論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Study on Talker Dependent Features of Speech and Its Application to Speech Recognition
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	, , No. 149, pp. 1-15
Citation(English)	E. C. L. Technical Publication, , No. 149, pp. 1-15
発行日 / Pub. date	1976,

E.C.L. Technical Publication No. 149

Study on Talker Dependent Features of Speech and Its Application to Speech Recognition

THE ELECTRICAL COMMUNICATION LABORATORIES

NIPPON TELEGRAPH AND TELEPHONE PUBLIC CORPORATION

Study on Talker Dependent Features of Speech and Its Application to Speech Recognition

1 Introduction

Information conveyed by speech waveform can be divided into linguistic and personal information. The two kinds of information are closely related to each other, and the problem of how to separate these two kinds of information, how to extract the personal information, and how to normalize the talker differences in speech spectrum are both important and difficult ones. Investigations on these problems have a connection with machine recognition of words or sentences uttered by unspecified talkers, text-independent talker recognition and high quality speech synthesis.

During this study, personal information in the speech waveform is extracted by two kinds of feature parameters; long-time averaged speech spectrum of a short sentence as a text-independent parameter, and a set of statistical parameters derived from time sequences of fundamental frequency and partial auto-correlation coefficients (PARCOR coefficients)[5] of several spoken words as a text-dependent parameter. Properties and measurements of intertalker and intratalker variability among these parameters are examined by talker recognition experiments and statistical analysis based on speech samples of a long period.

For intertalker normalization in speech recognition, individualities of the vocal-cord wave spectrum and the vocal-tract length are extracted and effects of these characteristics on speech spectrum are approximately normalized. Spoken digits recognition experiments were made to ascertain the effectiveness of this method.

2 Personal Information in Long-time Averaged Speech Spectrum

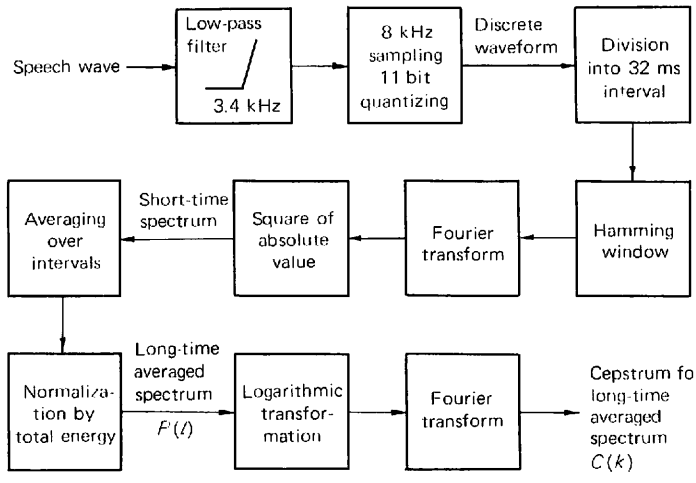
2.1 Measurement Method

As is shown in Fig. 1, a speech signal is converted into a discrete sequence and is divided into 32 ms long sections. In every section, short-time power spectrum is computed, and the spectra of all the sections are averaged to derive the long-time averaged spectrum. The long-time averaged spectrum is normalized in terms of total energy and then goes through logarithmic and Fourier transformations to obtain the cepstrum. As is easily seen from this definition, lower elements of the cepstrum give a general indication of the logarithmic spectrum pattern and higher elements give the microscopic structure. In other words, the transformation to the cepstrum makes it possible to separate these two aspects.[1]

The pronounced sentence used in the tests is a paragraph of a fairy tale (about 10 seconds long), and no specification is given to the talker concerning the utterance speed, pitch, or intensity of his speech. Speech samples were collected from nine males over a long period, at intervals of several days to three months, and for many other males at one session.

2.2 Recognition Model

A recognition model was constructed for the cepstrum of the long-time averaged spectrum, taking its long-term variation into consideration, and experiments were performed on the identification and verification of the talker. Here, identification of the talker means a process



$$C(k) = \frac{1}{N} \sum_{l=0}^{N-1} \{\log P(l)\} \cos(2\pi kl/N)$$

$$0 \leqq k,\ l \leqq N-1$$

Fig.1—Measuring method blockdiagram of cepstrum for long-time averaged spectrum.

whereby it is determined from which of the registered talkers the given speech came, and verification of the talker means a process whereby it is determined whether or not the given speech is from the registered talker.

Recognition or decision is made based on the weighted distance between an unknown sample and the average value of reference samples:

$$S(\boldsymbol{C}, \{\boldsymbol{D}_{rj}\}) = |V|^{1/N} (\boldsymbol{C} - \bar{\boldsymbol{D}}_r) V^{-1} (\boldsymbol{C} - \bar{\boldsymbol{D}}_r)^t \tag{1}$$

where $\boldsymbol{C}$ is the cepstrum of an unknown sample (N-dimensional vector), $\boldsymbol{D}_{rj}$ is the cepstrum of the j-th sample of registered talker r, $\bar{D}_r = \left(\sum_{j=1}^{n} D_{rj}\right)/n$ is the average value of reference samples, V is the covariance matrix of the cepstrum averaged over all registered talkers (an $N \times N$ matrix) and n is the number of reference samples for each registered talker.

In the case of talker identification, the talker whose unknown sample is closest to a particular standard pattern, the average value of reference samples, in regard to this distance is determined to be the talker corresponding to that standard pattern. In the case of talker verification, the sample, whose distance to the standard pattern is less than a certain threshold value, is determined to be the speech from that talker, and other samples are rejected as coming from other talkers. The threshold is conveniently 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.

For quantitative cepstrum temporal variation effect evaluation, three kinds of recognition experiments were performed. The time allocation of reference and unknown samples in each experiment is shown in Fig. 2. In each recording session, each talker utters speech samples 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,

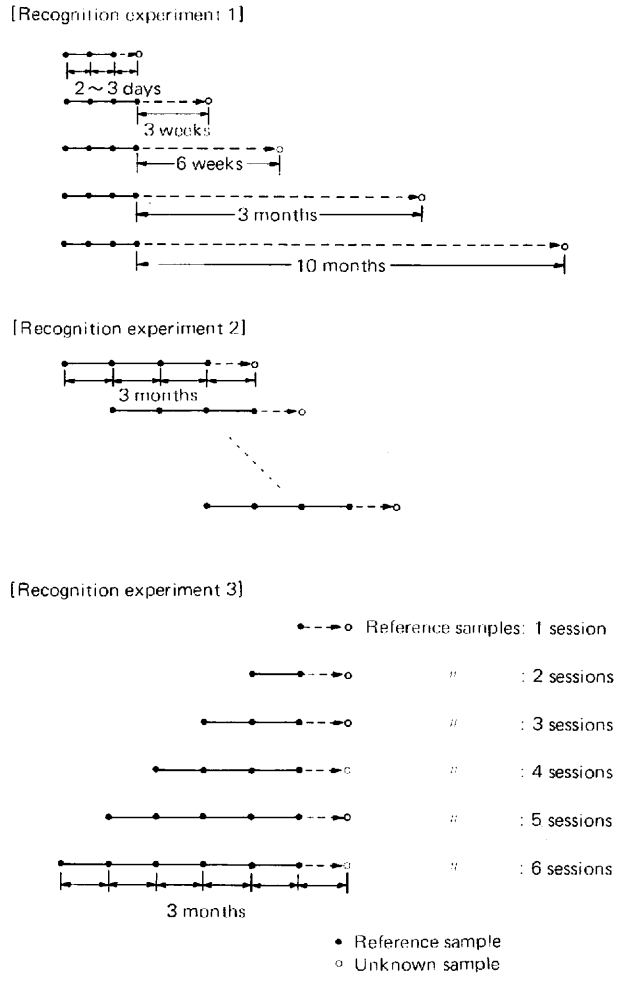


Fig.2—Relations between reference samples and unknown samples in three recognition experiments.

samples of three kinds of impostors; A (8 persons), B (26 persons) and C (111 persons), are used as samples to be rejected. Group A consists of 8 registered talkers other than the actual talker, group B consists of group A members and 18 other talkers, and group C consists of group B members and 85 other talkers.

As is seen from the description of the method in Section 2.1 and from the even property of the power spectrum, the cepstrum obtained effectively has 129 dimensions. Its higher components, however, have only to do with the microscopic structure of the spectrum pattern, and have less significance in the recognition process. Accordingly, referring to the results of preliminary experiments, the number of dimensions (N) of the cepstrum to be used in the reconginition is set at 31, neglecting the higher components.

2.3 Recognition Results

Figure 3 shows the results of recognition experiment 1. Average rates for the nine talkers are shown as functions of the time interval between reference and unknown samples. These

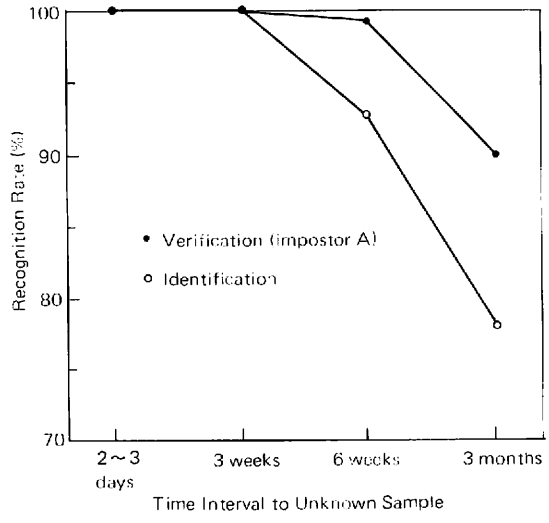


Fig.3—Identification and verification rate for recognition experiment 1.

results show that, when the time interval between reference and unknown samples is short, for example from several days to three weeks, the recognition rate is high, but it becomes low with increase of the interval and extremely so when the interval is of the order of three months.

Table 1 shows the results of recognition experiment 2. The unknown samples are the same as those in recognition experiment 1. It is seen that, by using the long-term data as reference samples and by setting the weighted distance with consideration of its variation, a better recognition rate can be achieved for unknown samples uttered three months later than the last reference samples.

Table 1 Identification and verification rate for recognition experiment 2.

	Recognition Rate
Identification	90.8%
Verification (impostor A)	95.5
Verification (impostor B)	93.5

By recognition experiment 3, it has been made clear that, if reference samples are taken only at one session, the recognition rate is very low, but it becomes higher with increase of the number of sessions and saturates at about four sessions.

2.4 Long-time Averaged Spectrum Variation

In order to examine the time-varying characteristics of cepstrum elements derived from long-time averaged spectrum, their variances over a short period (about 10 days) and a long period (about a year and a half) were estimated. And it was indicated that, common to all

talkers, the variances of lower elements are very large for a long period. This means that, as the period becomes longer, the variation in the general indication of the long-time averaged spectrum pattern becomes more significant than the variation in its fine structure.

3 Personal Information in Statistical Parameters of Fundamental Frequency and PARCOR Coefficients

3.1 Measurement Method

Several kinds of spoken words were used in the measurement. As is shown in Fig. 4, the speech waveform is converted into a discrete sequence, in the same manner as the measurement of the long-time averaged spectrum, and 12 PARCOR coefficients and fundamental period are analyzed every 10 ms. The fundamental period is extracted from the position of the peak of the correlation function of the prediction error, and the voiced-unvoiced decision is made based on the peak value. PARCOR coefficients and fundamental period are transformed to the log area ratios and fundamental frequency, respectively, to make the frequency distribution of each parameter approach normal distribution. Then, the acoustic feature at time t is represented by a 13-dimensional vector $\boldsymbol{X}_t$, consisting of the fundamental frequency f_{0t} and the 12 log area ratios $\{g_{it}\}_{i=1}^{12}$.

$$\begin{aligned}\boldsymbol{X}_t &= (x_{1t}, x_{2t}, \cdots, x_{13t}).\\ &= (g_{1t}, g_{2t}, \cdots, g_{12t}, f_{0t}).\end{aligned} \tag{2}$$

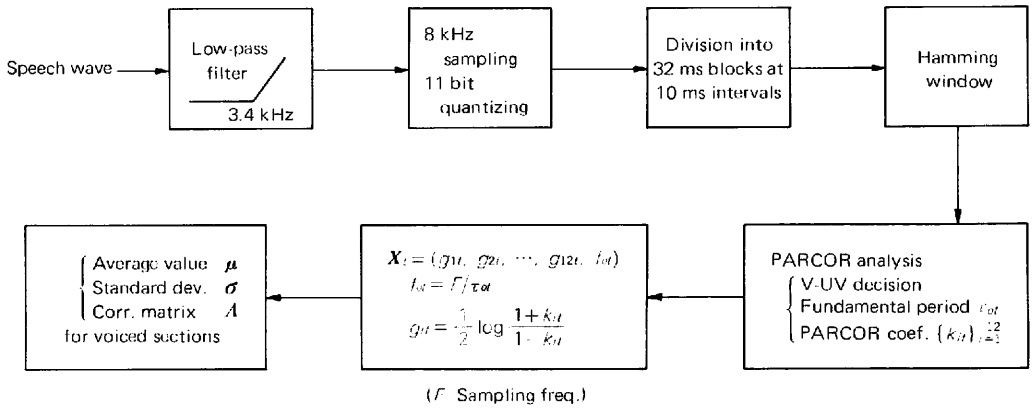


Fig.4—PARCOR coefficients and fundamental frequency statistical parameters measuring method blockdiagram.

From the voiced sections in a word, average $\boldsymbol{\mu}$, standard deviation $\boldsymbol{\sigma}$, covariance matrix $\boldsymbol{\Sigma}$ and correlation matrix $\boldsymbol{A}$ for $\boldsymbol{X}_t$ are measured:[2]

$$\boldsymbol{\mu} = \left(\sum_{t=1}^{M} \boldsymbol{X}_t\right)/M \tag{3}$$

$$\boldsymbol{\Sigma} = (\sigma_{ij}) = \left\{\sum_{t=1}^{M} (\boldsymbol{X}_t - \boldsymbol{\mu})^t (\boldsymbol{X}_t - \boldsymbol{\mu})\right\}/(M-1) \tag{4}$$

$$\boldsymbol{\sigma} = (\sigma_{11},\ \sigma_{22},\ \cdots,\ \sigma_{13,13}) \tag{5}$$

$$\boldsymbol{\Lambda} = (\lambda_{ij}) \qquad \lambda_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\,\sigma_{jj}}}, \tag{6}$$

where M is the number of voiced sections in a word. The same transformation as that applied to PARCOR coefficients is also applied to λ_{ij}.

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

3.2 Recognition Model

Talker identification and verification experiments are performed based on the statistical parameters. When a single word is used, recognition is made on the basis of the following method. First, analysis of variance concerning talker-effect is made for each $\boldsymbol{\sigma}$ and $\boldsymbol{\Lambda}$ element, making use of reference samples obtained from 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 $\boldsymbol{y}$:

$$\boldsymbol{y} = (y_1,\ y_2,\ \cdots,\ y_m). \tag{7}$$

A $(13+m)$-dimensional vector z is constructed by vector $\boldsymbol{y}$ and $\boldsymbol{\mu}$:

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

$$z_i = \begin{cases} \mu_i & : (1 \leqq i \leqq 13) \\ y_{i-13} & : (14 \leqq i \leqq 13+m). \end{cases} \tag{8}$$

The weighted distance indicated in Section 2.2 between an unknown sample and the average value of reference samples obtained from an individual talker is used in the decision making. When more than one word are used, the distances for all words are averaged algebraically. In the same way as in the experiment concerning the long-time averaged spectrum, three kinds of recognition experiments are performed.

3.3 Recognition Results

Figure 5 shows the results of recognition experiment 1, obtained for the cases of one word and four words. In the one word case, results are averaged over four kinds of words. The number m of $\boldsymbol{\sigma}$ and $\boldsymbol{\Lambda}$ elements used in the recognition is fixed at 20. These results are quite similar to the results obtained for the long-time averaged spectrum (Fig. 3).

The results of recognition experiment 2 are shown in Fig. 6. The number m is varied from 0 to 20 in order to examine the effectiveness of these parameters. The results reveal that, using 20 elements (m = 20) selected from $\boldsymbol{\sigma}$ and $\boldsymbol{\Lambda}$, the error rate is reduced to half of that obtained with m = 0 (i.e., only the average value $\boldsymbol{\mu}$ is used). When four words are combined, the error rate is from 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. These values are extremely higher than those of long-time averaged spectrum. This shows the effectiveness of the text-

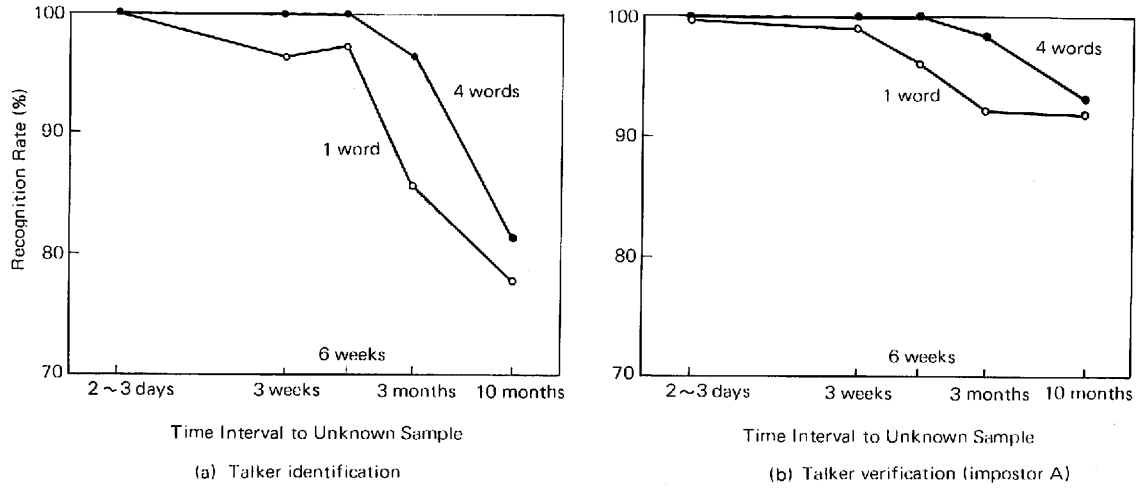


Fig.5—Identification and verification rate for recognition experiment 1 ($m = 20$).

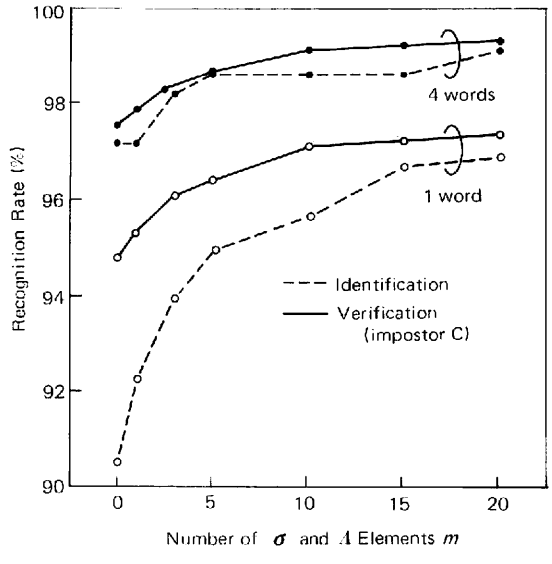


Fig.6—Identification and verification rate for recognition experiment 2.

dependent information.

Figure 7 shows the relation between the total number (T) of dimensions of feature vector and the error rate (e) of talker verification using impostor C, when the number of combined words (k) and the number of σ and Λ elements (m) are varied. The relation between k, m, and T is as follows:

$$T = (13+m) \cdot k. \tag{9}$$

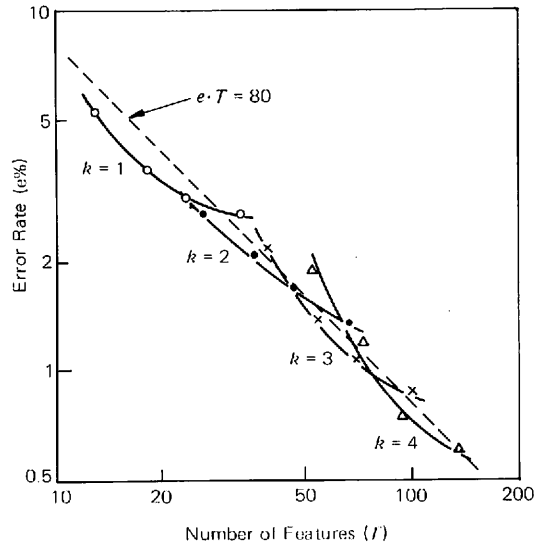


Fig.7—Error rate for talker verification (impostor C) vs. number of features.

Both axes are indicated by logarithmic scale. It can be seen that the product of the error rate in percentage and the number of dimensions is approximately constant:

$$e \cdot T \simeq 80. \tag{10}$$

Figure 8 shows the results of recognition experiment 3. Recognition experiment 2 corresponds to "4 sessions" in this experiment. When the reference samples are taken at only one session (three samples), the recognition rate is very low. Therefore, there is a large difference

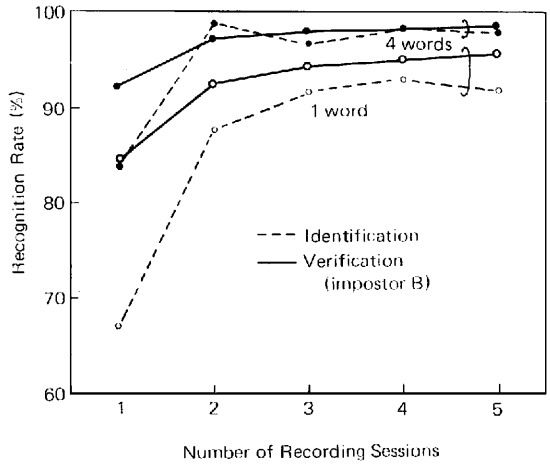


Fig.8—Identification and verification rate for recognition experiment 3 ($m = 20$).

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 at four sessions.

The results of experiments on the long-time averaged spectrum and the statistical parameters of PARCOR coefficients and fundamental frequency, reported from Sections 2.1 to 3.3, indicate that it is necessary to take reference samples for a certain long period to obtain a high accuracy after a long interval.

3.4 Parameter Temporal Variation

In order to examine the short-term and long-term variation of the 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 third principal components. A set of samples, $\alpha = \{A, B, C, \ldots L\}$ was collected over a period of about three years at intervals of about three months, as shown in Fig. 9(a). These data are used as long-term samples. Another set of samples, $\beta = \{a, b, c, \ldots, g\}$, collected over a period of about five months at intervals of two or three days to three months, as shown in Fig. 9(b), is used as short-term samples. Analyses are made after the covariance matrices are transformed into correlation matrices.

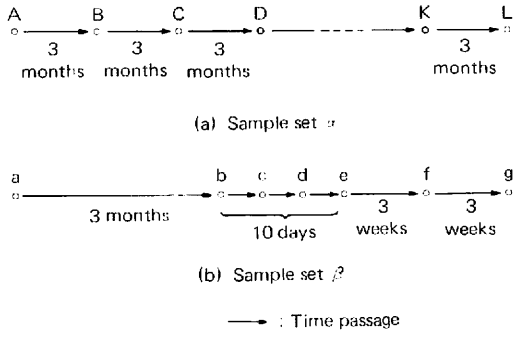


Fig.9—Long-term and short-term speech sample time allocation.

Figure 10 shows a part of the results, which is obtained for speech samples of the word /baNgo:/ uttered by a male talker. Cumulative contribution ratio (the ratio of principal components variance to total variance) of the first, second, and third principal components is 71.2%, both for short-term and long-term samples. In the results obtained for 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.

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

$$d(\tau) = E_t[\| \mu_t - \mu_{t+\tau} \|^2]. \qquad (11)$$

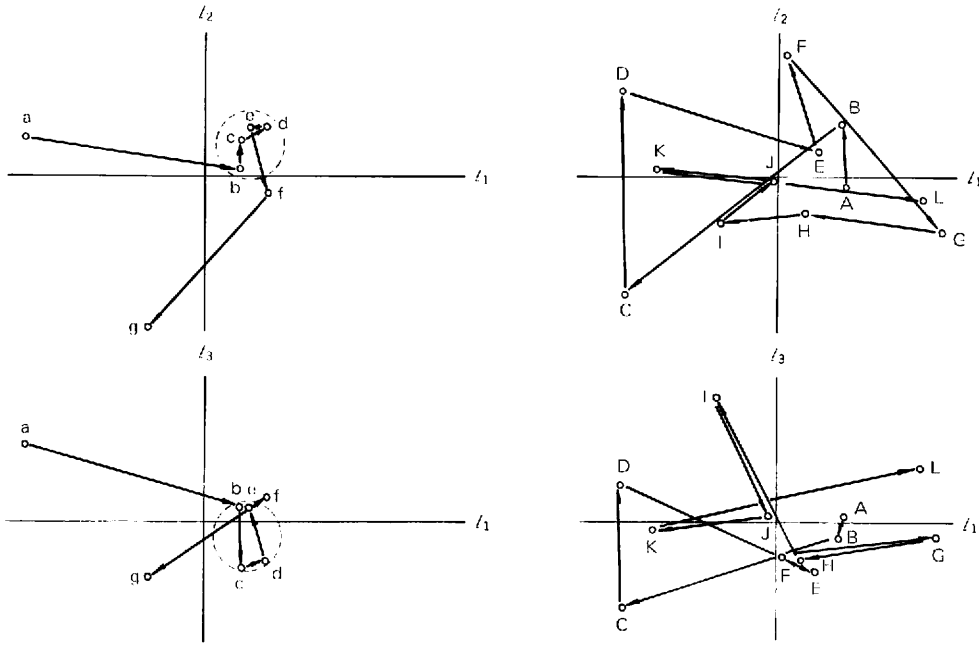


Fig.10—Sample points projected onto the space spanned by three principal components. (word /baNgo:/ uttered by a male talker).

The solid line curve in Fig. 11 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. The results shown in Figs. 10 and 11 are in good agreement.

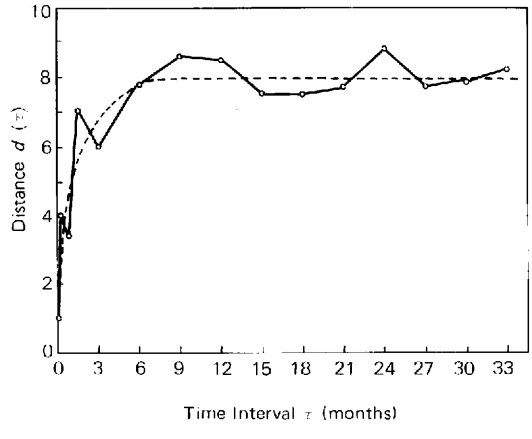


Fig.11—Variation amount for vector μ vs. time passage.

Euclidian (squared) distance can be related approximately to the auto-correlation coefficient of the parameter variation. The broken line curve in Fig. 11 corresponds to the distance variation curve derived from the auto-correlation coefficient of a random process, which is a 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.

4 Talker Dependent Parameter Temporal Variation Factors

4.1 Analyzing Method

Talker dependent parameter temporal variation factors and the method to remove these effects on the talker recognition are examined on the basis of the vocal system mechanism. Speech waveform could be regarded approximately as the output of the cascade connection of electrical circuits characterized by the vocal-cord, the vocal-tract and the radiation characteristic, driven by the source signal composed of the outputs of an impulse generator and a white noise generator. In actual practice, it is extremely difficult to accurately separate the vocal-cord, vocal-tract and radiation characteristics from the speech waveform, but an approximate separation can be achieved on some assumptions. The long-time averaged speech spectrum can be considered to approximately represent the vocal-cord characteristics, including the radiation characteristic, since the vocal-tract characteristics for the various phonemes are averaged.[3] As it is known that the radiation characteristic can be represented by an approximately 6 dB/octave in the spectral domain, independent of the talker and the speech material, the long-time averaged spectrum can be regarded to represent the vocal-cord characteristic of the talker.

Figure 12 shows the vocal-cord and vocal-tract characteristics separation method. In this method, the vocal-cord characteristic is approximated, for simplicity, as a 1st-order system or a 2nd-order critical damping (CD) system. By means of the inverse filter of this characteristic, the vocal-tract characteristic is extracted from the speech waveform. The same spoken sentences as used in Section 2 are used as long-time speech samples. The inverse filter transfer function can be expressed as follows:

$$\begin{cases} \text{1st-order system:} & 1+\beta Z \\ \text{2nd-order CD system:} & 1+\gamma Z+\dfrac{\gamma^2}{4}Z^2, \end{cases} \qquad (12)$$

where $Z=e^{-j\lambda}$ and λ = radian/unit time. From the speech wave after the inverse filtering, PARCOR coefficients and fundamental frequency are extracted.

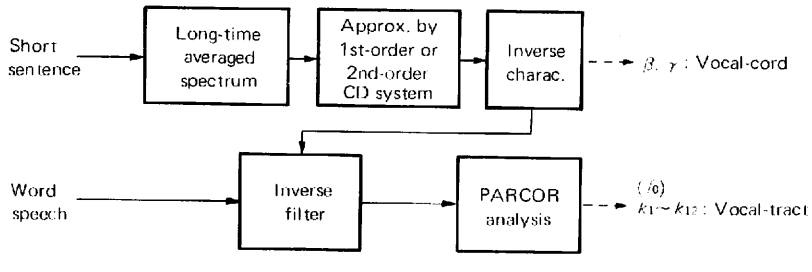


Fig.12—Vocal-cord and vocal-tract characteristics separating method.

4.2 Recognition Experiment

From the preliminary investigation, it has been made clear that the amount of temporal variation is greater for vocal-cord parameters β and γ than for vocal-tract parameters $\{k_i\}$. Also, when the vocal-cord characteristic is approximately removed from speech waveform by the inverse filtering, the temporal variation in the PARCOR coefficients is reduced, but the personal information in the parameters is scarcely reduced.

The effect of eliminating the vocal-cord characteristic on recognition accuracy is investigated by means of the talker recognition system examined in Section 3. Two words, /bango:/ and /namae/, are used and m, the number of elements extracted from σ and Λ, is fixed at 20. Two kinds of recognition experiments were performed. The results of the talker verification (impostor A) in recognition experiment 1, with or without inverse filtering, are shown in Fig. 13. It is seen that, by eliminating the vocal-cord characteristic, which has a strong and gradual temporal variation trend, the recognition errors can be reduced by about half.

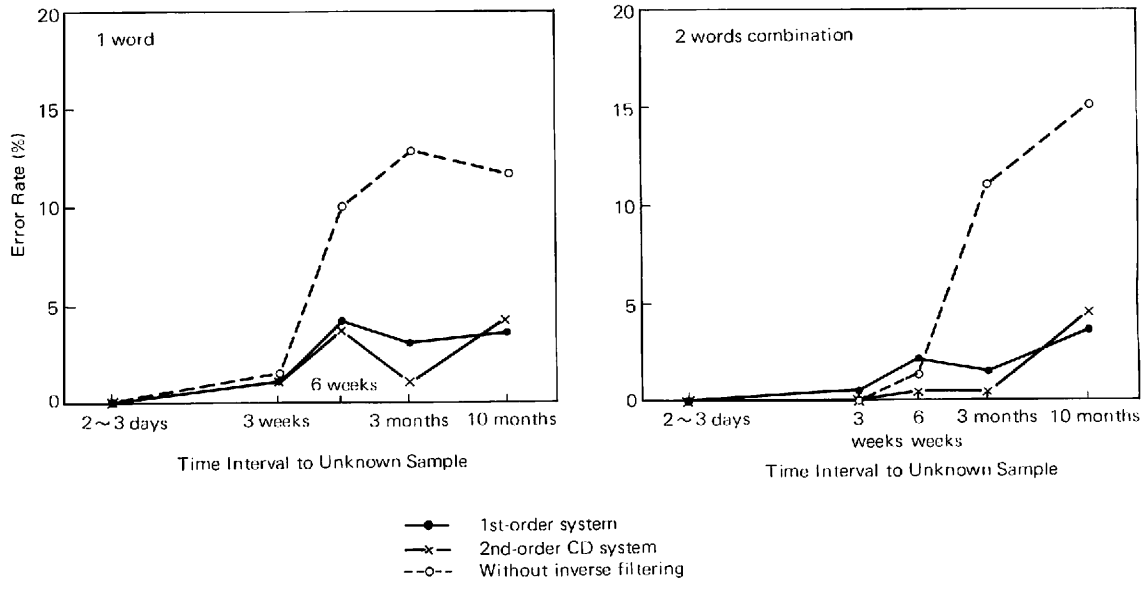


Fig.13—Effects of vocal-cord characteristic elimination on talker verification using short-term reference samples.

Next, using common unknown samples, a comparison is made between the case with the reference samples, covering a short period of about 10 days (short-term case), and the case where the reference samples cover a long period of about 10 months (long-term case). Time intervals between reference and unknown samples are varied from 3 months to 21 months. Based on the fact that the recognition error does not increase as the time interval between reference and unknown samples is varied from 3 months to 21 months, the results are averaged over all the unknown samples and presented in Table 2. The results show that, even in the short-term case, a recognition rate of about 99% can be achieved, using 2 words with the vocal-cord characteristic elimination technique; see the underlined part of Table 2. These rates are almost equal to the performance with long-term reference samples. It has been ascertained that the method whose investigation is reported in this section is highly effective to reduce speech feature parameters temporal variation and to extract stable personal information.

Table 2 Comparison of results with short-term and long-term reference samples.

Reference Samples	Inverse Filter	Talker Identification (9 indiv.)		Talker Verification (impostor A)	
		1 word	2 words	1 word	2 words
Short-term (10 days)	1st-order sys.	90.8%	99.4%	97.4%	99.0%
	2nd-order CD sys.	92.3	99.4	97.2	98.9
	None	88.3	94.5	92.7	92.8
Long-term (10 months)	1st-order sys.	97.2	99.4	97.6	99.5
	2nd-order CD sys.	96.0	98.8	97.1	98.8
	None	98.2	99.4	97.9	99.5

5 Intertalker Normalization in Spoken Word Recognition

5.1 Normalization Method

It is important to solve the problem of how to learn and normalize talker differences efficiently in order to achieve high accuracy in the recognition of spoken words, when the vocabulary is very large and the words are uttered by unspecified talkers. Among many speech waveform individuality factors, the vocal-cord wave spectrum and the vocal-tract length can be regarded as very important ones. Strictly speaking, these factors may vary from phoneme to phoneme, but they can be considered to be approximately invariant, at least, for vowels.

Figure 14 shows the spoken word recognition system, including intertalker normalization.[4] As a measure of the vocal-cord wave spectrum and the vocal-tract length, global pattern (inclination) of the speech spectrum, approximated as a 1st-order system, and the expansion-contraction rate of the auto-correlation pattern are estimated, respectively, based on learning

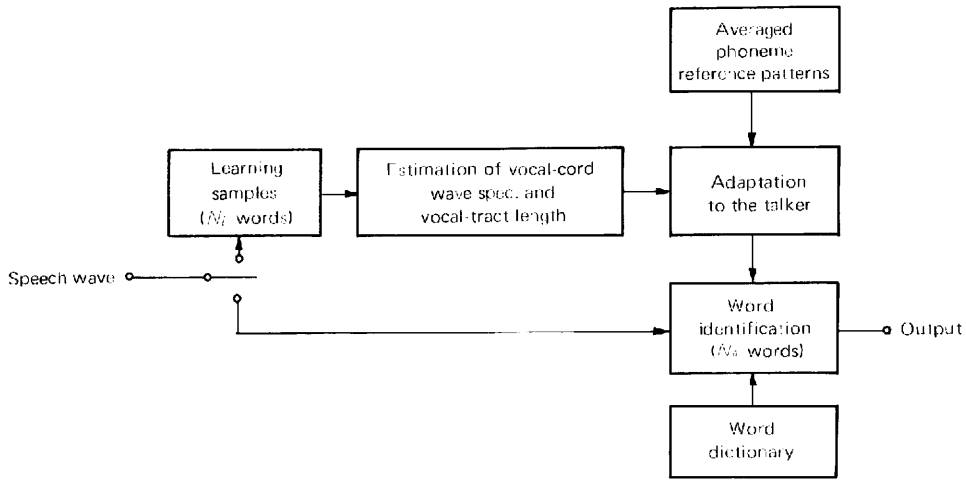


Fig.14—Spoken word recognition system, including intertalker normalization.

samples. The learning samples consist of N_l words, which is a part of the whole vocabularies of N_a words to be recognized. Reference patterns are stored by phoneme spectra averaged over many talkers. In the recognition of unknown samples, the reference patterns of vowels are modified and adapted to an individual talker, based on the vocal-cord wave spectrum and vocal-tract length extracted from the learning samples. Reference patterns of consonants are not modified. Word dictionary is stored by phoneme sequences and the word identification is performed, making use of the word dictionary and the similarity between input speech and phoneme reference patterns. Details of the identification process are presented in Reference (6).

5.2 Word Recognition Experiment

Recognition experiments were accomplished for spoken digits uttered by 30 male and 30 female talkers. The results of four experiments, without normalization, with vocal-cord wave normalization, with both vocal-cord wave and vocal-tract length normalization, and using individual reference patterns, are indicated in Fig. 15.

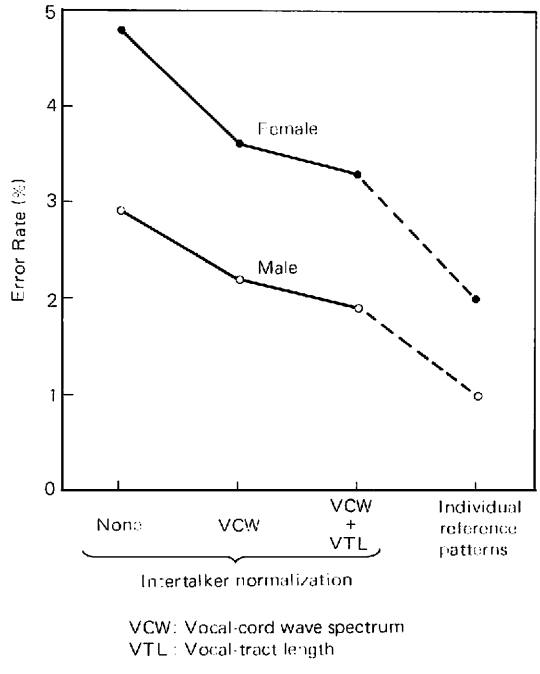


Fig.15—Intertalker normalization effects on spoken word recognition.

Error rate is reduced by one percent for male samples and by one and a half percent for female samples by normalization of the effects of the two factors. However, in comparison with the case using individual reference patterns, a gap of one percent or one and a half percent remains.

6 Conclusion

Personal information in speech waveform was extracted by two kinds of feature parameters; the long-time averaged spectrum of a short sentence and the statistical parameters derived

from PARCOR coefficients and fundamental frequency. Talker recognition experiments and several analyses were performed from the viewpoint of the temporal variation of these parameters. The effectiveness and properties of these parameters were made clear. Factors of the temporal variation were analyzed based on the vocal mechanism, and it was revealed that the method of eliminating the vocal-cord characteristic is useful for extracting the stable personal information. Intertalker normalization method in spoken word recognition by means of estimated vocal-cord wave spectrum and vocal-tract length was investigated and the effectiveness of this method was ascertained.

References

(1) 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.

(2) S. Furui and F. Itakura: Analysis of Talker Differences in Statistical Properties of Speech Spectra, *Review of the E.C.L.*, NTT, Jap., **24**, 5-6, p.418, 1976.

(3) S. Furui: An Analysis of Long-Term Variation of Feature Parameters of Speech and Its Application to Talker Recognition, *Trans, IECE Jap.*, **57-A**, 12, p.880, 1974.

(4) S. Furui: Effects of the Intertalker Normalization on Speech Recognition, *Acoust. Soc. of Japan Meeting*, 1-4-19, Oct. 1975.

(5) F. Itakura and S. Saito: Digital Filtering Techniques for Speech Analysis and Synthesis, *The 7th ICA*, 25CI, 1971.

(6) M. Kohda, S. Hashimoto and S. Saito: Spoken Digit Mechanical Recognition System, *Trans, IECE Jap.*, **55-D**, 3, p.186, 1972.

* * * *