

論文 / 著書情報
Article / Book Information

Title	Personal Information in the Long-Time Averaged Speech Spectrum
著者	古井 貞熙
Author	Sadaoki Furui, Fumitada Itakura, Shuzo Saito
掲載誌/書名	, Vol. 23, No. Nos. 9-10, September-October, pp. 1133-1141
Journal/Book name	Review of the Electrical Communication Laboratories, Vol. 23, No. Nos. 9-10, September-October, pp. 1133-1141
発行日 / Issue date	1975, 9

Personal Information in the Long-Time Averaged Speech Spectrum

Sadaoki FURUI, Fumitada ITAKURA and Shuzo SAITO

The long-time averaged speech spectrum of a short sentence is used for a feature parameter representing the characteristics of an individual talker. Variation in the spectrum pattern was investigated for a number of talkers during more than a year and a half and talker recognition experiments were performed based on the spectrum patterns. Results of the study indicate that, although variations are observed in patterns for an individual talker, it is possible to recognize the talker by defining a suitable measure of the multi-dimensional distance, even if the time interval between the known samples and the unknown sample is 3 months or longer.

1 Introduction

Many investigations have been made of feature parameters representing the personal characteristics of speech.⁽¹⁾ One of the most difficult problems in machine recognition of talkers is that the feature parameters frequently vary after a long time interval, even if the same sentence is pronounced by the same talker.⁽²⁾⁽³⁾ Very few investigations have been made on the effect of the time interval between known samples and unknown samples on the recognition of a talker, and no previous report on recognition tests after a long time interval of the order of 3 months is known.⁽⁴⁾⁽⁵⁾

This paper considers the long-time averaged speech spectrum⁽⁶⁾⁻⁽⁸⁾ of a short sentence as a context-free parameter representing the characteristics of the talker, and presents the results of experiments paying attention to the long-term variation of this spectrum pattern. More precisely, the variation of the long-time averaged spectrum pattern is examined for a number of talkers during more than a year and a half. Various properties of this pattern variation are stated, with application to recognition of the talker. The relation between the time interval and the recognition rate is examined for various

intervals, including those of more than 3 months between the known samples and the unknown samples.

2 Long-time Averaged Spectrum Measurement

2.1 Measurement Method

As is shown in Fig. 1, a speech signal uttered in a shielded room is passed through a 3.4 kHz low-pass filter and is converted into a discrete sequence by 8-kHz sampling and 11-bit amplitude quantization procedures. This sequence is divided into blocks of 256 samples (corresponding to 32 ms) and the short-time power spectrum is computed in every block. In order to avoid the distortion produced by the block division, the Hamming window is multiplied in the time domain. The short-time power spectra obtained in all the 32-ms intervals of the speech are then 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.⁽⁹⁾ As is easily seen from this definition, lower elements of the cepstrum

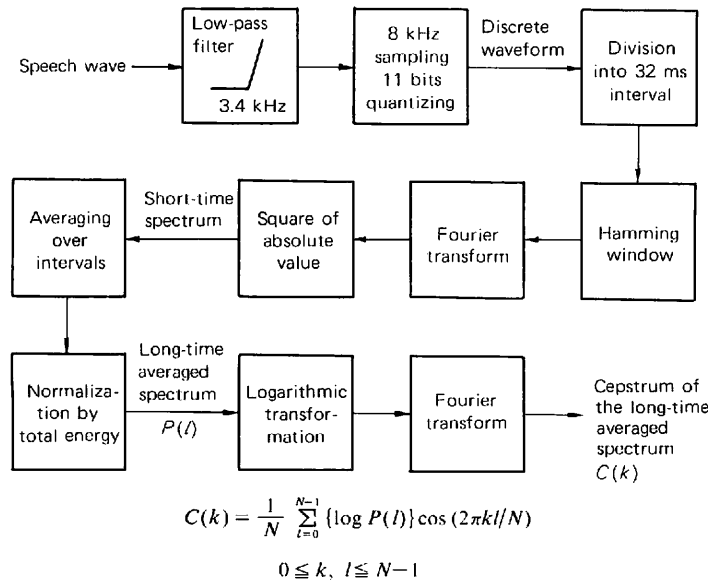


Fig. 1—Measuring system blockdiagram.

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. Since the cepstrum is obtained by a linear orthogonal transformation from the logarithmic spectrum, we see that distance in the cepstrum domain is proportional to distance in the logarithmic spectrum domain.

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. The measurements are usually made for 9 males with intervals of several days to three months.

2.2 Measurement Results

Figure 2 shows examples of measurements of the long-time averaged (logarithmic) spectrum. Part (a) of the figure shows the average pattern of samples from talkers *F*, *S*, *K* and *I* over about half a year. Part (b) shows the long-term variation of the pattern for talker *S*.

3 Recognition of the Talker

3.1 Recognition Model

A recognition model is constructed for the cepstrum of the long-time averaged spectrum, taking its long-term variation into consideration, and experiments are performed on the identification and verification of the talker. Here, identification of the talker means a process whereby it is determined from which of the registered *m* 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 following distances, (i) and (ii), between unknown sample and the known samples of the registered talker: ⁽¹⁰⁾

(i) The weighted distance to the mean value of known samples:

$$S(C, \{D_{r,j}\}) = |V_r|^{1/N} (C - \bar{D}_r) V_r^{-1} (C - \bar{D}_r)^t \quad (1)$$

where *C* is the cepstrum of unknown sample (*N*-dimensional vector), *D_{r,j}* is the cepstrum of *j*th sample of registered talker *r*, $\bar{D}_r = 1/n \sum_{j=1}^n D_{r,j}$ is the mean value of known samples, *V_r* is the covariance matrix of the cepstrum of *r*th talker (an *N* × *N* matrix) and *n* is the number of known samples for each registered talker.

In order for Eq. (1) to be well-defined, *V_r*

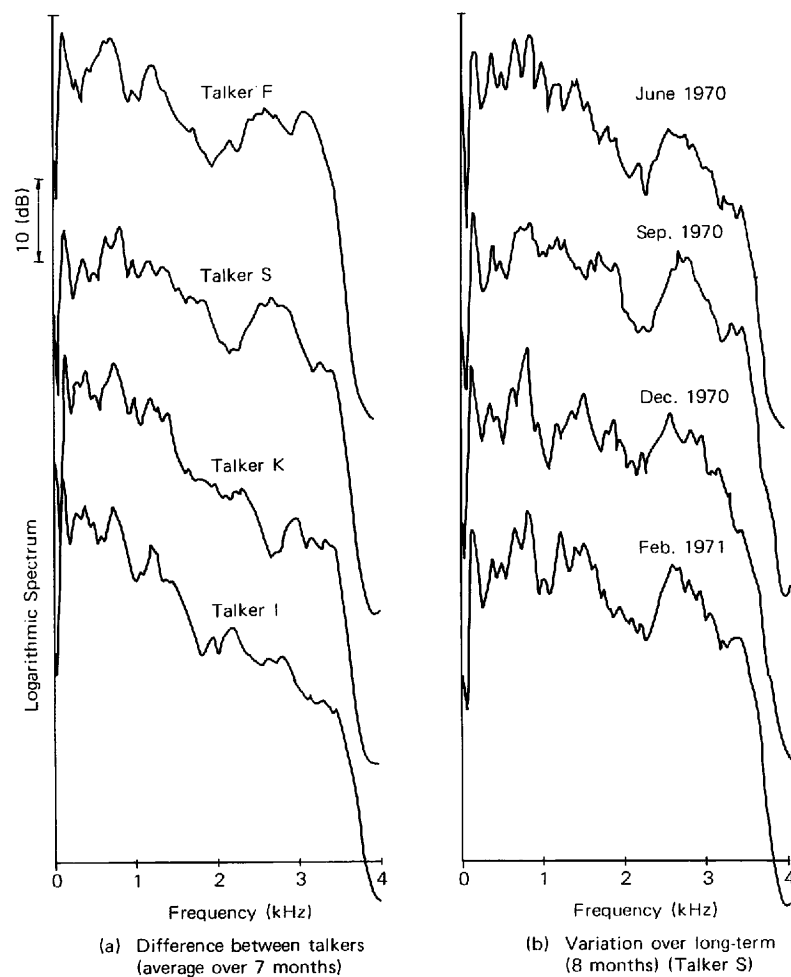


Fig. 2—Examples of long-time averaged spectra.

must be positive definite. This requires that $N \leq n$, where N is the dimension of the characteristic vector and n is the number of known samples. In practical situations, however, this condition is difficult to satisfy, since N is relatively large, while it is difficult to obtain a large number of speech samples. Even if this condition is satisfied, a large error will be included if the number of known samples is not sufficiently large. In order to avoid this, one of the following matrices is used, instead of V_r , on the assumption that the covariance matrices of the talkers are similar.

[Method I]

$$V_I = (V_{Iij}) = \frac{1}{m} \sum_{r=1}^m V_r \quad (2)$$

where m is the number of registered talkers.

[Method II]

$$V_{II} = (V_{IIij}) = \begin{pmatrix} V_{I11} & & 0 \\ & V_{I22} & \\ 0 & & V_{I NN} \end{pmatrix} \quad (3)$$

where V_{Iii} is the diagonal element of the covariance matrix. To replace V_I by V_{II} , the correlation between cepstrum elements must be sufficiently small.

[Method III]

$$V_{III} = (V_{IIIij}) = \begin{pmatrix} V_{Ia} & 0 \\ 0 & V_{IIa} \end{pmatrix} \quad (4)$$

$$V_{Ia} = (V_{Iij}) : 1 \leq i, j \leq M, \quad (M \times M \text{ matrix})$$

$$V_{ii} = \begin{pmatrix} V_{ii} \end{pmatrix}: M+1 \leq i, j \leq N, \\ ((N-M) \times (N-M) \text{ matrix})$$

(ii) The square of the (unweighted) Euclidian distance to the mean value of the known samples [Method IV]:

$$S(C, \{D_{rj}\}) = \|C - \bar{D}_r\|^2 \quad (5)$$

In the case of talker identification, the talker whose unknown sample is closest to a particular standard pattern (known sample) 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.

3.2 Specification of Known and Unknown Samples

In order to examine the effect of the amount of time over which the known (reference) samples are collected, the number of the known samples and the time interval between the known samples and the unknown samples on the recognition error rate, the known and unknown samples are used as shown in Fig. 3 in four kinds of recognition experiments.

As the known samples for each of the registered talkers, samples spoken three times for

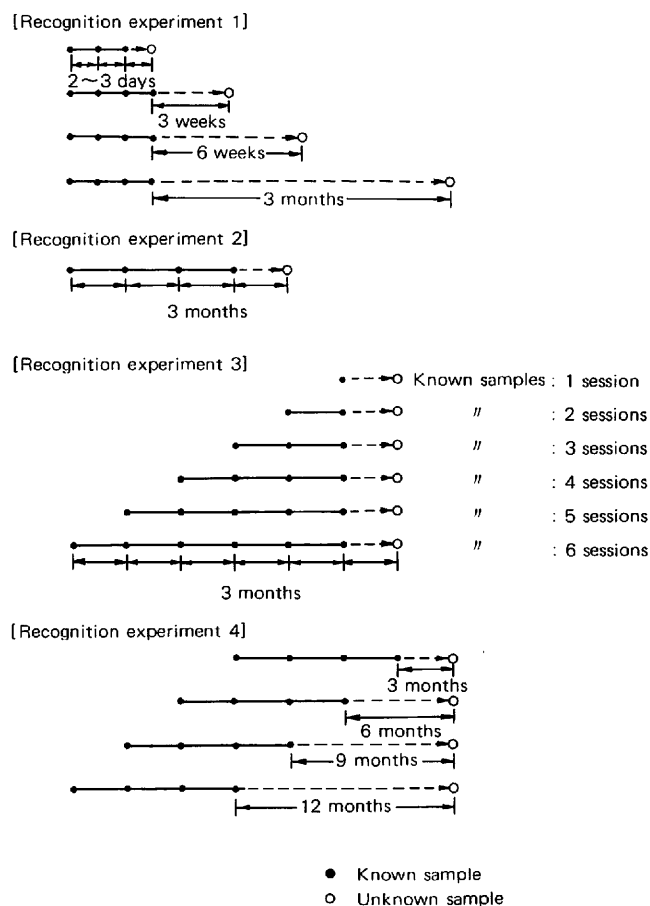


Fig.3—Relations between reference samples and unknown samples in four recognition experiments.

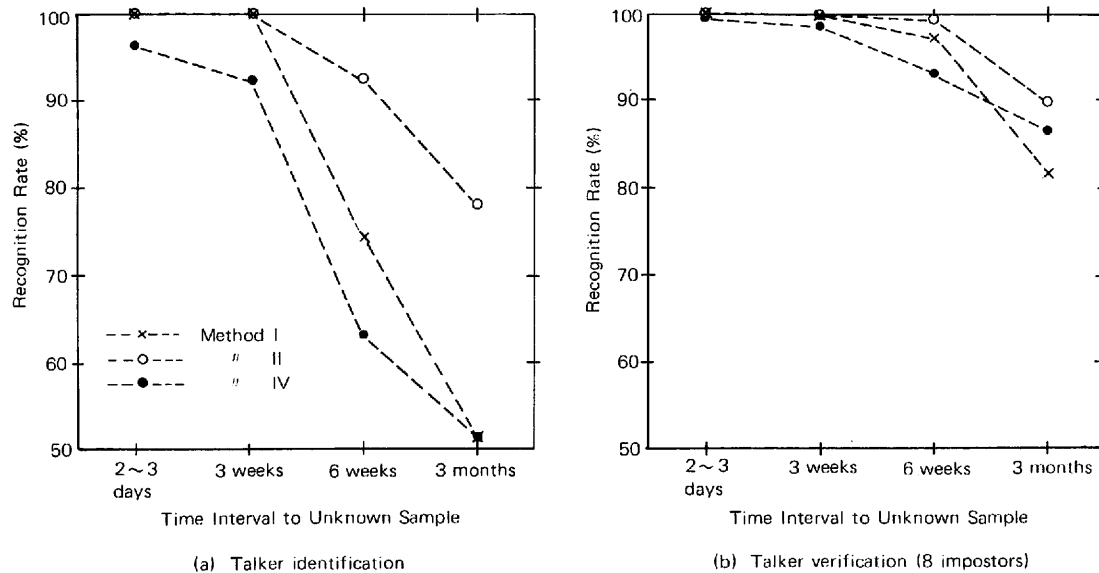


Fig 4—Identification and verification accuracy of 1st recognition experiment.

each recording session are used. In the case of talker identification, one out of nine registered talkers is determined. 27 samples consisting of 3 samples for each of the 9 talkers are used as the unknown samples. In the case of talker verification, for each of the 9 talkers, 3 samples spoken by that talker are used as the samples to be accepted, and 24 (or 78) samples consisting of 3 samples pronounced by each of the 8 other registered talkers (or 26 unregistered talkers) are used as samples to be rejected.

As seen from the description of the method in Section 2.1 and from the even property of the power spectrum, the cepstrum obtained has effectively 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 recognition is set at 64 (31 for Method I) for talker identification, and 41 (21 for Method I) for talker verification, neglecting the higher components.

4 Recognition Results

4.1 Recognition Experiment 1 —Recognition Using Known Samples Collected over a Short Period —

Figure 4 shows the average rates of recognition for the 9 talkers as functions of the time interval between the known samples and the unknown samples. From these results of recognition rate for each talker, analysis of variance is performed with regard to the talkers, recognition method (I, II and IV), interval between known and unknown samples, and their interactions. The results reveal that, both for identification and verification of the talker, the interval and the interaction between the interval and the talkers are significant with an error probability of 0.5%, and the talker and the recognition method are significant with an error probability of 1%.

From these results it is seen that, when the known samples are of short observation, for example 10 days, the recognition rate is high if the time interval from known to unknown samples is short, for example from several days to three weeks, while it becomes low with increase of the interval and extremely so when the interval is of the order of three months. It is also

seen from the results of the analysis of variance that the stability of the long-time averaged spectrum pattern depends on the talkers.

4.2 Recognition Experiment 2

— Recognition Using Known Samples
Collected over a Long Period —

Table 1 shows the results of talker identification and talker verification. The unknown samples are the same as those in recognition experiment 1, and the time interval between the known and the unknown samples is 3 months. It is seen that, by using the long-term data as the known samples and by setting the weighted distance with consideration of its variation, a better recognition rate is achieved, even if the interval from the last known sample to the unknown sample is as long as 3 months.

The results of Table 1 indicate that the definition of distance according to Method III gives a relatively good recognition rate, but the difference between Methods I, II and III is small.

4.3 Recognition Experiment 3

— Effects of the Number of
Known Samples —

Figure 5 shows the average recognition rate for 9 talkers for different distance measures. Figure 5 (a) shows the results of talker identification, (b) the results of talker verification with 8 false talkers (impostors), and (c) the results of talker verification with 26 impostors.

The results of Fig. 5 indicate that, for both talker identification and verification, the recognition rate is as low as comparable to the results of experiment 1 with the time interval

Table 1 RESULTS OF RECOGNITION EXPERIMENT 2

	Method I	Method II	Method III	Method IV
Identification	88.0%	89.8%	90.8%	76.0%
Verification (8 impostors)	94.5	95.4	95.5	93.1
Verification (26 impostors)	93.4	92.3	93.5	89.3

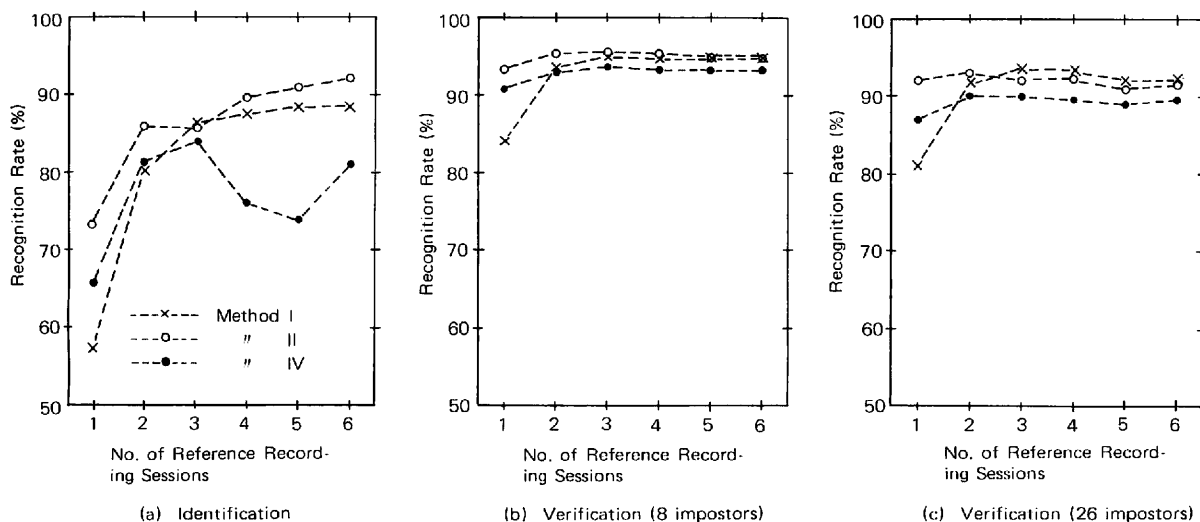


Fig. 5—Identification and verification accuracy of 3rd recognition experiment.

of 3 months, if known samples are taken only at one time. It is necessary to take samples for a certain long period in order to achieve a high recognition rate.

By increasing the number of impostors from 8 to 26, the recognition rate is seen to decrease by 1 to 3%. The results of statistical-hypothesis testing of talker verification by Methods I and II have revealed that the hypothesis, that the difference is insignificant between the two methods, is not rejected at the significance level of 5%. Accordingly, further study seems necessary on the effect of the number of impostors on the recognition rate.

4.4 Recognition Experiment 4 – – Recognition of the Unknown Sample after a Long Interval –

The results of this experiment are shown in Fig. 6. When the time interval between the last known sample and the unknown sample is made longer, from 3 months to 12 months, the rate of recognition decreases (solid points) unless one uses a weighted distance, while no such tendency is observed if one uses Methods I and II. Consequently, the results in recognition experiment 3, which are obtained by using weighted distances and known samples taken 4 or more times, can be considered effective, even if one makes the interval to the unknown sample longer,

from 3 months to 12 months.

5 Variation of Long-time Averaged Spectrum Pattern

In order to examine the time-varying characteristics of the cepstrum elements of the long-time averaged spectrum, their variances over short and long intervals are determined. For short-term variation, sample spectra taken four times with intervals of 2 to 3 days are used. For long-term variation, sample spectra taken six times at intervals of 3 months are used. Figure 7 shows the variances for each of two talkers and the pattern averaged over 9 talkers.

These results indicate that, as in the case of talker W, there are many talkers who have significant peaks nearly at the same elements in short and long-term variances. Also, common to all talkers, the variances of lower elements are greater for longer intervals. This indicates that for longer intervals, variation of the general indication of the long-time averaged spectrum pattern becomes more significant than the variation of its fine structure.

6 Conclusion

The results of experiments reported in this paper indicated that the long-time averaged spectrum pattern is quite stable for periods of several

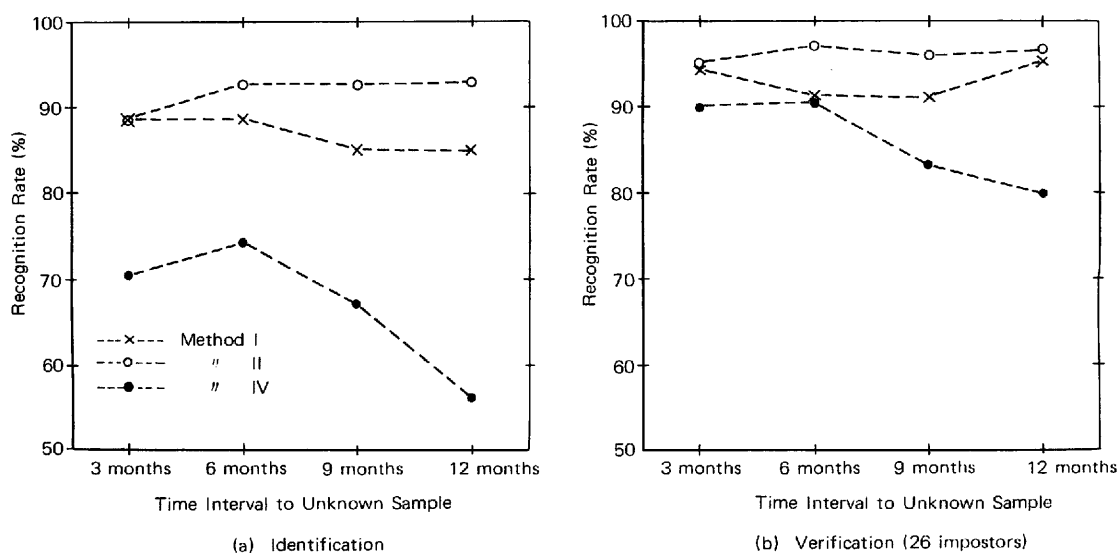


Fig. 6—Identification and verification accuracy of 4th recognition experiment.

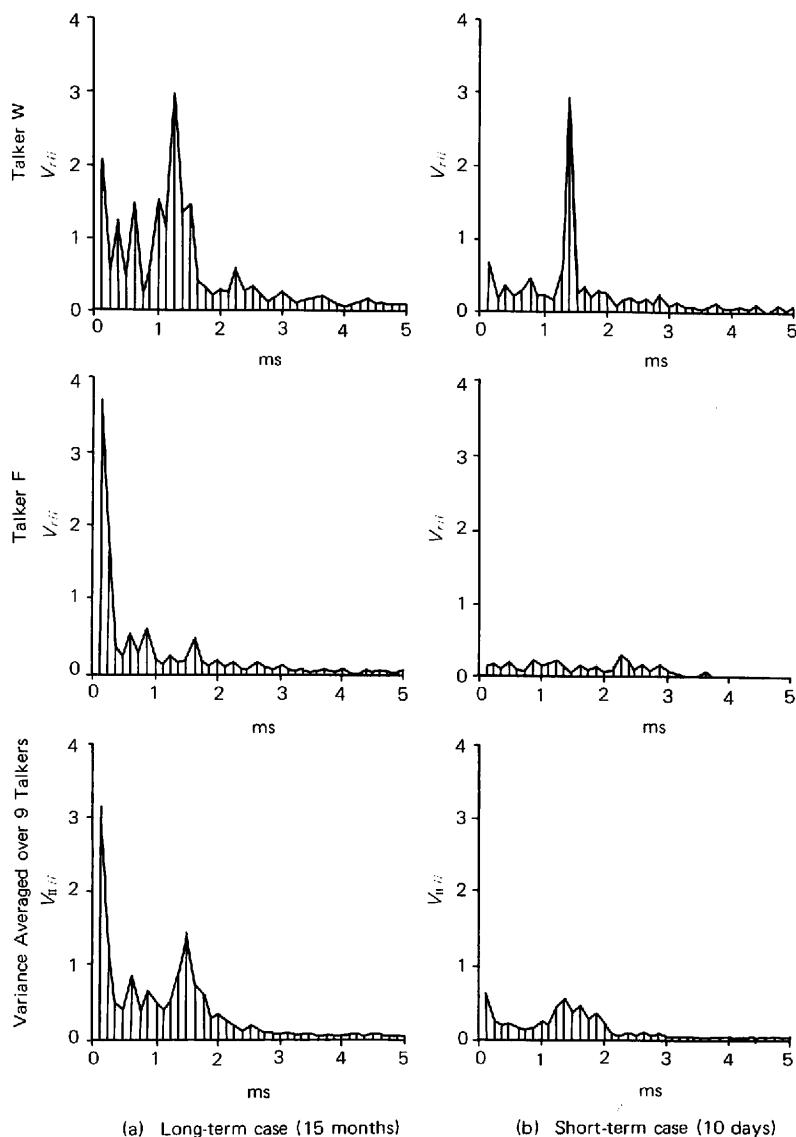


Fig.7—Variance of the cepstrum for the long-time averaged spectrum.

days to three weeks, while it exhibits a gradual variation for longer periods, even for the same talker. By setting up an appropriate measure of distance, however, taking the long-term variation of the pattern into consideration, it is ascertained that the recognition of the talker is very well achieved after an interval longer than three months.

Using the long-time averaged spectrum taken four times at three-month intervals as the known samples, and setting the weighted distance considering the long-term variation of the spectrum,

an average recognition rate of 91% for talker identification of 1 out of 9 talkers and 93 to 95% for talker verification with 8 to 26 impostors was achieved for unknown samples collected three months after the last known sample. If one shortens the measurement interval between the known samples to one month, and revises the standard pattern of each registered talker by accumulated data (known samples) over about 6 months at one month intervals, a better recognition rate may be obtained.

Acknowledgement

The authors wish especially to acknowledge the guidance by Dr. Noda, Director of the Research Division, and other members of Saito Research Section and the 4th Research Section.

References

- (1) S. Saito: Talker Recognition, *Mathematical Sciences*, No. 116, p. 49, 1973.
- (2) W.A. Hargreaves and J.A. Starkweather: Recognition of Speaker Identity, *Language & Speech*, 6, p. 63, 1963.
- (3) P.D. Bricker, et al.: Statistical Techniques for Talker Identification, *B.S.T.J.*, 50, 4, p. 1427, 1971.
- (4) S.K. Das and W.S. Mohn: A Scheme for Speech Processing in Automatic Speaker Verification, *IEEE Trans.*, AU-19, 1, p. 32, 1971.
- (5) J.E. Luck: Automatic Speaker Verification Using Cepstral Measurements, *J. Acoust. Soc. Amer.*, 46, 4, 2, p. 1026, 1969.
- (6) N. Ochiai and A. Fukumura: Basic Quality of Speech Communication, *E.C.L. Tech. Jour.*, NTT, Jap., 8, 12, p. 1469, 1959.
- (7) H. Mangold: Die Individuellen Unterschiede bei Langzeitmessungen am Sprachsignal, *NTZ*, 6, p. 364, 1969.
- (8) G. Ramishvili: Some Problems in Automatic Identification of Spoken Utterances and Speakers, *STL-QPSR*, 2, 3, p. 36, 1970.
- (9) B.P. Bogert, et al.: The Quefrency Alanysis of Time Series for Echoes, *Proc. of the Symposium on Time Series Analysis*, John Wiley & Sons, New York, p. 209, 1963.
- (10) S. Furui, F. Itakura and S. Saito: Talker Recognition by the Long-Time Averaged Speech Spectrum, *Trans. IECE.*, 55-A, 10, p. 550, 1972.
- (11) S. Furui and F. Itakura: Talker Recognition by Statistical Features of Speech Sounds, *Trans. IECE.*, 56-A, 11, p. 717, 1973.

* * * *

1934

1935