is drawn from a population of a multidimensional normal distribution and under the condition that the covariance matrix elements can be varied in

proportion to each other. In the elements of V , correlation elements for σ and $A(V_{ij}, i \neq j, i \text{ or } j > 13)$ are set to zeroes in order to reduce the processing time and memory capacity.

When more than one word is used, the distances defined by Eq. (10) for each word are averaged algebraically. In talker identification, the talker whose unknown sample is closest to a particular registered pattern, with respect to this distance, is determined to be the talker corresponding to that registered pattern. In talker verification, the sample whose distance to the registered pattern is less than a certain threshold value is determined to be the speech from that talker. Other samples are rejected as coming from other talkers. The threshold is set a posteriori so that the two kinds of error rate (the rate of rejecting samples, which should be accepted, and the rate of accepting samples, which should be rejected) are equal.

6 Recognition Experiment

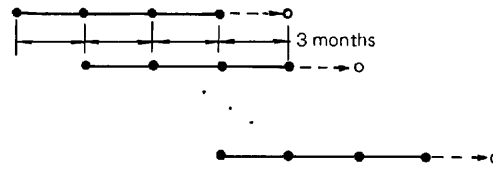
For quantitative evaluation of the effect of the combination of several words and long-term parameters variation, four kinds of recognition experiments were performed. The time allocation of known samples and unknown samples in each experiment is shown in Fig. 7. In each recording session, each talker utters every word three times at an interval of about one minute.

The registered talkers are nine male adults. In talker identification, one out of the nine registered talkers is determined. In talker verification, for each of nine registered talkers, samples of three kinds of impostors; A (8 persons), B (26 persons) and C (111 persons), are used as the samples to be rejected. Group A consists of 8 registered talkers other than the actual talker, group B consists of A and 18 other talkers, and group C consists of B and 85 other talkers.

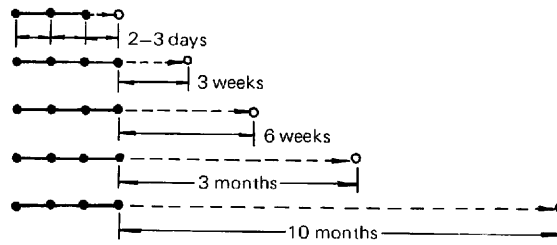
6.1 Experiment on Statistical Parameters Effectiveness

In this experiment, as shown in Fig. 7(a), reference samples were collected in four sessions at intervals of about three months. The time interval between the last reference sample and the test sample is three months. Experiments were performed for the unknown samples at eight different times for each of four words. In each experiment, 12 reference samples were used

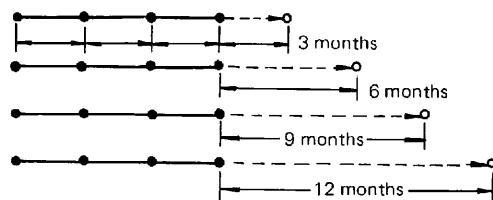
(a) Experiments on statistical parameters effectiveness



(b) Recognition based on short-term reference samples



(c) Recognition of unknown samples after a long time



(d) Experiments on the effects of the number of reference samples

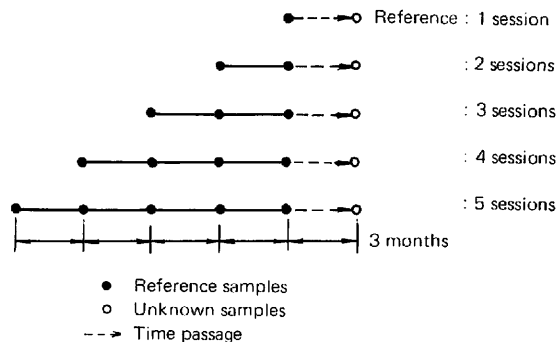


Fig. 7—Relations between reference samples and unknown ones in each recognition experiment.

for each word of each registered talker. In the identification, the number of unknown samples is 216 (9 talkers $\times$ 3 samples $\times$ 8 times) for each word. In the verification, the number of samples which should be accepted is also 216 and the number of samples which should be rejected is

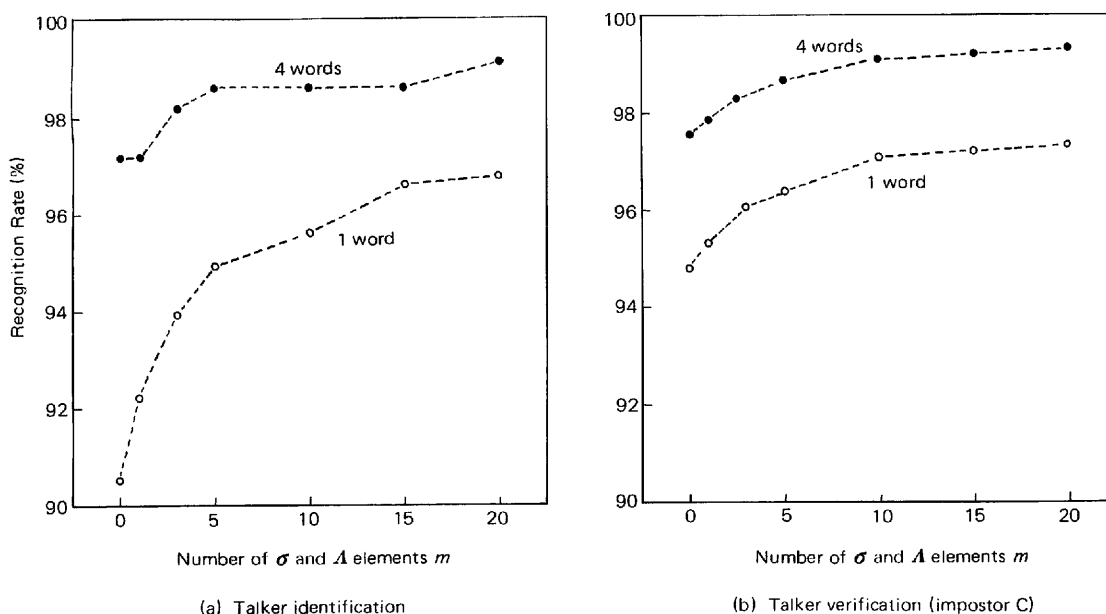


Fig.8—Results of experiment on statistical parameters effectiveness.

represented by 9 persons $\times$ 3 samples $\times$ 8 times $\times$ (the number of impostors), when impostor C is used, it is 23,976 for each word.

The number m of elements selected from σ and λ is varied from 0 to 20 in order to examine the effectiveness of these parameters. Figure 8 shows the results averaged over nine registered talkers. The results obtained by single word and those obtained by four words in combination are compared. The results reveal that, using 20 elements ($m = 20$), the error rate is reduced to one half of that obtained with $m = 0$ (i.e., only the average value μ is used). When four words are combined, the error rate is from one half to one fourth of that obtained based on only one word, independently of number m . When only one word is used, the average accuracy with $m = 20$ is 96.8% in identification and 97.3% in verification with impostor C. When four words are used, the average accuracy with $m = 20$ is 99.1% in identification and 99.3% in verification.

The confusion matrix in the identification and error rate for each registered talker in the verification for each of the four words and those obtained when four words are combined are examined. And the results reveal that there are some talkers whose samples often cause errors. On the

other hand, there are some talkers whose samples hardly ever cause errors. Moreover, the distribution of the error rates among individual talkers is different for different words.

6.2 Combining Words Effects

6.2.1 Relations between the Number of Characteristic Vector Dimensions and Error Rate

The results in the previous section suggest that the characteristics used for each word are fairly independent. For this reason, recognition accuracy may be improved by combining words in decision making. In order to examine this effect quantitatively, the number of combined words n and the number of σ and λ elements m are varied, and relations between the total number of dimensions $T = (13 + m) \cdot n$ and error rate are investigated. For example, if m is equal to 20 and n is equal to 4, T becomes equal to 132.

In Fig. 9, the relations between the total number of dimensions and error rates in verification are plotted on logarithmic scales. In part (a) of this figure, the lines join points with the same number of n . In part (b), the lines join points with

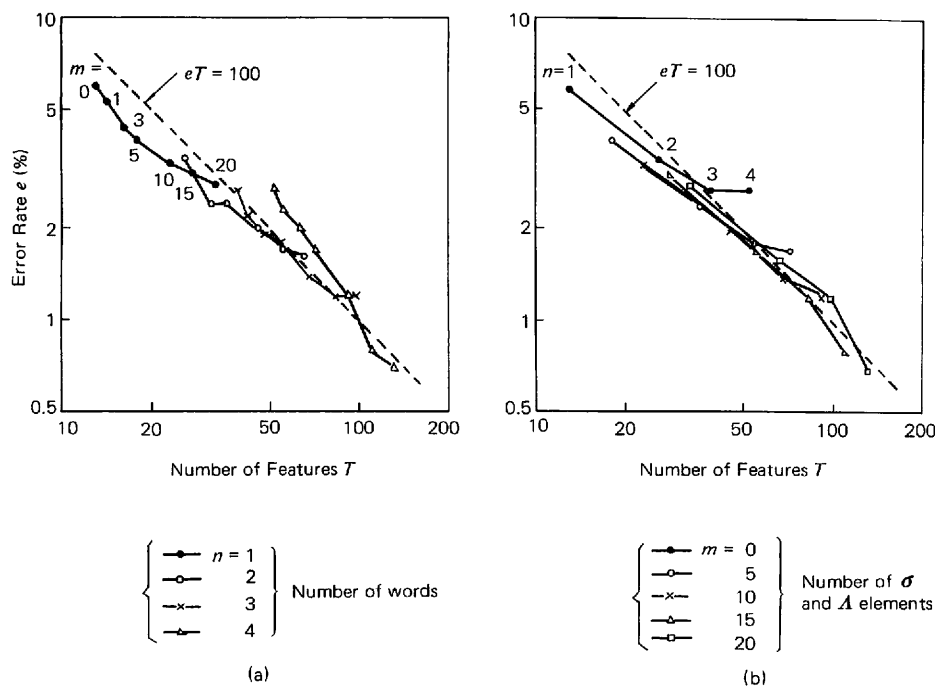


Fig. 9—Error rate for talker verification vs. number of features.

the same number of m . The product of the error rate in percentage and the number of dimensions is approximately constant:

$$eT \approx 100 \quad (11)$$

Similar results are obtained in identification.

As shown in Fig. 9(b), when m is less than 10, the improvement in error rate saturates with an increase in n . However, when m is 15 or 20, the error rate decreases linearly. For example, when m is equal to 20 and five words are combined, T becomes 165 and linearly extrapolated error rate e is 0.6%.

6.2.2 Origin for Words Combination Effects

Additional experiments are made, in which the distances for four words are not averaged. However, the four words are connected into one sample which is two to three seconds long. These results are compared with the above experiments using one to four words separately. The error rate for the pooled sample is about 3/4 of that obtained based on one word, but about 1.5 times

of that obtained based on four separate words. These results reveal that the pooled sample leads to better accuracy than a single word, but it is less accurate than when several words are combined in making the distance.

6.3 Recognition Based on Short-term Reference Samples

Recognition experiments were performed with reference samples collected in three or four sessions over a period of about ten days. The time interval between the last reference sample and the test sample was varied from two or three days to 10 months. This effect on accuracy is examined. Figure 10 shows the results obtained for the cases of one word and four combined words. In the one word case, results are averaged over four kinds of words. Number m is fixed at 20.

When the time interval between reference and test samples is two or three days, it is possible to obtain an accuracy of nearly 100% with one word, but this decreases with the increase of the interval of the order of three months. With four words in combination, when the time interval is six weeks or less, an accuracy of 100% is ob-

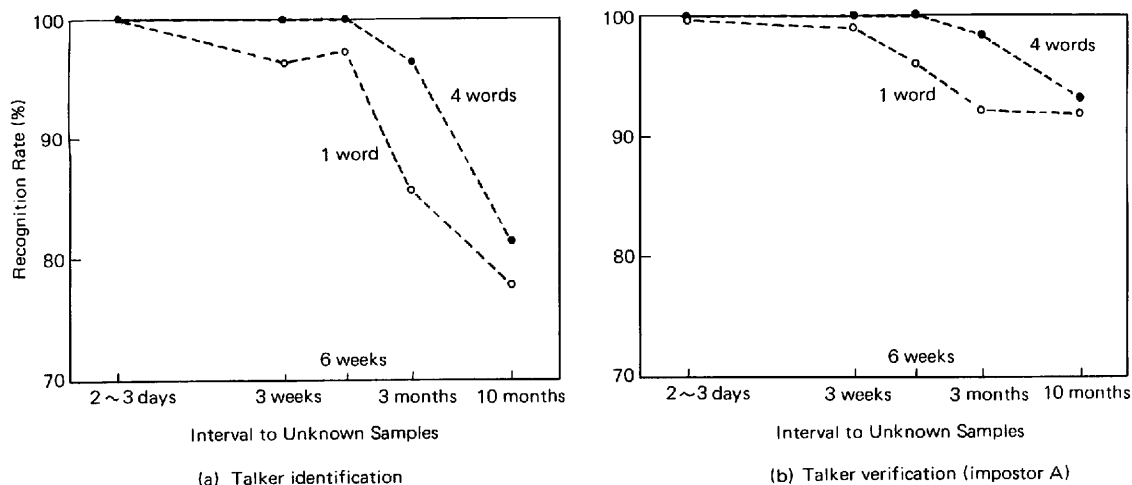


Fig.10—Results of recognition based on short-term reference samples ($m = 20$).

tained, but again, the accuracy becomes extremely low, when the time interval is as long as three months. These characteristics can be considered as a result of the time-varying characteristics of the parameters.

6.4 Recognition of Unknown Samples after a Long Time

Recognition experiments were made with

reference samples collected in four sessions at intervals of about three months, as in the recognition experiment covered in Section 6.1. The time interval between the last reference sample and the test sample ranged from 3 months to 12 months. Recognition experiments were performed for the test samples at five different times under the same conditions. The averaged recognition rate is shown in Fig. 11.

In all cases, the error rate is almost doubled,

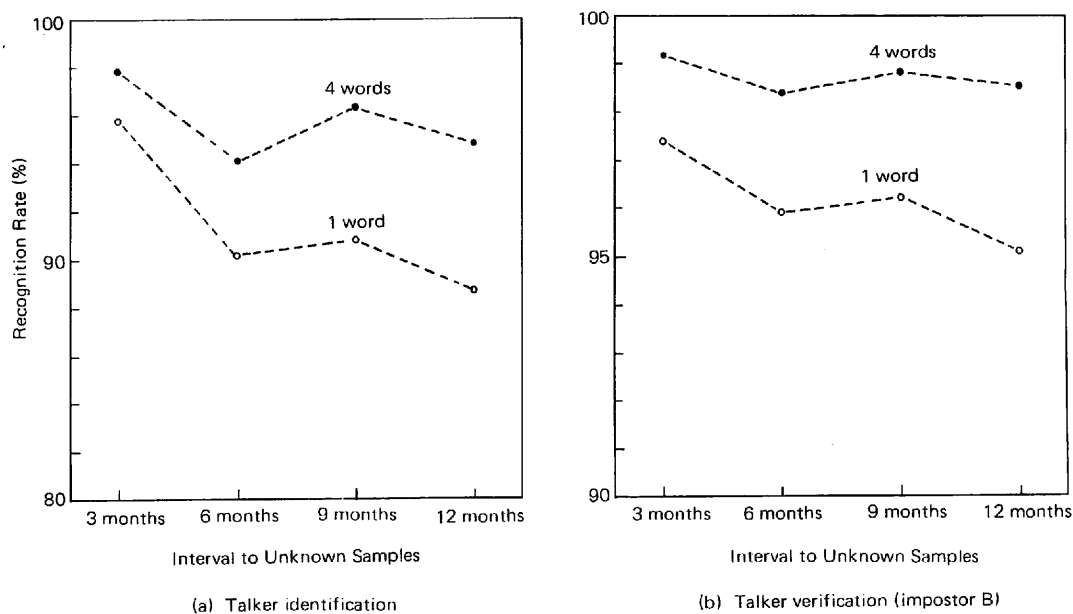


Fig.11—Results of recognition of unknown samples after a long time ($m = 20$).

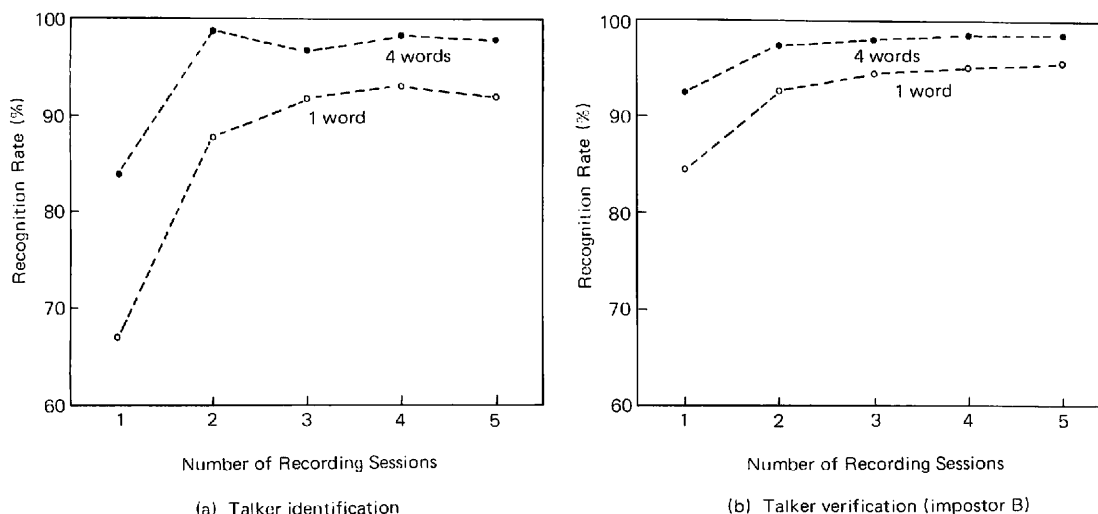


Fig.12—Results of experiments on effects of number of reference samples ($m = 20$).

as the interval increases from three months to six months. However, when the time interval is more than six months, the error rate is nearly constant. These results are consistent with the relation between the time interval and the amount of variation mentioned in Section 4.2.

6.5 Experiments on the Effects of the Number of Reference Samples

The time interval between the last reference sample and the test sample is three months. The number of recording sessions for reference samples is varied from 1 to 5 (every three months). Results are shown in Fig. 12. Experiments reported in Section 6.1 correspond to “4 sessions” in this experiment.

When the reference samples are taken at only one session (three samples), the recognition rate is very low and less than the results of the experiment reported in Section 6.3 with an interval of three months. There is a large difference from the results obtained with reference samples of two sessions or more. It is also seen that it is almost sufficient to collect reference samples in four sessions.

The results of experiments reported from Sections 6.1 to 6.5 indicate that it is necessary to take reference samples for a certain long period. These samples must be renewed successively in

order to obtain a high accuracy after a long interval. For example, it is necessary to collect samples four times with 3-month intervals in order to recognize unknown samples three months later. These results are similar to those obtained in the previous investigation of talker recognition by the long-time averaged spectrum.⁽¹⁾⁽²⁾

7 Conclusion

Speech sounds of several words were represented by the time sequences of the partial autocorrelation coefficients (PARCOR coefficients) and the fundamental frequency. Analysis of the personal information and talker recognition experiments were carried out, based on feature parameters which include several statistical measures, such as the average value, standard deviation and correlation coefficients between parameters. Multiple words were used for the decision. The results have shown that, when the reference samples are obtained in four sessions at 3-month intervals and four words are combined, a recognition rate of 99.1% is obtained for identifying a talker from nine talkers, and 99.3% for verifying a talker in 112 talkers, after three months from the last reference sample. Several causes of word combinations effectiveness have been made clear. Time-varying characteristics are similar to that of the random signal passed through an RC filter with a time constant of about two months.

Acknowledgement

The authors wish especially to acknowledge the guidance provided by Dr. Ken-ichi Noda, Director of the Research Division, Dr. Shuzo Saito, Director of the Saito Research Section, Dr. Shin-ichiro Hashimoto, Chief of 4th Research Section and other members of the 4th Research Section and the Saito Research Section.

References

- (1) S. Furui, F. Itakura and S. Saito: Talker Recognition by the Long-time Averaged Speech Spectrum, *Trans. IECE Jap.*, **55-A**, 10, p. 550, 1972.
- (2) S. Furui, F. Itakura and S. Saito: Personal Information in the Long-time Averaged Speech Spectrum, *Review of the E.C.L.*, NTT, Jap., **23**, 9-10, p. 1133, 1975.
- (3) K. P. Li and G. W. Hughes: Talker Differences as They Appear in Correlation Matrices of Continuous Speech Spectra, *J. Acoust. Soc. Amer.*, **55**, 4, p. 833, 1974.
- (4) R. C. Lummis: Speaker Verification by Computer Using Speech Intensity for Temporal Registration, *IEEE Trans.*, **AU-21**, 2, p. 80, 1973.
- (5) F. Itakura and S. Saito: Digital Filtering Techniques for Speech Analysis and Synthesis, *7th Int. Congr. Acoust.*, 25C1, 1971.
- (6) R. A. Fisher: Statistical Methods for Research Workers, Oliver & Boyd Ltd., 1958.
- (7) S. Furui and F. Itakura: Talker Recognition by Statistical Features of Speech Sounds, *Trans. IECE Jap.*, **56-A**, 11, p. 717, 1973.

* * * *