論文 / 著書情報
Article / Book Information

Title	Personal Information in Dynamic Characteristics of Speech Spectra
Author	Shuzo Saito, Sadaoki Furui
Journal/Book name	Proc. 4th IJCPR, Vol. , No. , pp. 1014-1018
発行日 / Issue date	1978,
権利情報 / Copyright	(c)1978 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

PERSONAL INFORMATION IN DYNAMIC CHARACTERISTICS OF SPEECH SPECTRA

Shuzo Saito and Sadaoki Furui
Musashino Electrical Communication Laboratory
Nippon Telegraph and Telephone Public Corporation
Musashino, Tokyo, 180, Japan

Summary The dynamic programming procedure is applied for extraction of dynamic characteristics of personal information. Feature parameters used for speech analysis are the PARCOR parameter and the fundamental frequency. Deriving the dynamic programming matched path between the reference pattern and speech input of unknown talker, it is found that the matched path is near by the diagonal in the case of within talker, but much more spread around the diagonal in the case of between talkers, and that there is the specific region for individual talker where the rate of similarity in dynamic programming matched path is very high in the case of within talker. By the use of the dynamic programming matched path as one of measures of personal information, the error rate of talker recognition is reduced to about 2 % in the test performed by two words.

INTRODUCTION

Personal information contained in speech sounds is divided into the two components, that is, the speech spectrum envelope and the voice excitation characteristics. Physical characteristics of these two components vary temporarily in accordance with the movement of the articulatory organ. Effects of these characteristics on talker recognition problem have been studied on the viewpoints of the static and dynamic characteristics of personal information. Studies based on the static characteristics of personal information have been reported by authors using the long-term averaged speech spectrum [1] and the time averaged characteristics of the PARCOR parameters and fundamental frequency of the voiced portion of spoken words. [2] As the testing words are fixed in the latter study, dynamic characteristics of personal information was involved implicitly, but not used explicitly for talker recognition. In this paper, the dynamic characteristics of personal information is analyzed by the use of dynamic programming procedure and tested its contribution to talker recognition.

SPEECH ANALYSIS PROCEDURE FOR EXTRACTION OF PERSONAL INFORMATION

For the extraction of personal information, the PARCOR parameters [3] and fundamental frequency of speech sounds are used as the feature parameters. PARCOR parameter is defined as follows: Consider a time sequence of (n+1) samples of a speech signal: X_{t-n}, $X_{t-(n-1)}$, $\cdots$ X_{t-1}, X_t. Values X_{t-n} and X_t can be predicted from sample values between samples X_{t-n} and X_t by means of the least squares method. The predicted values are denoted by $\hat{X}_{t-n}$ and $\hat{X}_t$ respectively. The PARCOR parameter is defined as the correlation coefficient between the prediction error values of $X_{t-n}-\hat{X}_{t-n}$ and $X_t-\hat{X}_t$. The correlation coefficient is called n-th order PARCOR coefficient. Several PARCOR coefficients are obtained by varying the value of n. The basic principle of extracting PARCOR parameter from a speech signal is shown in Fig. 1. In practice PARCOR parameters can be easily obtained by the recursive implementation of the autocorrelation equations. The frequency spectrum envelope of speech signal is expressed more precisely and efficiently by PARCOR parameter than the conventional autocorrelation coefficient. The fundamental frequency of speech signal is extracted from the peak period of correlation function of the prediction residual.

Speech signal is fed through a low pass filter of 3400 Hz cutoff, then sampled at 8000 Hz and digitized its amplitude in twelve bits. Using such discrete speech samples, the PARCOR parameters and fundamental frequency are extracted in every 10 ms of speech samples. In other words, the frame frequency of speech analysis is 100 Hz. Number of speech samples used for extraction of feature parameters in each frame is 256 and the time window of Hamming shape is applied. Then PARCOR parameter is transformed into the log-area-ratio (LAR) as shown in the following equation.

$$x_{ti} = \operatorname{arctanh} k_{ti} = \frac{1}{2} \log \frac{1 + k_{ti}}{1 - k_{ti}} \qquad (1)$$

where x_{ti} : LAR of i-th order at t-th frame,
k_{ti} : PARCOR parameter of i-th order at t-th frame.

The fundamental frequency at t-th frame is denoted as f_t. Then the personal information vector at t-th frame is defined as follows,

$$\mathbf{X}_t = (x_{t1}, x_{t2}, \cdots x_{tN}, f_t) \qquad (2)$$

where N : maximum order of the PARCOR parameter.

The reference pattern registered in advance for a known talker is defined as a similar expression to Equation (2) and is denoted by $\mathbf{Y}_t$. The measure of pattern matching between the reference pattern and speech input is defined by a distance in a vector space as shown in Equation (3).

$$d_{tkj} = (\mathbf{X}_{tkj} - \mathbf{Y}_{t'k}) \mathbf{W} (\mathbf{X}_{tkj} - \mathbf{Y}_{t'j})^T \qquad (3)$$

where d_{tkj} : distance measure of personal information between speech input $\mathbf{X}_{tkj}$ and reference pattern $\mathbf{Y}_{t'k}$,

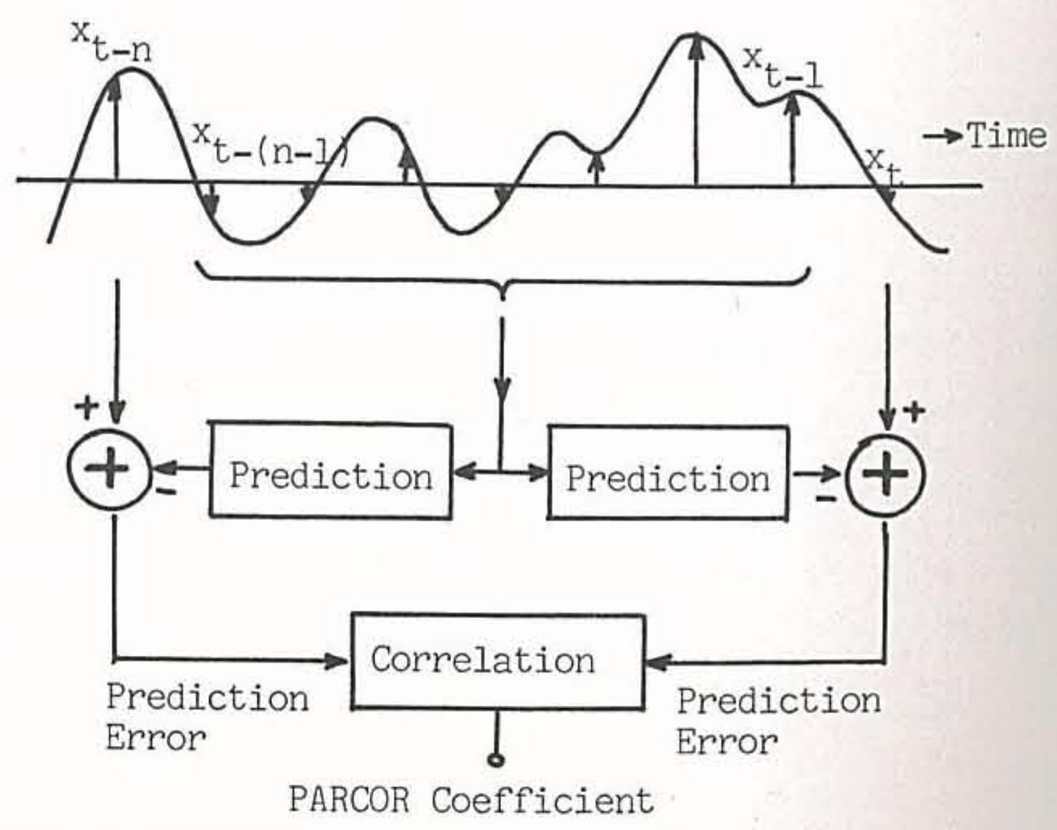


Fig.1 - Principle of extracting PARCOR coefficient.

$\mathbf{X}_{tkj}$: personal information vector of k-th talker at t-th frame in j-th speech sample sequence,

$\mathbf{Y}_{t'k}$: personal information vector of k-th talker at t'-th frame of the reference pattern,

$\mathbf{W}$: weighting matrix of personal information.

$$\mathbf{W} = \begin{pmatrix} w_1 & \cdot & \cdot & \cdot & 0 \\ \cdot & w_2 & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & w_{N+1} \end{pmatrix}$$

The diagonal elements w_i (i=1,2, .. N+1) are derived as follows. Let μ_{kj} denotes the averaged value of personal information vector over the voiced portion of reference pattern uttered by k-th talker,

$$\mu_{kj} = \mathop{E}_{t} \mathbf{X}_{tkj} \tag{4}$$

where $\mathop{E}_{t}$: average on the voiced portion of speech input.

Variance of μ_{kj} on speech sample sequence for k-th talker is denoted as σ_k^2, that is,

$$\sigma_k^2 = \mathop{Var}_{j} \mu_{kj} \tag{5}$$

Then σ_k^2's of whole talker are averaged and denoted by $\bar{\sigma}^2$. The weighting matrix elements w_i is defined as the reciprocal of element of $\bar{\sigma}^2$. That is,

$$\bar{\sigma}^2 = \mathop{E}_{k} \sigma_k^2 \tag{6}$$

$$w_i = \frac{1}{\bar{\sigma}_i^2} \tag{7}$$

In the dynamic programming procedure for matching between the reference pattern and speech input, matching path is selected so as to minimize the distance of personal information defined as Equation (3). Let the reference pattern and speech input are arranged on orthogonal axis as shown in Fig. 2, the dynamic programming matched path traces from the left-lower corner to right-upper corner. As the dynamic programming matched path for the case of same talker will differ from that of different talker, it seems to be available to talker recognition as one of the measure of dynamic characteristics of personal information.

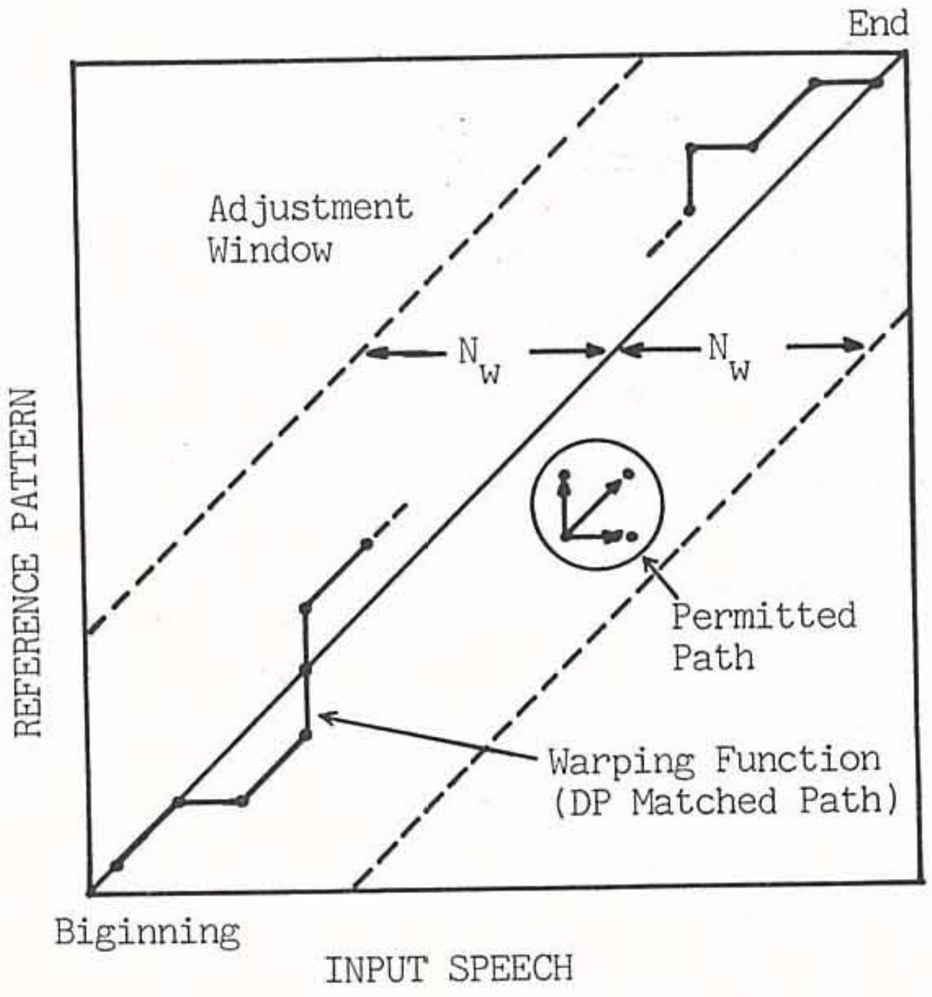
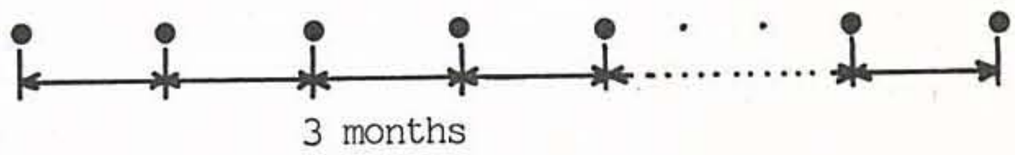


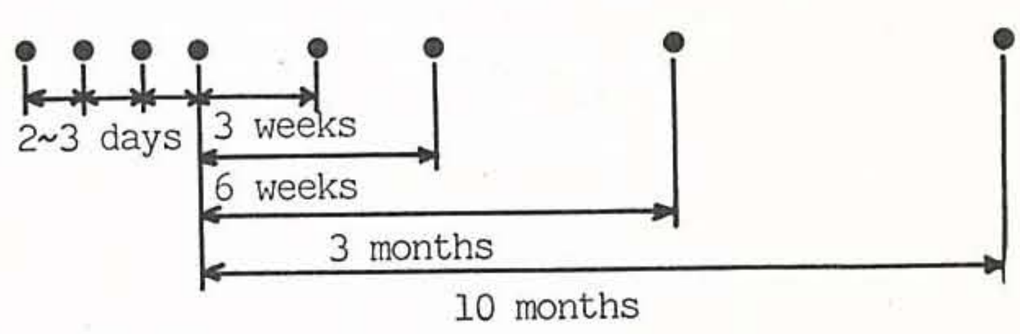
Fig.2 - Dynamic programming procedure between the reference pattern and input speech.

EXPERIMENT ON EXTRACTION OF DYNAMIC PERSONAL INFORMATION

Testing speech materials are utterances of nine male talkers of Japanese words /namae/ (name in English) and /baNgo:/ (number in English). These speech samples are uttered in two kinds of time sequences, that is, the long-term samples and the short-term samples as shown in Fig. 3 (a), (b). In the long-term samples, each talker uttered at eight sample sequences, each sequence was set apart at intervals of three months. In each sequence every talker uttered three times, so the total number of speech samples was 9×8×3 = 216. The reference pattern for each talker is derived as expressed in Equation (2) by the use of three speech samples, each one is extracted from the former three sample sequences. The rest 21 speech samples of each talker are used as speech input for dynamic programming procedure. In the short-term samples, each talker uttered at three sample sequences during 2 or 3 days at first, and these are used for the reference pattern. Then each talker uttered testing speech samples at four time sequences, each sequence was set apart at the intervals of 3 weeks, 6 weeks, 3 months and 10 months.



(a) Long-term Samples



(b) Short-term Samples

Fig.3 - Time sequences of speech samples.

Results of the dynamic programming procedure are shown in Fig. 4. Fig. 4 (a) and (b) are results for the cases of within talker and between talkers, respectively. It is seen that the dynamic programming matched paths are near to the diagonal in the case of within talker, but much more spread around the diagonal in the case of between talkers. It appears that there are three or four constricted parts, that is, invariant parts in time in the case of within talker. For the measure of similarity of personal information between the reference pattern and speech input, the restriction for dynamic programming matched path as shown in the following equation is considered.

$$D(i,j) = \begin{cases} D(i-1,j-1)+d(i-1,j)+d(i,j) \\ D(i-1,j-1)+2d(i,j) \\ D(i-1,j-1)+d(i,j-1)+d(i,j) \end{cases} \tag{8}$$

$$\text{where } D(i,j) = \text{Min}[D(i-1,j)+d(i,j), \; D(i-1,j-1)+2d(i,j), \; D(i,j-1)+d(i,j)] \tag{9}$$

Using the dynamic programming matched path, the rate of similarity between the reference pattern and speech input, which is defined as the frequency of occurrence of matched path that satisfies Equation

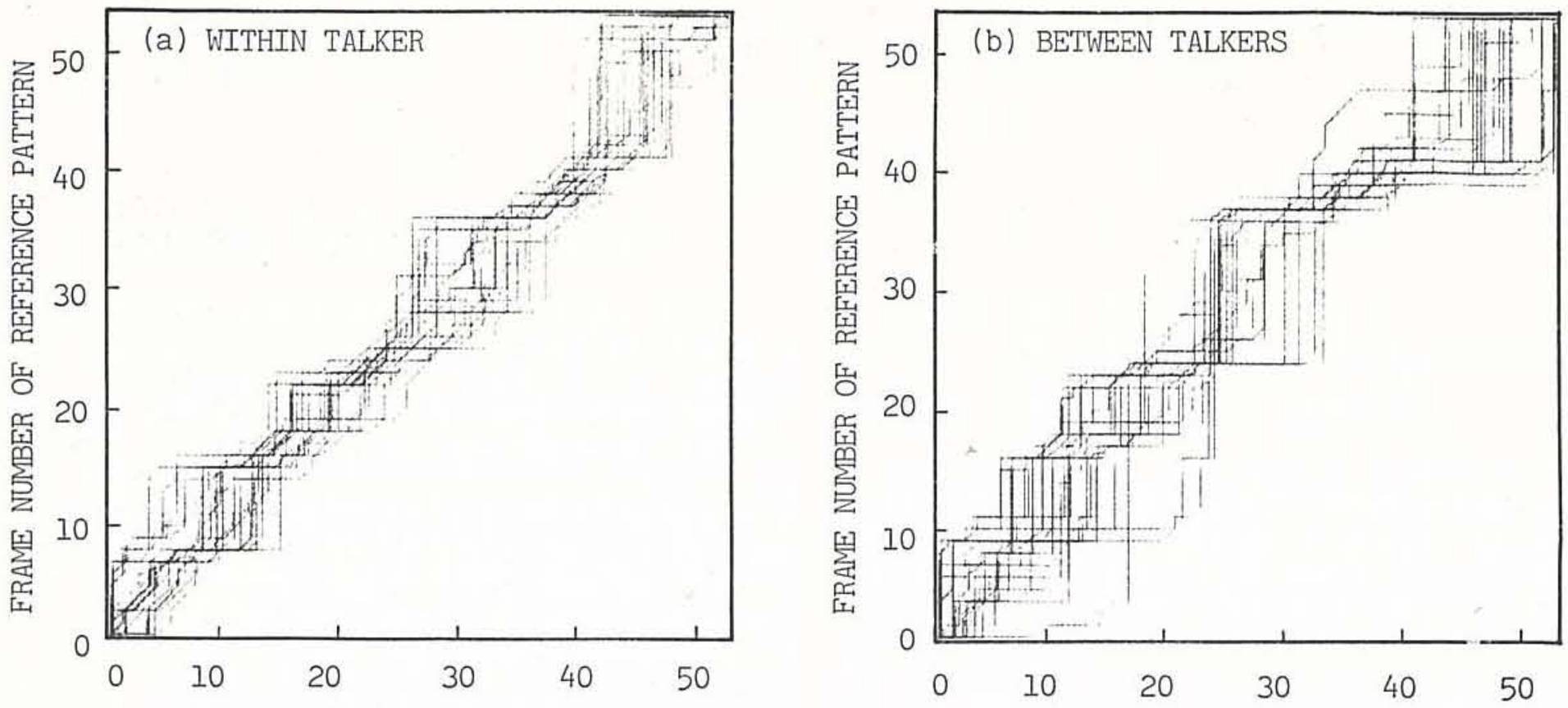


Fig.4 - Results of the dynamic programming procedure.

(8), is calculated in every frame number of speech input and is shown in Fig. 5. Rates of similarity are shown for the cases of within and between talkers, respectively. It appears that there are several frames of higher rates of similarity at the regions of phoneme concatenation, that is the transitional parts of a word in the case of within talker. It need scarcely be said that rate of similarity is lower in the case of between talkers. Assuming the rate of similarity of 50 % is a threshold, the frame number of more than 50 % similarity rate is extracted for each talker. Such a specific region of speech input for individual talker seems to be available for talker recognition. Example of such specific regions are shown in Fig. 6 accompanying several results of speech spectrum analysis.

APPLICATION OF DYNAMIC PROGRAMMING PROCEDURE TO TALKER RECOGNITION

For a speech input of unknown talker, the dynamic programming procedure is applied and talker recognition is executed for each test word and their combination. Results are shown in Fig. 7 (a) and (b). Fig. 7 (a) is the results for speech input of the long-term samples. In this figure, the error rates of talker recognition obtained by the dynamic programming procedure are compared with those obtained by the time averaged characteristics of frequency spectrum.[2] It is seen that the recognition rates for test word /baNgo:/ are a little superior than those of test word /namae/ and combining the two test words reduces the error rate to half of it, and that the recognition rates obtained by the dynamic programming procedure

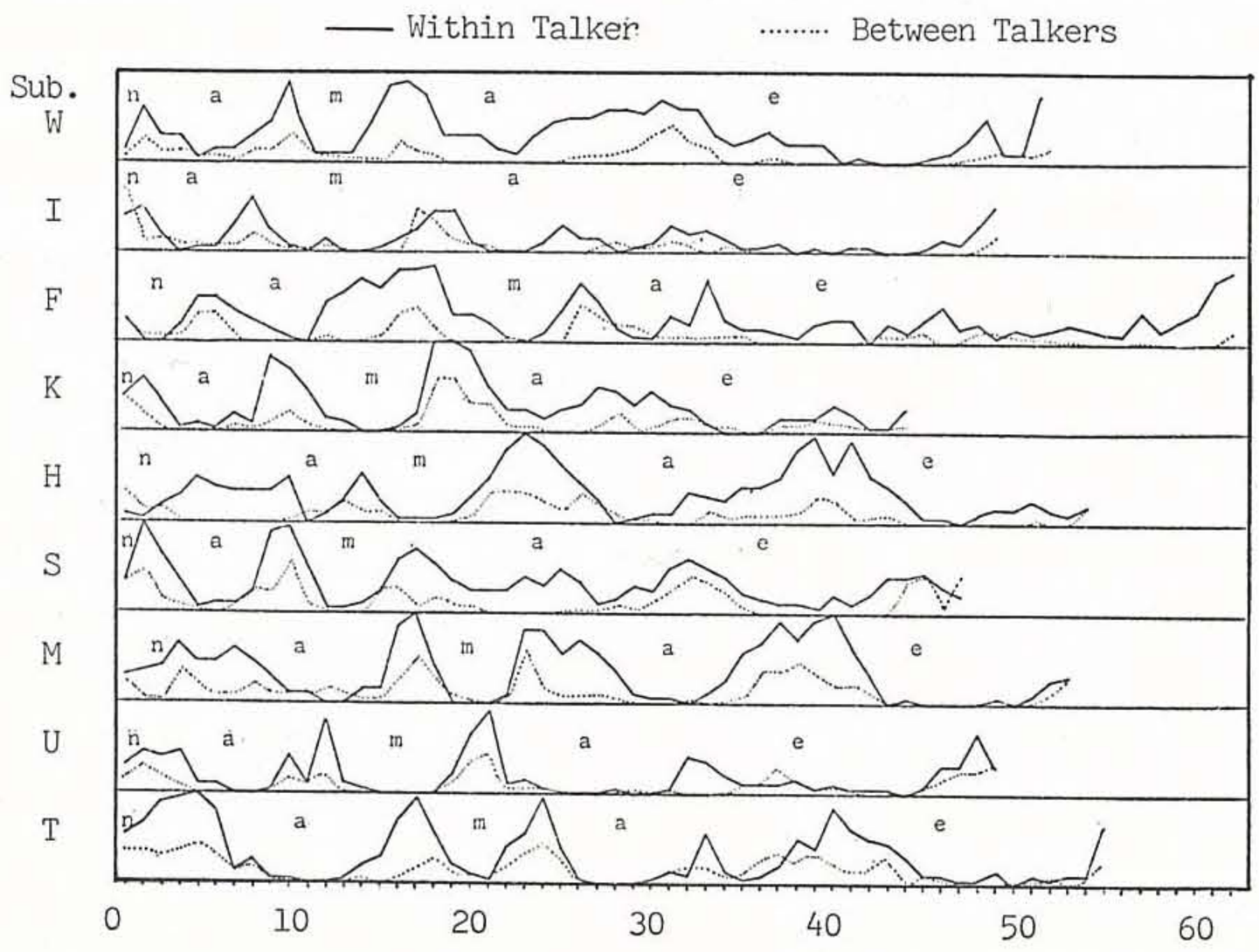


Fig.5 - Rate of similarity in dynamic programming matched path.

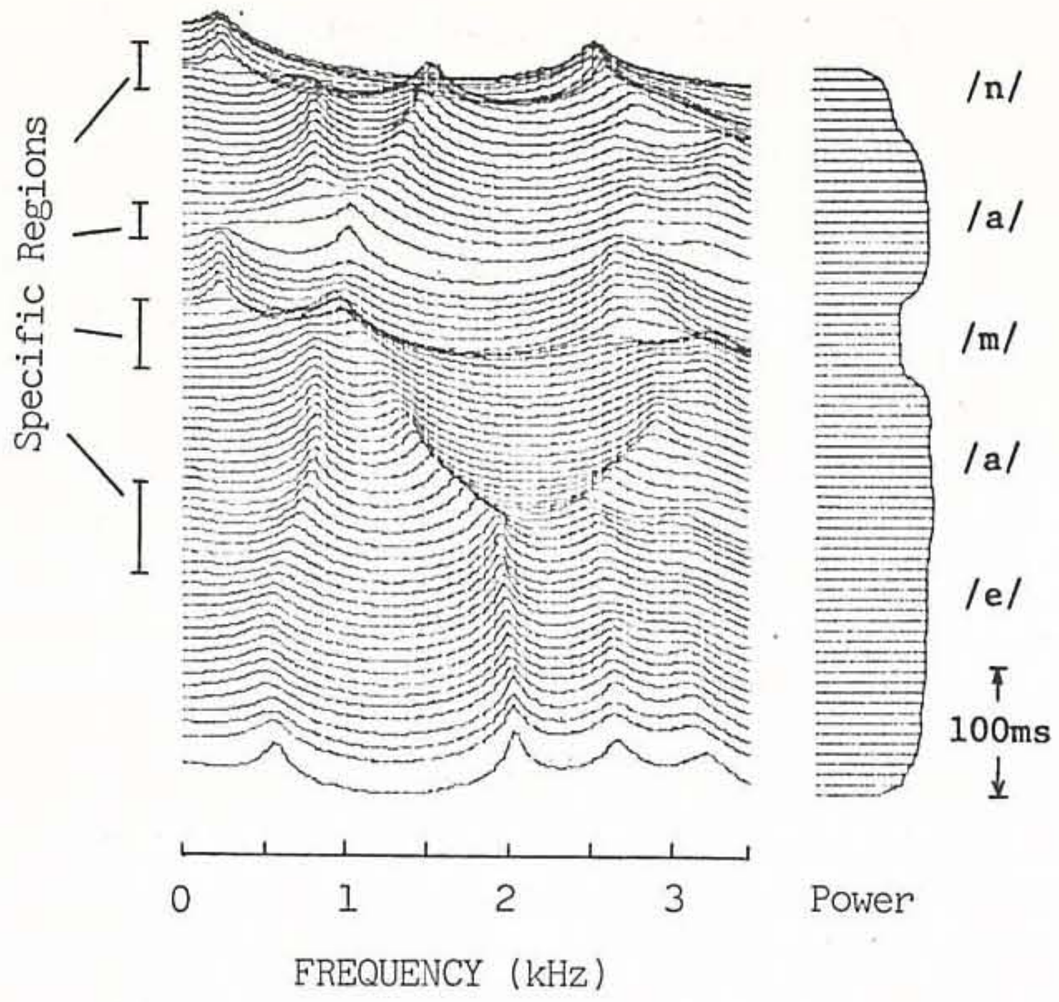


Fig.6 - An example of specific regions in an utterance /namae/.

are better than those by the static characteristics of frequency spectrum.

Fig. 7 (b) is the results for speech input of short-term samples. In this figure, error rates of talker recognition obtained by the time averaged characteristics of frequency spectrum are also plotted for comparison, and they are expressed in the average value of the two testing words. As the speech samples used for the reference pattern of talker recognition are uttered in such a short time interval as about ten days, error rates of talker recognition are increased monotonously in accordance with the increase of time interval between the reference pattern and speech input of unknown talker, and error rates for speech input after 6 weeks is comparable with those obtained for speech input of long-term samples. It is also seen that the error rates for the dynamic programming procedure are less than those for the time averaged characteristics of frequency spectrum and its reduction rate is about one third. It appears that talker recognition based on the dynamic programming procedure is an efficient method to reduce the effect of temporal variation of speech input characteristics, especially in the short-term samples.

As is shown in Fig. 5 and 6, there are several specific regions of the dynamic programming matched path inherent to individual talker. The rate of similarity in specific regions defined in the preceding section may be used as one of measures of talker recognition. Calculating the rate of similarity in specific regions for speech inputs of nine talkers, ranking of similarity of speech input to the talker of reference pattern is performed and then the distribution of correct estimation of talker relative to the rank order is derived. The cumulative distributions of talker estimation are shown in Fig. 8 (a) and (b). Fig. 8 (a) is the results for speech input of the long-term samples. It is seen that the rate of talker estimation is about 63 % by the use of the rate of similarity in specific regions for the speech input of two words combination, when only the first rank of estimated talker is considered. When the estimated talkers untill third rank are considered, the rate of talker estimation becomes up to about 88 %. Fig. 8 (b) is the results for speech input of short-term samples. It is seen that the rate of talker estimation is much similar to that for speech input of long-term samples and is relatively stable for the change of time interval between the reference pattern and speech input of unknown talker. It appears that the rate of similarity in specific region inherent to individual talker is useful as a supplementary measure for talker recognition, although it is insufficient to be used as an independent measure for talker recognition.

Combining the similarity measure in specific regions of dynamic programming matched path with the distance measure of the dynamic programming procedure, experiments of talker recognition are performed for speech inputs of two kinds of time sequences. Results are shown in Fig. 9 (a) and (b). Fig. 9 (a) is the results for speech input of the long-term samples. It is seen that the rate of talker recognition could be improved in about 2 % by the aid of the similarity measure for speech input of single word, although the improvement in talker recognition by the similarity measure is far less for speech input of two words combination. Fig. 9 (b) is the results for speech input of the short-term samples. It is seen that the similarity measure is efficient to reduce the effect of time interval between the reference pattern and speech input of unknown talker.

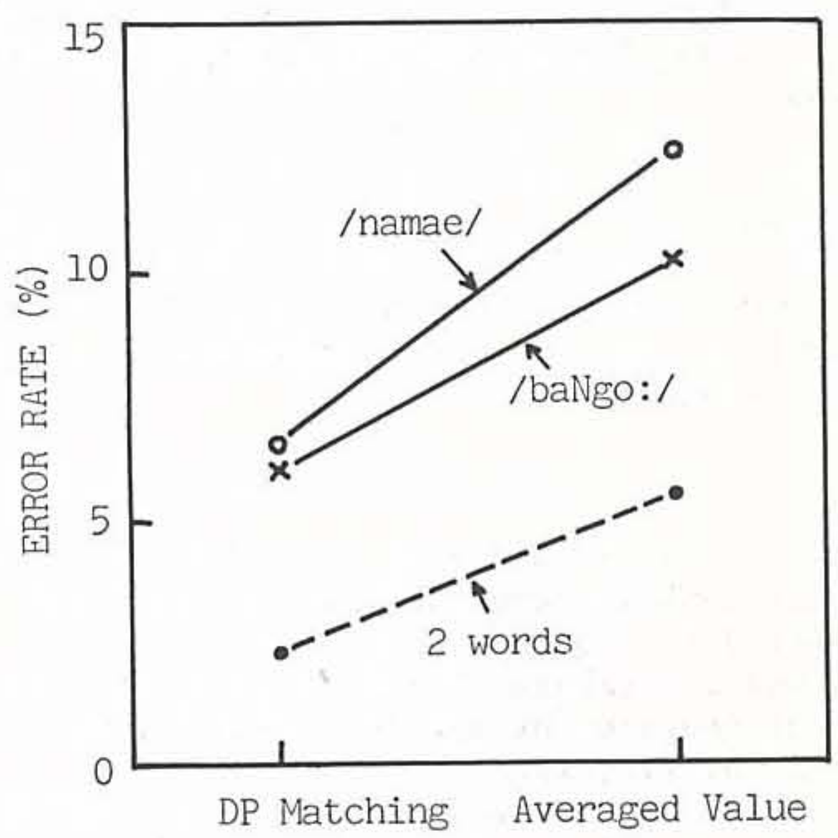


Fig.7 (a) - Error rates for speech input of the long-term samples in the dynamic programming procedure.

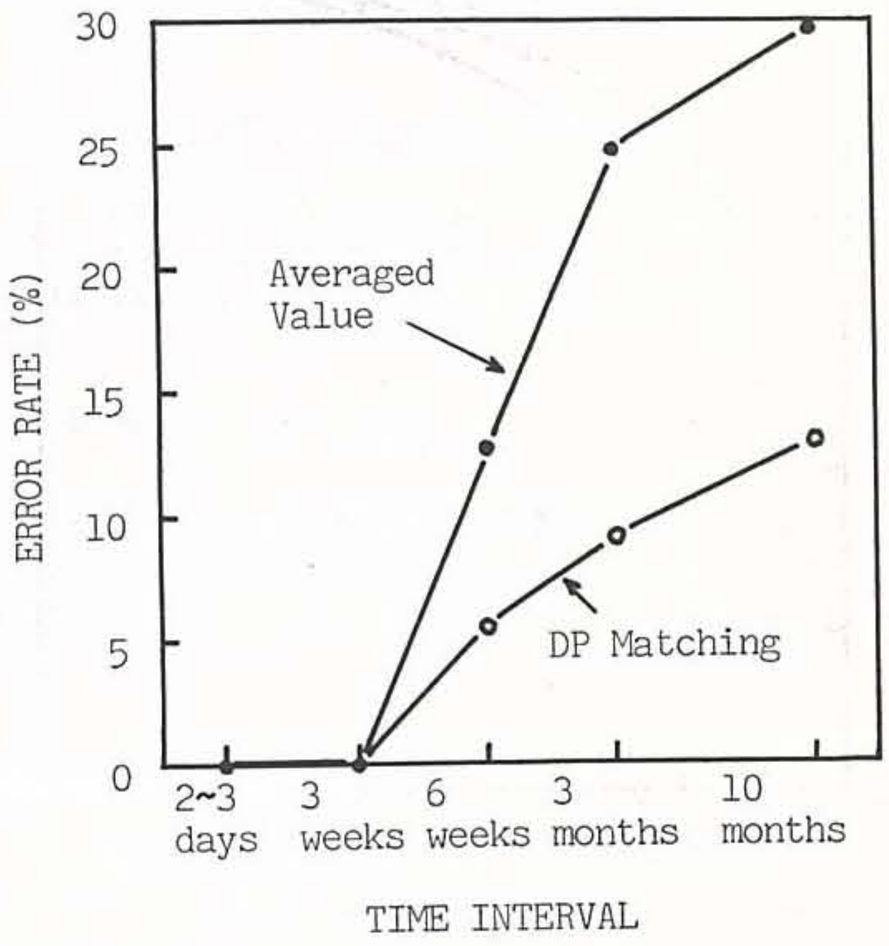


Fig.7 (b) - Error rates for speech input of the short-term samples in the dynamic programming procedure.

CONCLUSION

The dynamic programming procedure is used to extract the dynamic characteristics of personal information and to apply them to talker recognition. Following results are derived from this study.

(1) PARCOR parameter and fundamental frequency are used as feature parameters for analyzing personal information. Based on these feature parameters, the dynamic programming procedure is used for matching between the reference pattern and speech input of unknown talker and the matched distance with weighting on feature parameters is derived. It is found that this dynamic distance measure is more efficient for talker recognition than the conventional static measure on speech spectrum.

(2) In the dynamic programming matched path within talker, there are several specific regions inherent to individual talker, where the rate of similarity is high, that is, the dynamic programming matched path traces along the diagonal. The similarity measure in the specific regions is useful as a supplementary measure for talker recognition.

(3) Combining the dynamic distance measure and the similarity in the specific regions on the dynamic programming matched path, error rate of talker recognition is reduced to about 2 % in the test performed by two testing words.

Acknowledgement

The authors wish to express their thanks to members of Speech Research Group of Musashino Electrical Communication Laboratory for enthusiastic cooperation and discussion.

References

1. S. Furui, F. Itakura and S. Saito, "Talker recognition by the longtime averaged speech spectrum", Trans. IECE Jap., 55-A, p. 550, 1972
2. S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds", Trans. IECE Jap., 56-A, p. 717, 1973
3. F. Itakura, S. Saito, T. Koike, H. Sawabe and M. Nishikawa, "An audio response unit based on partial autocorrelation", IEEE Trans. COM-20, p. 792, 1972

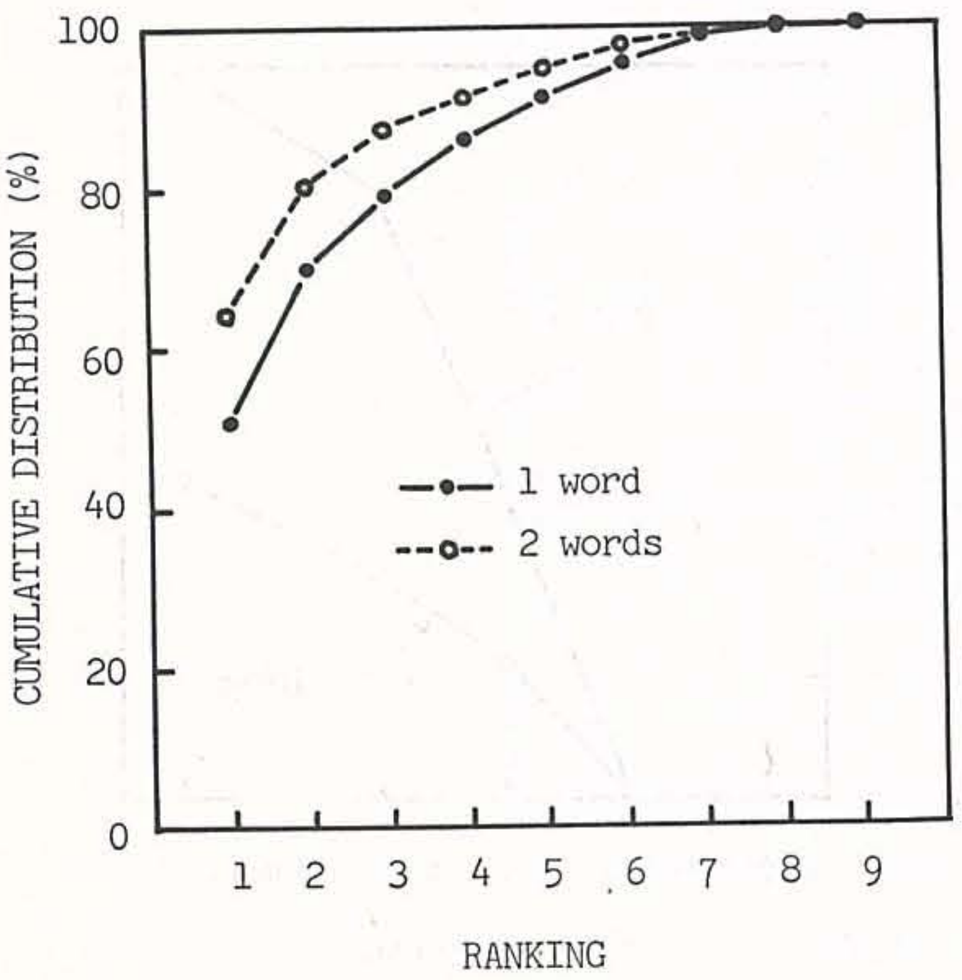


Fig.8 (a) - Cumulative distribution of frequency of talker estimation using the similarity measure for long-term samples.

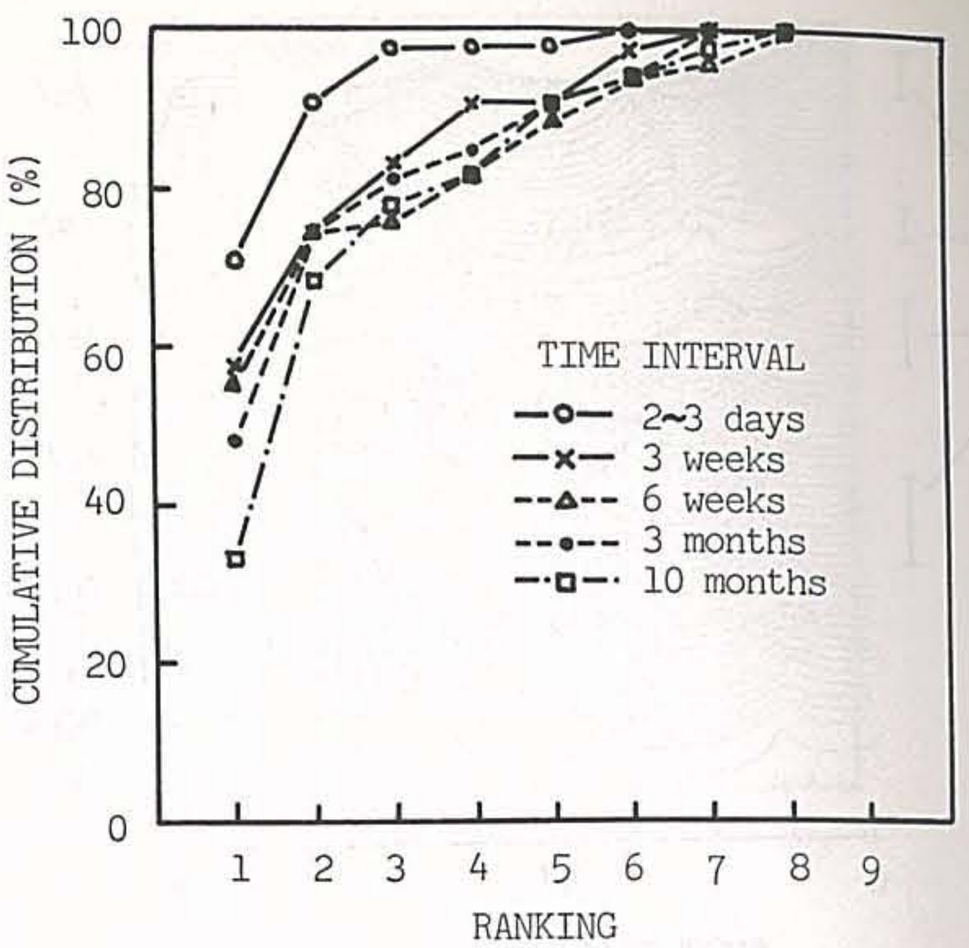


Fig.8 (b) - Cumulative distribution of frequency of talker estimation using the similarity measure for short-term samples.

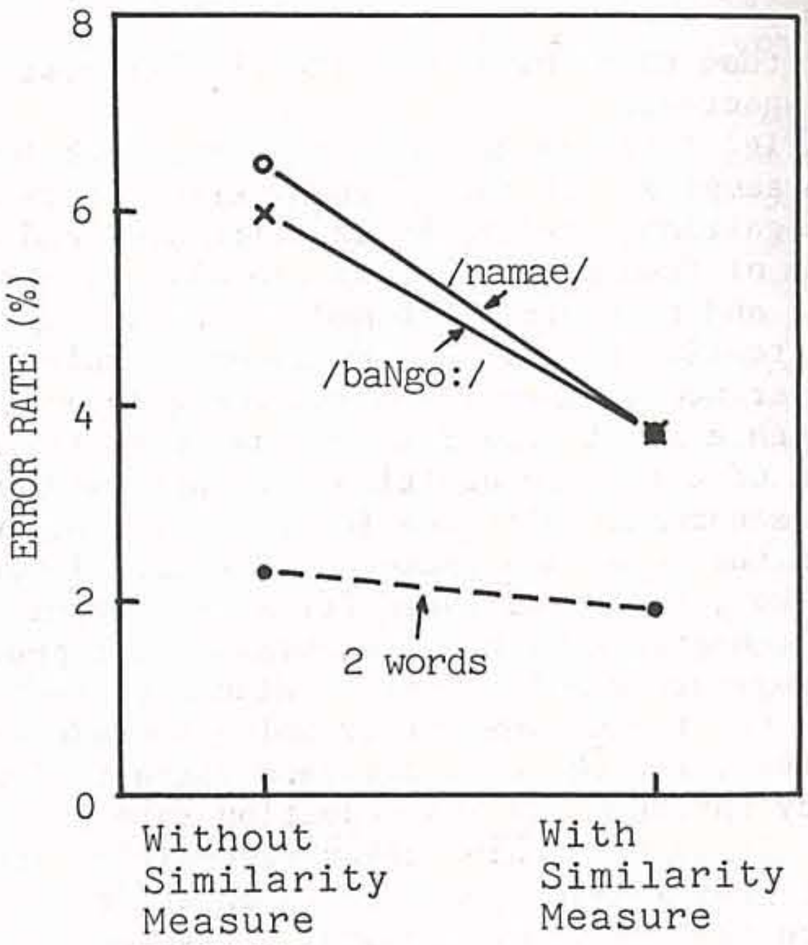


Fig.9 (a) - Effects of the similarity measure on talker recognition of long-term samples.

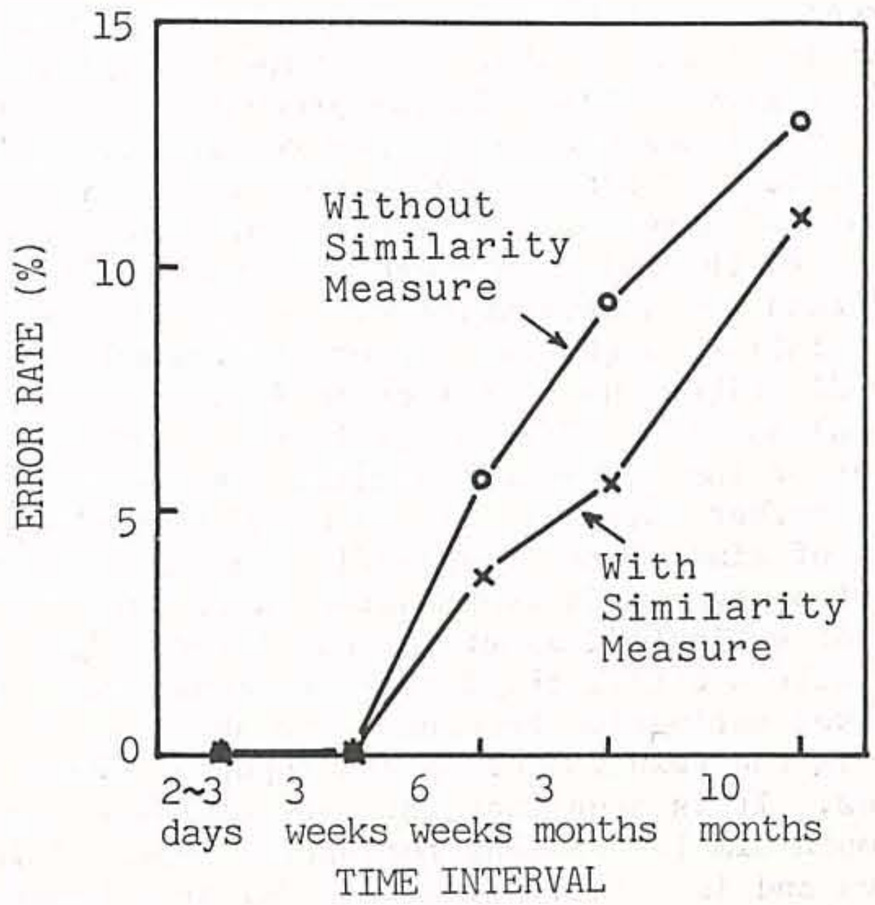


Fig.9 (b) - Effects of the similarity measure on talker recognition of short-term samples.