

論文 / 著書情報
Article / Book Information

Title	Analysis of Temporal Variation of Talker Dependent Features
著者	古井 貞熙
Author	Sadaoki Furui
掲載誌/書名	, Vol. 25, No. 3-4, March-April, pp. 231-244
Journal/Book name	Review of the Electrical Communication Laboratories, Vol. 25, Nos. 3-4, March-April, Vol. 25, No. 3-4, March-April, pp. 231-244
発行日 / Issue date	1977, 7

U.D.C. 534.322:801.4''4'':534.62:519.26

Analysis of Temporal Variation of Talker Dependent Features

Sadaoki FURUI

Two kinds of feature parameters related to the exciting source and vocal tract characteristics of speech sound are analyzed from the viewpoint of temporal variation within and between talkers. It is made clear that the longer the time interval becomes, the larger the variation of exciting source parameters becomes within a talker, although its variation is rather small in a short time interval. Consequently, the variation in exciting source parameters between talkers becomes small over a long time interval. Results of talker recognition experiments reveal that the removal of exciting source characteristics from original speech signal improves talker recognition rate.

1 Introduction

A speech wave conveys information about talker differences, in addition to linguistic information. How to extract the personal information from a speech wave and how to represent it by some objective parameters is a problem which is closely related to the advanced machine recognition of speech and to the synthesis of high quality speech, as a part of the study of the man-machine system using speech. In the future, the results of the study may be applied to the machine recognition of talkers.⁽¹⁾

In general, one of the most difficult problems in extracting personal information from speech wave and represent it by objective parameters is that the feature parameters frequently vary after a long time interval, even if the same sentence is pronounced by the same talker. As parameters representing personal information, the author et al. have been investigating the cepstrum of the long-time averaged speech spectrum computed from a short sentence, and statistical features of fundamental frequency and partial auto-correlation coefficients (PARCOR coefficients)⁽²⁾ in voiced portion of spoken words, and have made various analyses on the characteristics of long-term variation and effectiveness of these parameters.⁽²⁾⁻⁽⁵⁾ It has been made clear that, although

these parameters are stable for a short period of about ten days or less, they vary slowly but largely for a long period. For this reason, in the talker recognition experiments based on the statistical parameters of spoken words, the following results were obtained for unknown samples 3 months later than the last reference samples. When the period of collecting reference samples is as long as 10 months, the recognition accuracy of 99% is achieved. However, when the period is as short as 10 days, the accuracy becomes very low.

A parameter which is stable for a short period, but shows large change over a long period, cannot be said to be truly representative of talker individuality. And, in the practical talker recognition system, it is undesirable that reference samples must be collected over a long period for every registered talker. Very few investigations have been made on the factors of the temporal variations of the talker dependent parameters and on the method of how to remove these effects. In this paper, exciting source (glottal) and vocal tract characteristics are extracted separately from the speech wave, based on the mechanism of the vocal system, and primary factors of the temporal variation of the speech parameters are analyzed. The results of the analysis are applied to the talker recognition procedure.

2 Acoustic Analysis

2.1 Separation of Glottal and Vocal Tract Characteristics

The speech waveform $s(t)$ can be regarded approximately as the output of the electrical circuits characterized by the glottal characteristic, the vocal tract characteristic and the radiation characteristic, having impulse responses $c(\tau)$, $t(\tau)$ and $r(\tau)$, respectively, driven by the source signal $e(t)$, composed of the outputs of an impulse generator and a white noise generator. This system can be represented in the frequency domain by

$$S(\omega) = R(\omega) T(\omega) C(\omega) E(\omega), \quad (1)$$

where $S(\omega)$, $E(\omega)$, $C(\omega)$, $T(\omega)$ and $R(\omega)$ are the Fourier transforms of $s(t)$, $e(t)$, $c(\tau)$, $t(\tau)$ and $r(\tau)$ respectively. In actual speech, $E(\omega)$, $C(\omega)$, $T(\omega)$ and $R(\omega)$ vary with time.

In actual practice, it is extremely difficult to accurately separate from the speech waveform $s(t)$ or its Fourier transform $S(\omega)$, the glottal, vocal tract and radiation characteristics, but an approximate separation can be achieved by making some assumptions. It is known that the radiation characteristic can be approximated by about a 6 dB/octave curve in frequency domain. Consequently, if either the glottal characteristic or the vocal tract characteristic is determined approximately, the other can be deduced. Here, we consider several methods of separating the glottal and vocal tract characteristics to clarify the time-varying characteristics of speech parameters.

[Method I]

In this method the vocal tract is regarded as being formed by extremely small changes from a uniform, smooth, lossless, semi-infinite acoustical pipe, and poles are considered to be distributed almost uniformly in the frequency domain, so that it is assumed that the general outline of frequency transfer characteristic is flat. In other words, it is assumed that the general outline of the spectral pattern of the speech signal represents the glottal characteristic including the radiation characteristic. Based on this assumption, the combined radiation and glottal characteristics are estimated in every 10 ms. By means of the inverse filter of that characteristic, the vocal

tract characteristic is extracted from the speech waveform.⁽⁷⁾

[Method II]

The averaged speech spectrum over a long time period may be considered to approximately represent the glottal characteristic, including the radiation characteristic, since the vocal tract characteristics for the various phonemes are averaged. In this method, the glottal characteristic with the radiation characteristic is estimated from the shape of the long-time averaged spectral pattern and, by means of the inverse filter of this characteristic, the vocal tract characteristic is extracted from the speech waveform. Spoken sentences, about 10 seconds long, are used as long-time speech samples, based on the results of the analysis of long-time averaged spectrum.⁽²⁾⁽³⁾

[Method III]

This method is intermediate between Methods I and II. The glottal characteristic with the radiation characteristic is estimated from the time-averaged spectral pattern of a spoken word. By means of the inverse filter of this characteristic, the vocal tract characteristic is extracted from the speech waveform.⁽⁶⁾

A blockdiagram of each of the three methods is shown in Fig. 1. In all cases, the glottal characteristic is approximated, for simplicity, as a 1st-order system or a 2nd-order critical damping (CD) system. The transfer function of the inverse filter can be expressed as follows:

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

where $Z = e^{-j\lambda}$ and λ = radian/unit time.

For the speech wave after the inverse filtering, PARCOR analysis is applied. The glottal spectral characteristic, including the radiation characteristic, is represented by a parameter β or γ , and the vocal tract characteristic is represented by PARCOR coefficients $\{K_m\}$. PARCOR coefficients can be interpreted as reflection coefficients of the vocal tract, so that vocal tract area function can be estimated from PARCOR coefficients. Additionally, fundamental frequency (f_0) is extracted from the speech wave after the inverse filtering. The same analyzing methods are applied to β , γ , $\{K_m\}$ and f_0 parameters.

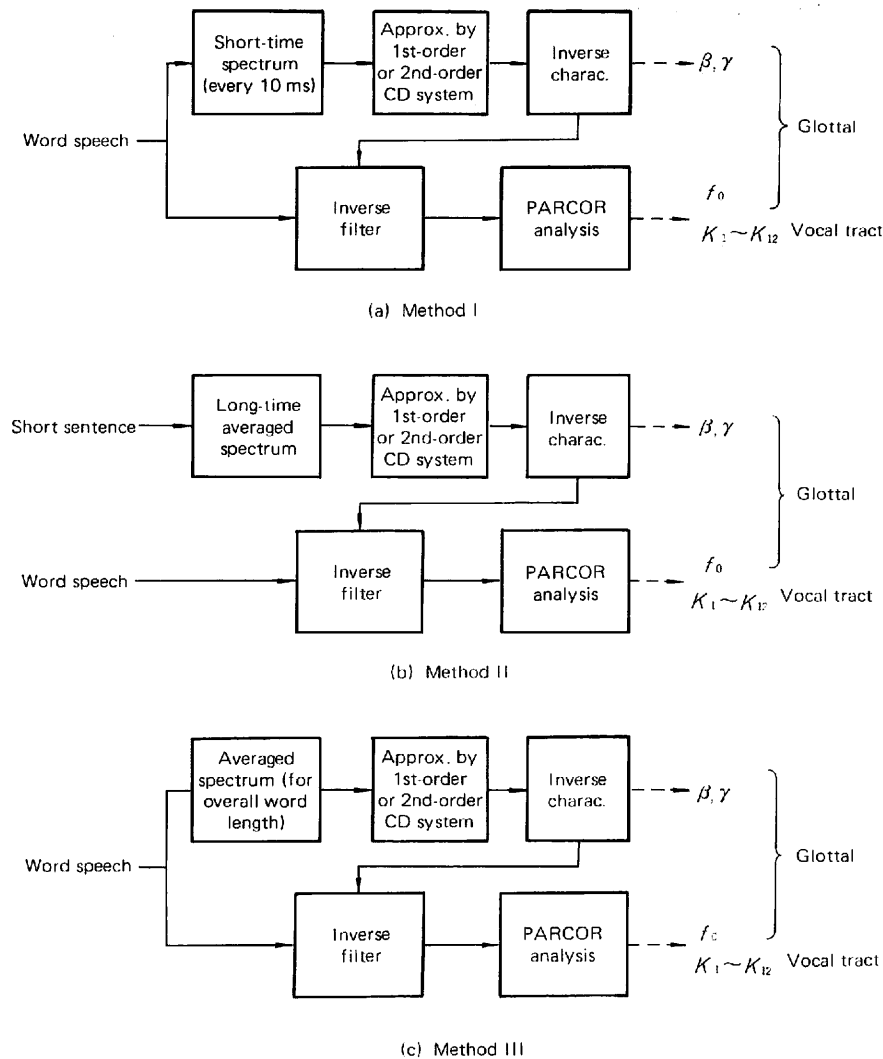


Fig. 1—Glottal and vocal tract characteristics separating methods.

2.2 Speech Samples

5 vowels and several words are uttered in an anechoic chamber. The speech 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 (frames) are extracted every 10 ms and multiplied by the Hamming window. The glottal and vocal tract characteristics in the speech wave are separated, and β , γ , f_0 and $\{K_m\}$ parameters are extracted. The average value over the whole voiced portion in the utterance is measured for each parameter. The following several analyses are applied to this value.

5 vowel samples of /i/, /e/, /a/, /o/ and /u/, uttered separately, and 2 word samples of /bango:/ and /namae/ are used in the experiment. They are spoken by 9 male talkers repeatedly at many sessions over a long period, ranging from about 10 months to 3 years.

3 Method of Evaluating Temporal Variation Characteristics

Let X_{ijk} denote the k th ($1 \leq k \leq 3$) sample, which is one of the 3 samples uttered in repetitions at session j ($1 \leq j \leq n$: 2 or 3 day to 3 month intervals) by talker i ($1 \leq i \leq 9$). This is either a parameter (scalar) of β , γ , $\{K_m\}$ or f_0 , or N -

dimensional vector composed of these values.

$$X_{ijk} = (X_{ijk1}, X_{ijk2}, \dots, X_{ijkN}). \quad (3)$$

The average value of the 3 samples at a given session is denoted as X_{ij} :

$$X_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijN}) = \sum_{k=1}^3 X_{ijk}/3 \quad (4)$$

We define the simultaneous variation parameter as the squared Euclidian distance between the 3 samples spoken at the same session, averaged over the n sessions, the 9 talkers and 3 times, which is the number of combination of 2 out of 3. $D_{0i}(l)$, the simultaneous variation parameter of l th component, and D_0 , the simultaneous variation parameter of the vector, are given as

$$D_{0i}(l) = E_{ijk} \{(X_{ijk1} - X_{ijk2})^2\} \\ = \sum_{i=1}^9 \sum_{j=1}^n \sum_{1 \leq k < k' \leq 3} (X_{ijk1} - X_{ijk'1})^2 / 27n \quad (5)$$

$$D_0 = \sum_{l=1}^N D_{0i}(l). \quad (6)$$

In the same manner, we determine the distance between samples separated by time interval τ , and call it the temporal variation parameter. In order to remove the effect of simultaneous variation, the temporal variation parameter is defined as the average of the distance between the average value of the 3 samples spoken at the same session:

$$D_i(l, \tau) = E_{ij} \{(X_{ij,t} - X_{ij',t})^2\} \quad (7)$$

$$D(\tau) = \sum_{l=1}^N D_i(l, \tau) \quad (8)$$

$$\tau = t(j) - t(j'),$$

where $t(j)$ is the real time corresponding to session j .

The temporal variation parameter, divided by the simultaneous variation parameter, is called the temporal variation ratio and is given as

$$V_i(l, \tau) = D_i(l, \tau) / D_{0i}(l) \quad (9)$$

$$V(\tau) = D(\tau) / D_0. \quad (10)$$

It has been experimentally determined that the temporal variation ratio is practically constant

over the range where τ is more than or equal to 6 months, independent of the kind of parameter. The long-term variation ratio is defined as the averaged value of the temporal variation ratio over this range. The long-term variation ratio of a component of X_{ijk} is denoted as $V_{il}(l)$, and that of the vector is denoted as V_L .

$$V_{il}(l) = E_{\tau \geq 6 \text{ months}} \{V_i(l, \tau)\} \quad (11)$$

$$V_L = E_{\tau \geq 6 \text{ months}} \{V(\tau)\}. \quad (12)$$

A parameter whose temporal variation ratio or long-term variation ratio is small can be called a stable parameter whose variation does not increase with the length of the collecting period. Parameters to be used for representation of the talker differences or for talker recognition should have small long-term variation ratio.

4 Measured Results for Temporal Variation Ratio and Long-Term Variation Ratio

For two cases of interest — one using an inverse filter to separate the glottal and vocal tract characteristics for the speech signal and the other without the filter — the temporal variation ratios and the long-term variation ratios have been studied.

4.1 Temporal Variation Ratio

Figure 2 shows temporal variation ratio $V(\tau)$ of the vector composed of the 1st to 12th PARCOR coefficients, extracted without inverse filtering, with the values being averaged over the 2 words

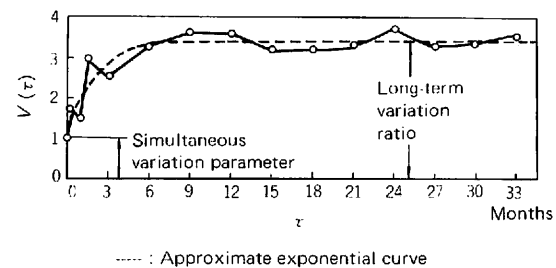


Fig.2—Temporal variation ratio of vector which consists of PARCOR coefficients averaged over voiced portion of spoken word.

and 9 talkers. When τ is less than 6 months, the ratio increases exponentially. When τ is more than 6 months, the ratio is nearly constant. The dashed line is an exponential curve fitted approximately to the result.

Figure 3 shows the temporal variation ratio of coefficient β of the 1st-order inverse filter (Method I), averaged over 2 words and 9 talkers. In comparison with the variation ratio of the total vector in Fig. 2, it can be seen that the variation ratio of parameter β , which represents the general shape of the glottal spectrum, is quite large.

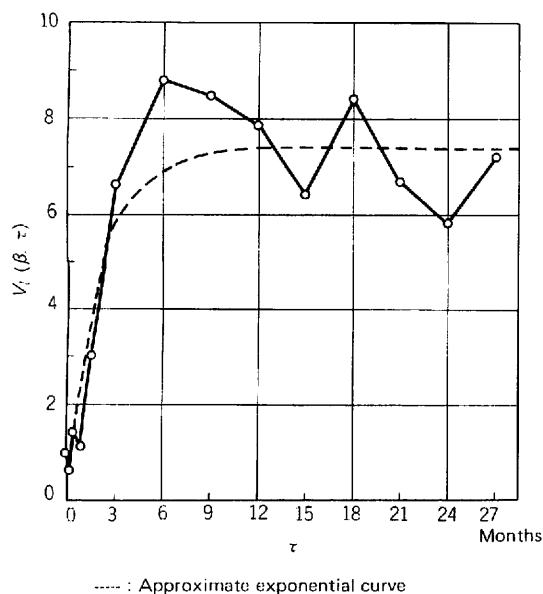


Fig.3—Temporal variation ratio of coefficient β , parameter of 1st-order inverse filter.

4.2 Long-Term Variation Ratio

Figure 4 shows the long-term variation ratio of the vector composed of the 1st to 12th PARCOR coefficients for each of the vowels. Results are given for the various types of filters, as well as for the case with no inverse filter. It can be seen from these results that, by removing the glottal characteristics and thus approximating the vocal tract characteristic alone, the long-term variation ratio is greatly reduced for all the vowels. The resultant long-term variation ratios are almost equal in all the vowels. This effect varies slightly with the type of inverse filter.

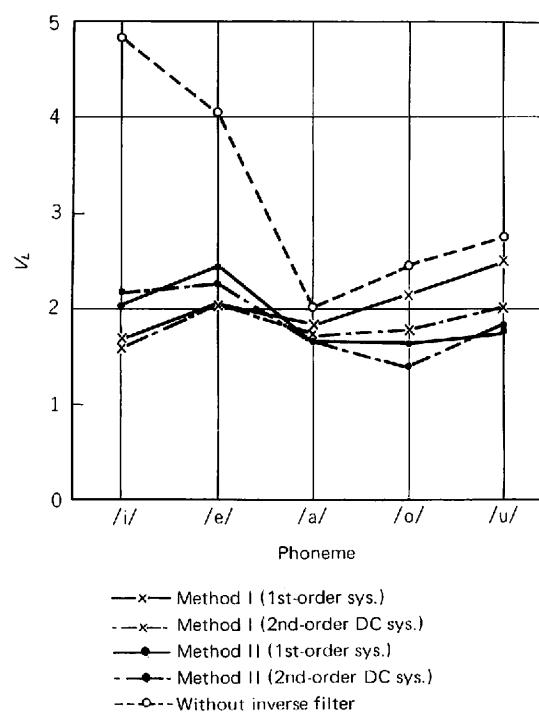


Fig.4—Long-term variation ratio of 5 vowels.

When the glottal characteristic is not removed, the long-term variation ratio is strongly dependent on the phoneme, and phonemes having large dispersion in formants or phonemes having the articulation point near the lips seem to have large long-term variation ratio. It might be considered that this is due to the fact that the contribution of the glottal spectrum is modified by the shape of the vocal tract, but this has not been firmly established.

Figure 5 shows the long-term variation ratio of the 1st to 12th PARCOR coefficients, with the values being averaged over the 2 words and 9 talkers. The results, with or without inverse filtering of Method II, are shown. It can be seen from these results that, with the inverse filtering the long-term variation ratio is reduced for all the coefficients, especially for 1st and 2nd ones.

Table 1 presents long-term variation ratios of all the parameters. Results for 5 vowels and 2 words are presented. In every case, the long-term variation ratios of parameters β , γ and f_0 , which are related to the glottal characteristic, are greater than those of parameters K_1 to K_{12} , which are related to the vocal tract characteristic. The results for the 5 vowels and 2 words are quite

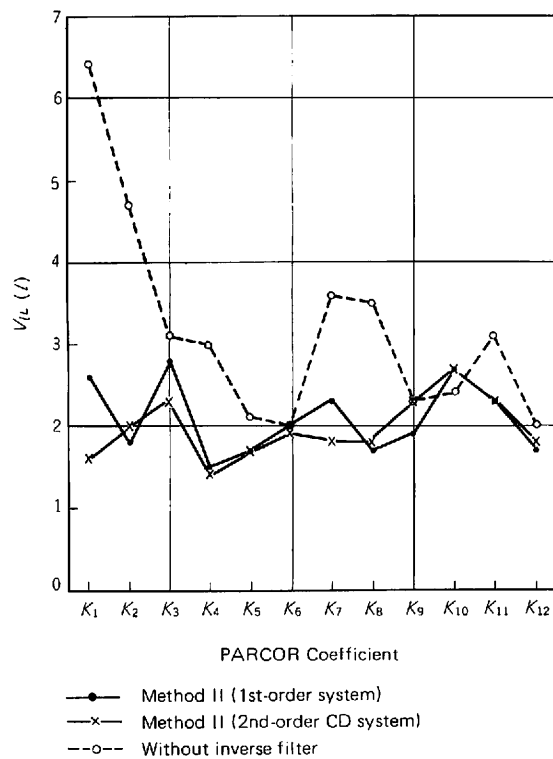


Fig.5—Long-term variation ratio of each PARCOR coefficient.

similar. These results indicate that the glottal characteristic changes more slowly and in a larger way than the vocal tract one with the time passage. In other words, the property wherein the amount

of variation increases with the length of the period, over which the speech samples are collected, is more evident for the glottal characteristic than the vocal tract one.

From the results shown in Fig. 3 to Fig. 5 and Table 1, it can be concluded that the temporal variation of the speech wave or speech spectrum is mostly due to variation in the glottal spectrum.

5 Analysis of Variance

5.1 Analysis of Variance for Each Parameter

An analysis of variance was conducted in regard to talker factors for each of the parameters, β , γ , f_0 and $\{K_m\}$, which were extracted from the speech samples of 2 words using the inverse filter of Method II. Two sets of samples, which cover about 10 days and 10 months, respectively, both taken at 4 different sessions, are analyzed and the effects of the length of the period are examined. Talker effects, divided by significance level F_0 at the probability of 1% of false rejection, are shown in Fig. 6. The results for glottal parameter β , γ and f_0 are presented in part (a), and those for vocal tract parameters, K_1 to K_{12} , obtained after passing through the inverse filter of 2nd-order CD system, are presented in part (b). From these results, we make the following observations. Glottal spectrum parameters β and γ show extremely strong talker effects over a short

Table 1 FEATURE PARAMETER LONG-TERM VARIATION RATIOS

Speech Sample	Inverse Filter	Parameter	Long-Term Variation Ratio	
			Method I	Method II
5 Vowels	1st-order system	β	5.7	27.8
		f_0	4.6	4.6
		$K_1 \sim K_{12}$	2.0	1.9
	2nd-order CD system	γ	4.2	24.8
		f_0	4.6	4.6
		$K_1 \sim K_{12}$	1.9	1.9
2 Words	1st-order system	β	6.9	27.8
		f_0	4.9	4.9
		$K_1 \sim K_{12}$	2.6	2.1
	2nd-order CD system	γ	—	24.8
		f_0	4.9	4.9
		$K_1 \sim K_{12}$	—	2.0

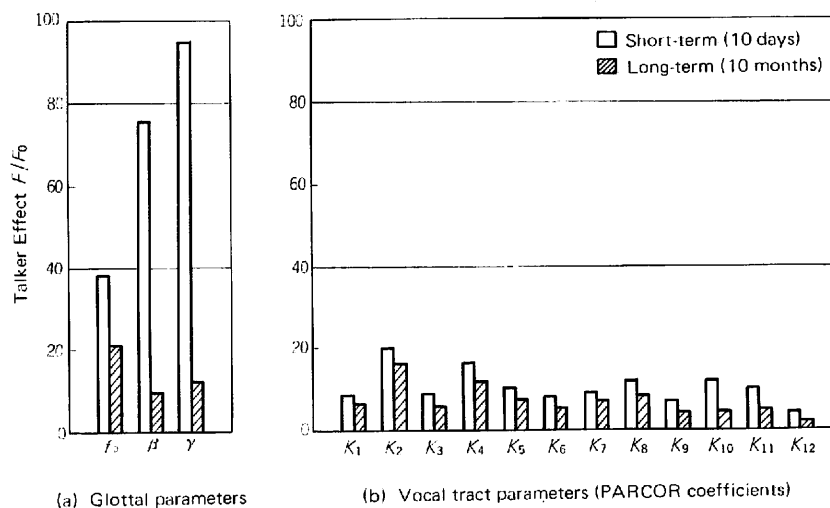


Fig.6— Relations between talker effect and period over which samples are taken.

period of only 10 days, but the effect decreases to about 1/8 the magnitude over a period of 10 months, affected by the increase in variation. On the other hand, vocal tract parameters of PARCOR coefficients K_1 to K_{12} do not show much of these characteristics and, for the long-term, show about the same magnitude of talker effect as β and γ .

5.2 Multivariate Analysis of Variance

A multivariate analysis of variance was performed with regard to phoneme factors, talker factors and phoneme-talker interaction for vectors composed of PARCOR coefficients K_1 to K_{10} , obtained with or without inverse filtering. 5 vowels spoken at 12 sessions over a period of about 3 years by 9 male subjects are used. Figure 7 shows the magnitude of each effect divided by the significance level at the probability of 1% of false rejection. From these results it can be seen that, for long-term samples, both the phoneme effect and the talker effect show little change, even if the glottal characteristic is eliminated.

6 Recognition Tests

From the results discussed above, it has been revealed that, when the radiation and glottal characteristics are approximately removed from

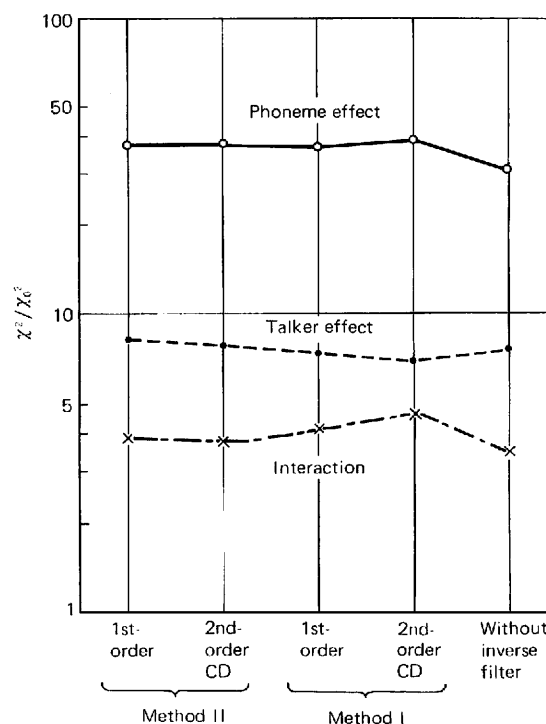


Fig.7— Results of multivariate analysis variance for vector which consists of PARCOR coefficients.

speech wave by the inverse filtering, and only the vocal tract characteristic is extracted, the temporal variation of the feature parameters is reduced but the personal information in the parameters is scarcely reduced. This effect is ascertained by talker recognition experiments in this section. In the usual talker recognition, when the reference samples are taken only at one session or over a short period of about 10 days for each registered talker, the recognition accuracy extremely drops for unknown samples uttered after a long time interval of 3 months or more, as mentioned in Section 1.

In this section, the effect of eliminating the glottal characteristic on the recognition accuracy is investigated by means of a talker recognition system which uses the statistical parameters (average, standard deviation, correlation matrix) of the fundamental frequency and the 1st to 12th PARCOR coefficients in the voiced portion of spoken words.⁽⁴⁾⁽⁵⁾ From the components of the standard deviation and the correlation matrix, 20 components are selected, based on the analysis of variance using the reference samples, and are used together with the average values as components of a feature vector. Two words, /bango:/ and /namae/, are used, and recognition tests are conducted by means of the 1-word method and the 2-word combination method.

The period over which reference samples for registered talkers are collected and the interval between the last reference sample and unknown samples are varied in a series of experiments. These effects on recognition accuracy are investigated.

6.1 Recognition with Short-Term Reference Samples 1 —Glottal Characteristic Elimination Effect—

As shown in Fig. 8, reference samples were collected at 3 or 4 sessions at intervals of 2 or 3 days. Following the last reference samples, recognition tests were made 2 or 3 days later, 3 weeks later, 6 weeks later, 3 months later and 10 months later. The results of the talker verification (with 8 impostors), with or without the inverse filtering of methods I to III, are presented in Fig. 9 to Fig. 11.

From these results, it is seen that, by eliminating the glottal characteristic, which has a strong

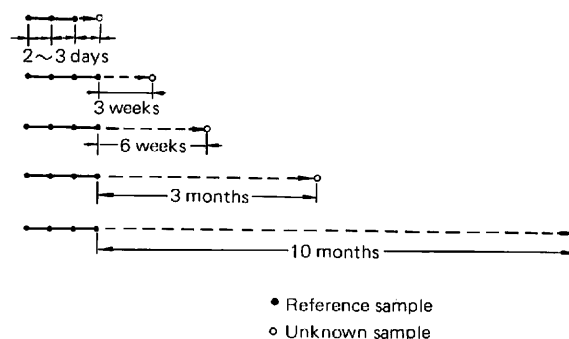


Fig.8—Relations between reference and unknown samples in recognition experiment using short-term reference samples.

and gradual temporal variation trend, the recognition errors can be reduced by about 1/2. This means that the temporal variation of feature parameters is greatly reduced by the above-mentioned method. Comparing the results for Methods I, II and III, the differences are small, but Methods II and III give relatively good results, compared to Method I.

6.2 Recognition Using Long-Term Reference Samples

As shown in Fig. 12, reference samples were taken at 4 sessions separated by intervals of 3 months. The recognition tests were made on unknown samples taken 3 months after the last reference sample. The results are presented in Fig. 13. As can be seen from this figure, there is practically no difference between the recognition rate with the inverse filter and the rate without the filter. It may be considered that this is due to the fact that the variation in parameters over a long term of about 10 months is essentially random, and that, over the long term, the talker effect (personal information) in the glottal spectrum is quite small. This result is consistent with the results of the analysis of variance presented in Fig. 7.

6.3 Comparison of Results with Long-Term and Short-Term Reference Samples

As shown in Fig. 14, using unknown samples

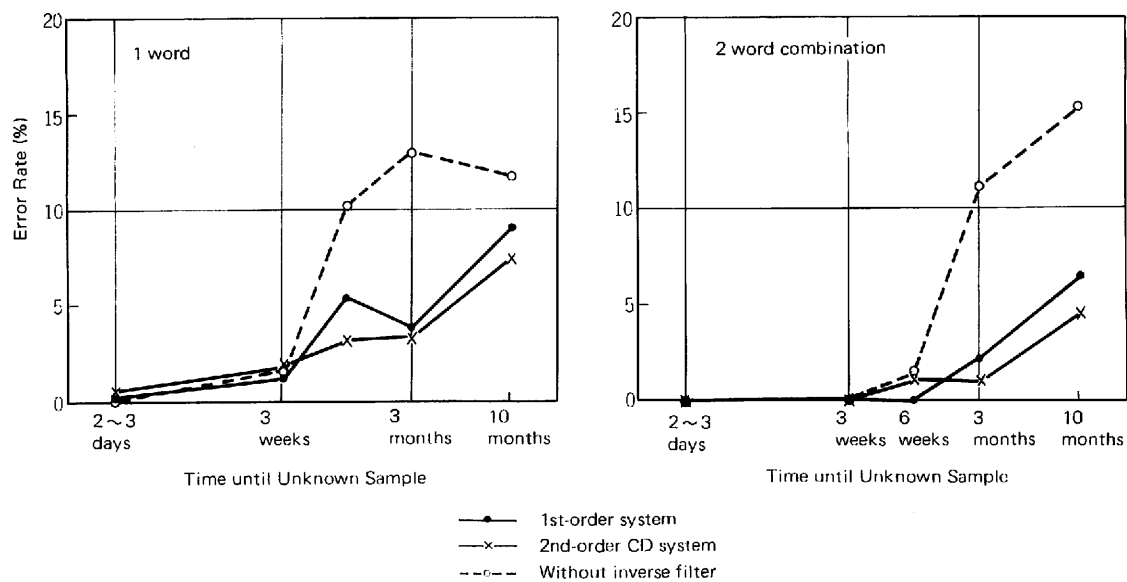


Fig.9—Results of talker verification using short-term reference samples.
(Effects of inverse filter of Method I).

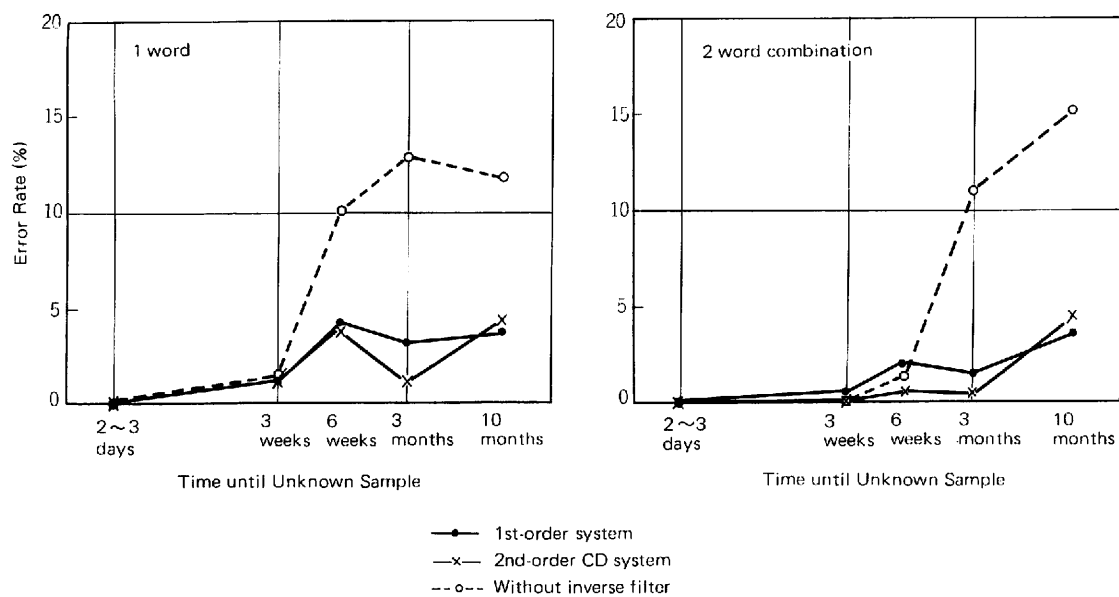


Fig.10—Results of talker verification using short-term reference samples.
(Effects of inverse filter of Method II).

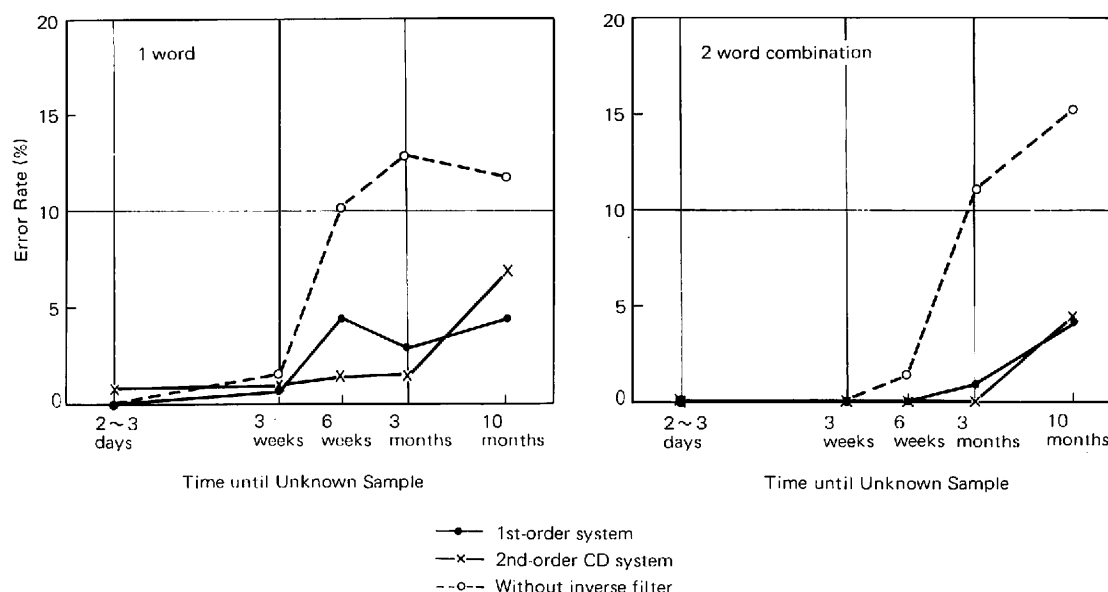


Fig.11—Results of talker verification using short-term reference samples.
(Effects of inverse filter of Method III).

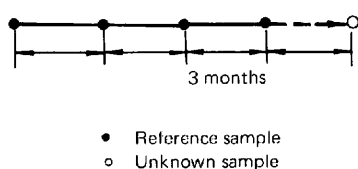


Fig.12—Relations between reference samples and unknown samples in recognition experiment using long-term reference samples.

(A, , F) of the same 6 sessions, a comparison was 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 covered a long period of about 10 months, long-term case. The inverse filter of Method II was used. In the short-term case, the reference samples were fixed and the time interval between reference and unknown samples were varied from 3 months to 21 months. In the long-term case, the time interval between reference and unknown samples was fixed at 3 months and the reference samples were varied. It has already been found, in the recognition experiment without inverse filtering, that, even if the time interval between

reference and unknown samples is varied over the range of 3 to 12 months, there is practically no change in recognition rate when the reference samples are collected over about 10 months. ⁽⁴⁾⁽⁵⁾

From the recognition tests, it was found that, in short-term case, recognition errors do not increase monotonically as the time until the unknown samples is varied from 3 months to 21 months. Therefore, recognition results are averaged over unknown samples of 6 sessions, both in short-term and long-term case, and presented in Table 2. The results show that, even if short-term reference samples are used, recognition rates of about 99% can be achieved (see the underlined part of Table 2), using 2-word combinations with the glottal characteristic elimination technique. These recognition rates are about equal to the performance with long-term reference samples. It is ascertained that the inverse filtering method reported in this paper is highly effective to reduce the influence of the temporal variation of the parameters on the talker recognition and to shorten the period over which reference samples should be collected.

6.4 Recognition with Short-Term Reference Samples 2

—Reference Samples Number Effect—

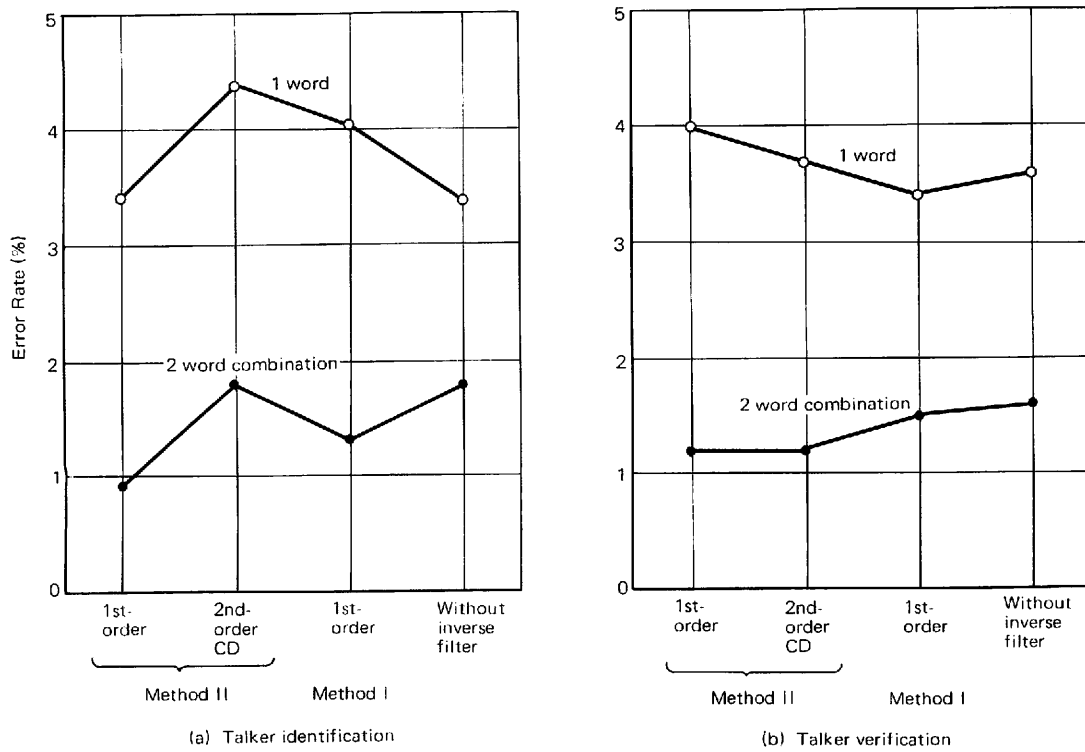


Fig.13—Results of talker identification and verification using long-term reference samples.

Table 2 COMPARISON OF RESULTS WITH SHORT-TERM AND LONG-TERM REFERENCE SAMPLES (INVERSE FILTER: METHOD II)

Reference Samples	Time Until Unknown	Inverse Filter	Taker Identification (9 indiv.)		Talker Verification (8 impostors)	
			1 Word	2 Words	1 Word	2 Words
Short-Term (10 days)	3 ~ 21 months	1st-order sys.	90.8 %	<u>99.4</u> %	97.4 %	<u>99.0</u> %
		2nd-order CD sys.	92.3	<u>99.4</u>	97.2	<u>98.9</u>
		None	88.3	94.5	92.7	92.8
Long-Term (10 months)	3 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

In the tests discussed above, the short-term reference samples were taken at 4 sessions with 2 or 3 day intervals over a period of about 10 days. The following supplemental tests were performed to determine at how many sessions the reference samples must be taken to achieve a

good recognition rate when the unknown samples are taken after a long time. The time separation between the reference samples and unknown samples was varied between 3 months and 21 months, as shown in Fig. 14(a). The number of sessions at which the reference samples are taken

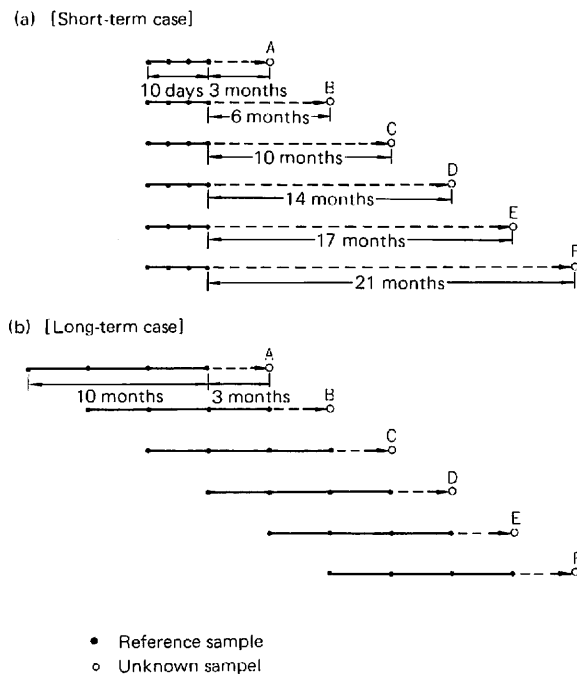


Fig.14—Relations between reference and unknown samples in recognition experiment about the effects of period over which reference samples are taken.

was varied from 1 to 4 with 2 or 3 day intervals. The filter of Method II (2nd-order CD system) is used in the recognition tests.

Results are shown in Fig. 15. Although the error rate is extremely high, when the reference samples are taken only at one session, it becomes lower with the increase in the number of sessions and is almost saturated at about 3 or 4 sessions. These results show that, even in the short-term case, it is both necessary and sufficient to use about 4 reference sample sessions.

6.5 Adding Inverse Filter Coefficients β and γ to the Feature Vector

Recognition tests were performed with parameters β and γ , which estimate the shape of the glottal spectrum, added to the feature vector. It was found that the effect is very small. With short-term reference samples, there is even a slight loss in recognition rate and, with long-term reference samples, there is little change in recognition rate. This may be considered to be mainly due to the very small talker effect and

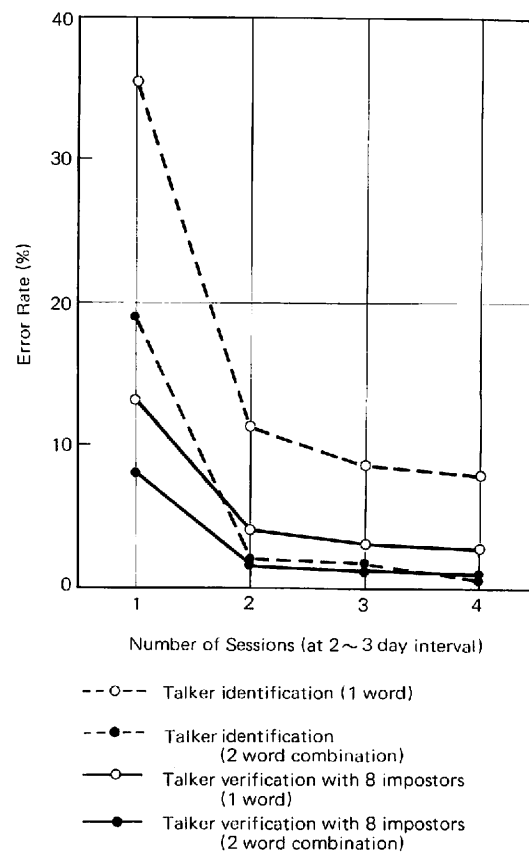


Fig.15—Relations between the number of sessions of short-term reference samples and error rate. (Inverse filtering by 2nd-order CD system of Method II).

short personal information of β and γ for a long period, as having been made clear previously in the analyses of variance and the recognition experiments.

6.6 Discussion

From the several talker recognition experiments, it was ascertained that the inverse filtering to approximately remove the radiation and glottal spectral characteristics in speech wave is highly effective to reduce the temporal variation of the feature parameters of speech and to extract the stable personal information.

Among the 3 methods of the inverse filtering, Methods II and III give relatively good results, compared to Method I, and differences between the results of the 2 methods are very small. In

Method I, the general shape of the vocal tract frequency transmission characteristic is assumed to always be flat, but this assumption is not always adequately satisfied in the narrow bandwidth of about 0–3.4 kHz, as considered in this study. Because of the impact of the formants, the glottal characteristic determined by this method varies quite widely from phoneme to phoneme or from frame to frame. This defect may affect talker recognition accuracy.

There is no such defect in Method II. In the sense of approximate glottal and vocal tract characteristics separation, this method may be the most reasonable, but it involves a problem wherein a long speech sample is necessary in addition to the sample words to be analyzed. This problem may be an obstacle to its application to the talker recognition system. From this point of view, Method III can be said to be the most practical and best method. In Method III, the inverse filter can be constructed by the averaged spectrum of the sample word to be used in recognition, and the same result as Method II can be obtained. The resemblance of the results of Methods II and III may be due to the fact that only the general features of the glottal spectrum are considered.

7 Conclusion

Exciting source (glottal) and vocal tract characteristics were extracted separately from the speech wave based on the mechanism of the vocal system, and temporal variation of the parameters was analyzed, from the viewpoint of the personal information in the speech wave. "Temporal variation ratio" and "long-term variation ratio" were defined and measured in order to evaluate quantitatively the amount of temporal variation in talker dependent parameters. It has been found that the glottal characteristic shows a stronger long-term variation trend than the vocal tract characteristic. An analysis of variance was then made for the talker factor. It was found that, although the talker effect of the glottal characteristic is quite large for samples over a short period, it becomes very small over a long period, because of the increase in variation.

Based on these results, talker recognition experiments were conducted. It was ascertained that inverse filtering, to approximately remove the radiation and glottal characteristics, is very ef-

fective to extract from short-term speech samples reliable personal information which is stable over a long period. It was found that a recognition accuracy of about 99% could be achieved in talker identification or verification, using reference samples collected over a period of about 10 days and unknown samples taken 3 to 21 months later, when 2 words were used in combination.

The above results show that the temporal variation of the spectral pattern mainly comes out about its broad outline and the stable feature is maintained in the relatively detailed structure, in agreement with the results obtained in the study of the long-time averaged speech spectrum.⁽²⁾⁽³⁾ Human listeners can recognize talkers by voice after a long time. Therefore, it is difficult to say that parameters which represent individual differences only over a short period and exhibit large changes over a long period truly represent speech individuality. That is, individual differences in the glottal spectral characteristic are said to be unessential. In listening tests, even with glottal individuality removed, high talker recognition rates have been reported.⁽⁹⁾

In order to determine whether or not the method used in this paper is appropriate to separating glottal and vocal tract characteristics, the vocal tract area function was determined from the PARCOR coefficients, which were obtained in Method II after elimination of the glottal and radiation characteristics. The result was compared with the observations of vocal tract shape during speech using X-ray photography by Chiba, Kajiyama⁽¹⁰⁾ and Fant,⁽¹¹⁾ and it was found that the results were quite similar. Details will be presented at some other time. Further studies are planned in regard to methods of separating the glottal information and the vocal tract information and on individuality analysis.

Acknowledgement

The author wishes especially to acknowledge the guidance provided by Dr. Shuzo Saito, Director of the Saito Research Section, and other members of the section.

References

- (1) S. Saito: Talker Recognition, Mathematical Sciences, No. 116, p. 49, 1973.
- (2) S. Furui, F. Itakura and S. Saito: Talker Re-

- cognition by the Long-Time Averaged Speech Spectrum, *Trans. IECE Jap.*, 55-A, 10, p. 550, 1972.
- (3) 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.
 - (4) S. Furui and F. Itakura: Talker Recognition by Statistical Features of Speech Sounds, *Trans. IECE Jap.*, 56-A, 11, p. 717, 1973.
 - (5) 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.
 - (6) 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.
 - (7) T. Nakajima, et al.: Estimation of Vocal Tract Area Functions by Adaptive Inverse Filtering Methods, *Bull. Electrotech. Lab.*, 37, 4, 1973.
 - (8) F. Itakura: Communication by the Feature Parameters of Speech, *Joint Conv. Four Elec. Inst.*, 219, 1972.
 - (9) R. O. Coleman: Speaker Identification in the Absence of Inter-Subject Differences in Glottal Source Characteristics, *J. Acoust. Soc. Amer.*, 53, 6, p. 1741, 1973.
 - (10) T. Chiba and M. Kajiyama: The Vowel, Its Nature and Structure, 1942, Tokyo-Kaiseikan Pub. Co., Ltd.
 - (11) G. Fant: Acoustic Theory of Speech Production, 1960, Mouton and Co.

* * * *