Ann. Bull. RILP, No. 13, 145-153 (1979)

ANALYSIS OF DYNAMIC CHARACTERISTICS OF SPEECH SPECTRUM CHARACTERIZING INDIVIDUAL SPEAKERS

Shuzo Saito and Sadaoki Furui*

Introduction

Speech information characterizing an individual speaker is composed of two kinds of tonal characteristics: the static and the dynamic characteristics of speech sounds. Various researchers have hitherto studied the problem primarily from the viewpoint of the static characteristics of the speech spectrum, namely, the long-term averaged speech spectrum [1] and time averaged speech spectrum of voiced portions of words. [2]

Recently, however, the dynamic characteristics of the speech spectrums have begun to be analyzed using the dynamic programming procedure with this then being applied to speaker recognition.

In this paper, a speech analysis procedure for extraction of individual speaker information will be outlined and then will follow a description of experimental procedures and results.

Speech Analysis Procedure

Feature parameters used for the extraction of speaker information are the PARCOR coefficient and fundamental frequency of speech sounds. The PARCOR coefficient is a kind of Linear Prediction Coefficient and is defined in Fig. 1. [3] Speech signals are sampled to (n+1) discrete samples periodically. The conventional autocorrelation coefficient is defined as the correlation between two sampled values such as x_{t-n} and x_t, but the PARCOR coefficient is defined as the correlation between the two prediction error values. Prediction error values are calculated as differences between the actual sample values and the predicted values by the forward and backward predictions, respectively. In practice PARCOR parameters can be easily obtained by the recursive implementation of the autocorrelation equations. By the use of the PARCOR coefficient, a frequency spectrum envelope of the speech signal is represented more precisely and effectively than the conventional autocorrelation coefficient. The fundamental frequency of the speech signal is derived from the calculation of the peak period of the prediction residual wave.

The speech signal is filtered in the 3.4 kHz frequency band, sampled at 8 kHz and then its amplitude digitized into twelve bits. Using such discrete speech samples, the PARCOR parameters and fundamental frequency are extracted from every 10 ms of the speech samples; that is, the frame frequency of speech analysis is 100 Hz. The number of speech samples used for extraction of feature parameters in each frame was 256 and the time window of Hamming shape was applied. The PARCOR parameter was then transformed into the log-area-ratio (LAR) as shown in the following equation.

* Musashino Electrical Communication Laboratory, N.T. & T.

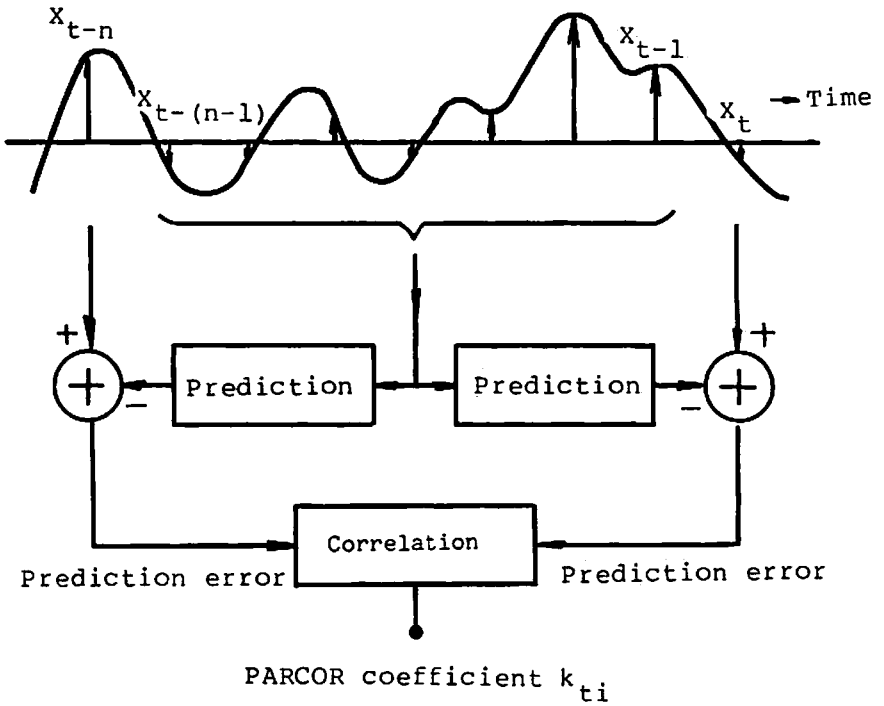

Fig. 1 Principle of extracting PARCOR coefficient

$$x_{ti} = \text{arctanh } k_{ti} = \frac{1}{2} \log \frac{1+k_{ti}}{1-k_{ti}} \tag{1}$$

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

k_{ti} : PARCOR coefficient of i-th order at t-th frame.

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

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

where N : maximum order of the PARCOR coefficient.

The reference templet registered in advance for a known speaker is defined as a similar expression to Equation (2) and is denoted by $\boldsymbol{Y}_t$. The first measure of similarity between the reference templet and speech input of an unknown speaker is defined by the distance in a vector space as shown in Equation (3).

$$d_{tkj} = (X_{tkj} - Y_{t'k}) W (X_{tkj} - Y_{t'k})^T \quad (3)$$

where d_{tkj}: distance measure of speaker information between speech input X_{tkj} and reference templet $Y_{t'k}$

X_{tkj}: speaker information vector of the k-th speaker at the t-th frame in the j-th speech sample sequence,

$Y_{t'k}$: speaker information vector of k-th speaker at t'-th frame of the reference templet,

W : weighting vector of speaker information,

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

The diagnoal elements w_i (i=1, 2, .. , N+1) of the weighting vector W are derived as follows. Let μ_{kj} denote the averaged value of the speaker information vector over the voiced portion of a specific word uttered by the k-th speaker,

$$\mu_{kj} = \underset{t}{E} X_{tkj} \quad (4)$$

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

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

$$\sigma_k^2 = \underset{j}{Var}\, \mu_{kj} \quad (5)$$

The σ_k^2s of all speakers are averaged and denoted by $\overline{\sigma}^2$.

$$\overline{\sigma}^2 = \underset{k}{E}\, \sigma_k^2 \quad (6)$$

Then, the weighting matrix elements w_i is defined as the reciprocal of element of $\overline{\sigma}^2$.

$$w_i = \frac{1}{\overline{\sigma}_i^2} \quad (7)$$

In the dynamic programming procedure for matching the reference templet and speech input, a matching path was selected so as to minimize the distance of speaker information defined in Equation (3). The reference templet and speech input were arranged on the orthogonal axis: the dynamic programming matched path then traces from the left-lower corner to the right upper corner. As the dynamic programming matched path for a speaker will differ from that of a different speaker, this seems to be one of the

measures of the dynamic characteristics of speaker information useful for speaker recognition.

Experiment on Extraction of Dynamic Speaker Information

The speech materials used in the testing were the utterances by nine male speakers of the two Japanese words, /namae/ and /baNgo:/. These test sounds were uttered in two time sequences, at long-term and short-term intervals. In the case of the long-term interval, each speaker uttered test sounds every three months over a period of 21 months. The reference templet for each speaker was derived as expressed in Equation (2) by the use of three speech samples. The rest of the speech samples of each speaker were used as speech input for the dynamic programming procedure. In the case of the short-term interval, each speaker uttered test sounds three times in advance, which were used as the reference templet then uttered test sounds every three weeks during a 10 month period.

Results of the dynamic programming procedure are shown in Fig. 2. Fig. 2 (a) and (b) are results for the cases of within and between speakers, respectively. It can be seen that the dynamic programming matched paths are near the diagonal in the case of within speaker, but much more spread around the diagonal in the case of between speakers. It appears that there are three or four constricted parts, that is, invariant parts in the case of within speaker. For the second measure of similarity of speaker information between the reference templet and speech input, the rate of the matched path on the diagnoal is considered. Such a rate is calculated in every frame of the reference templet and is shown in Fig. 3.

In this figure, the abscissa represents the frame number of the reference templet and the ordinate is the rate of the matched path on the diagonal. The solid line is the result of within speaker and the dotted line is between speakers. It can be seen that the rate of the matched path on the diagonal is lower in the case of between speakers and 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 speakers. Assuming the rate of 50% of the matched path on the diagonal is a threshold, the frame numbers exceeding this threshold were extracted as the specific regions of speaker information for the individual. This specific region of speech input seems to be available for speaker recognition.

Application of Dynamic Programming Procedure to Speaker Recognition

The dynamic programming procedure was applied to speech input of unknown speakers and a speaker recognition experiment was executed. Results are shown in Fig. 4. In Fig. 4(a), the results of speech input for the long-term interval are shown and these are compared with those of the speaker recognition experiments based on the static characteristics of the frequency spectrum. It seems that the recognition rates of the two Japanese words are rather similar and the combined use of two words reduced the error rate by half. It can also be seen that the recognition rates obtained by the dynamic programming procedure were much better than those of the static characteristics of the frequency spectrum. The results of speech input for the short-term interval are shown in Fig. 4(b), and are compared with those for the static characteristics of the frequency spectrum. Speech

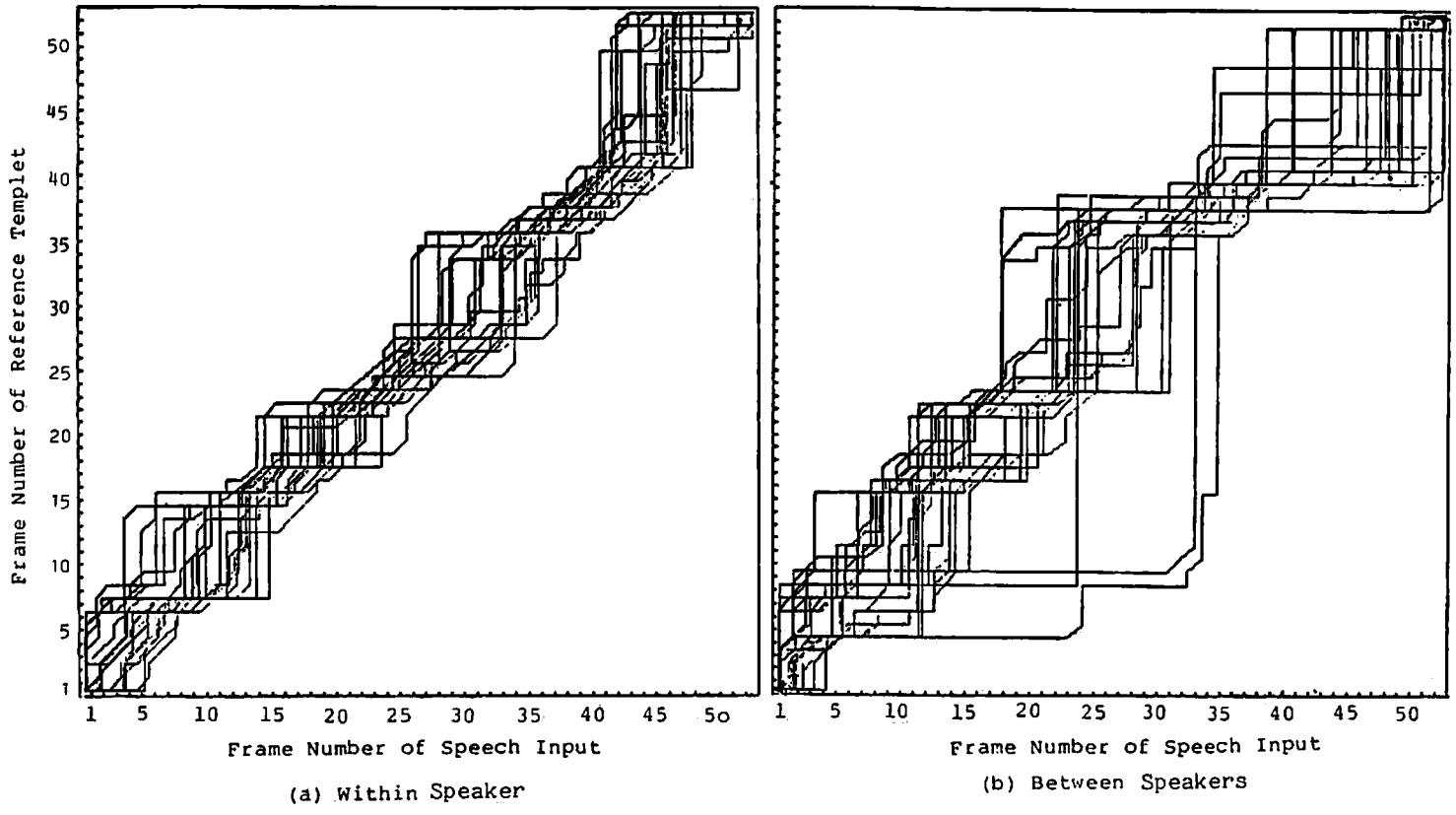

Fig. 2 Examples of dynamic programming matched path

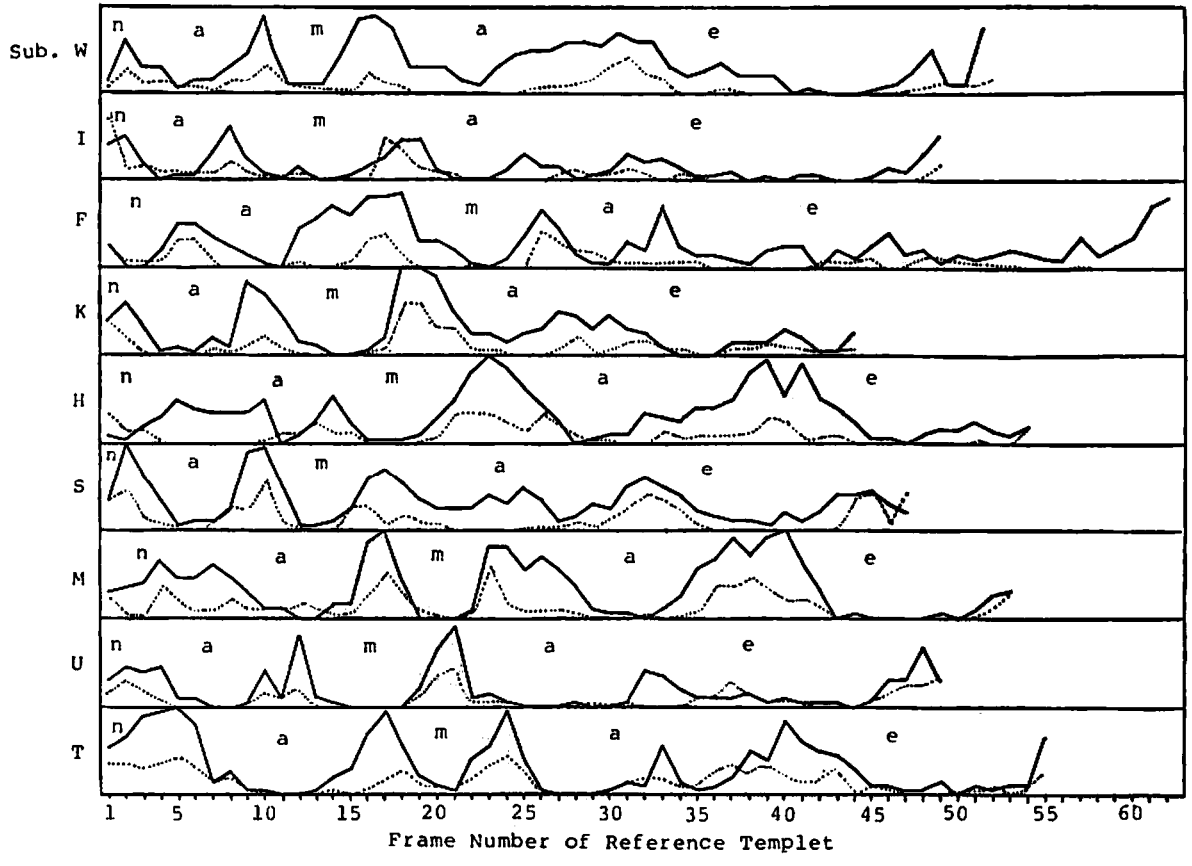

Fig. 3 Rates of the matched path on diagonal in dynamic programming procedure

——— Within Speaker ····· Between Speakers

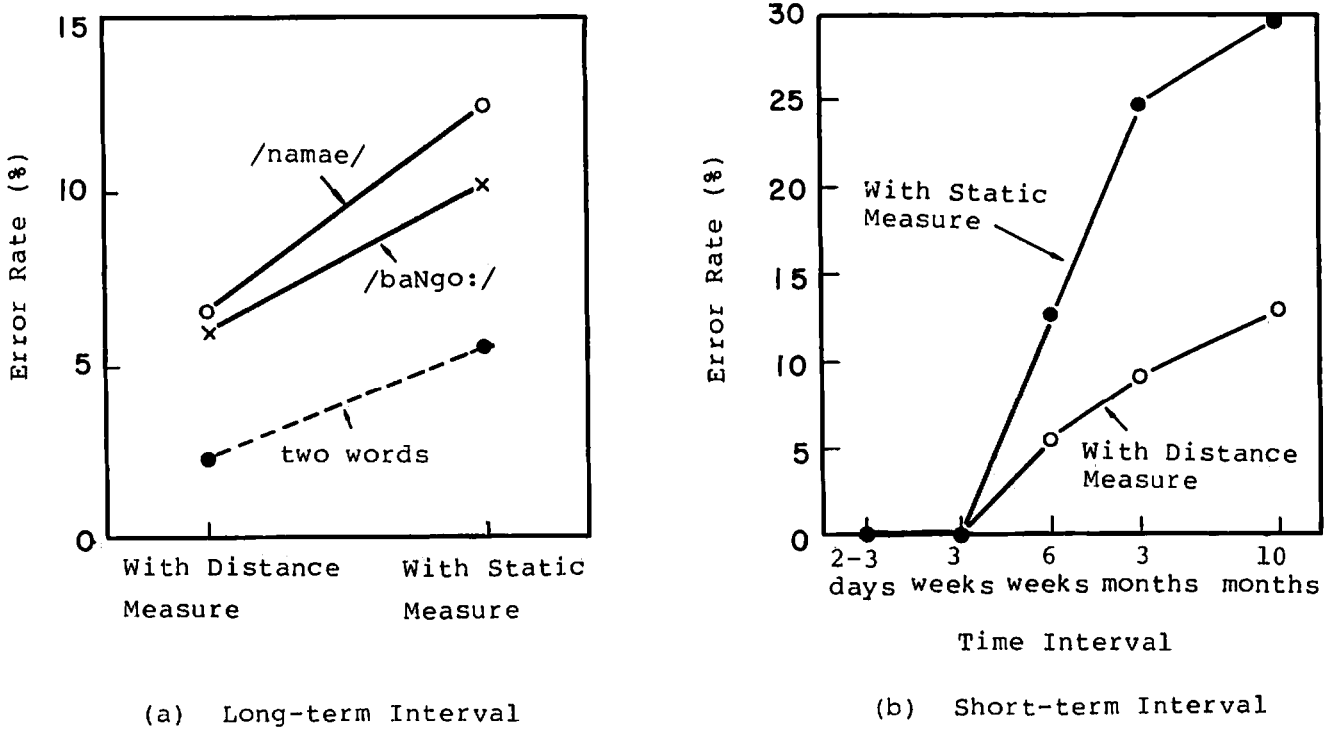

Fig. 4 Error rates of the speakeer recognition using distance measure

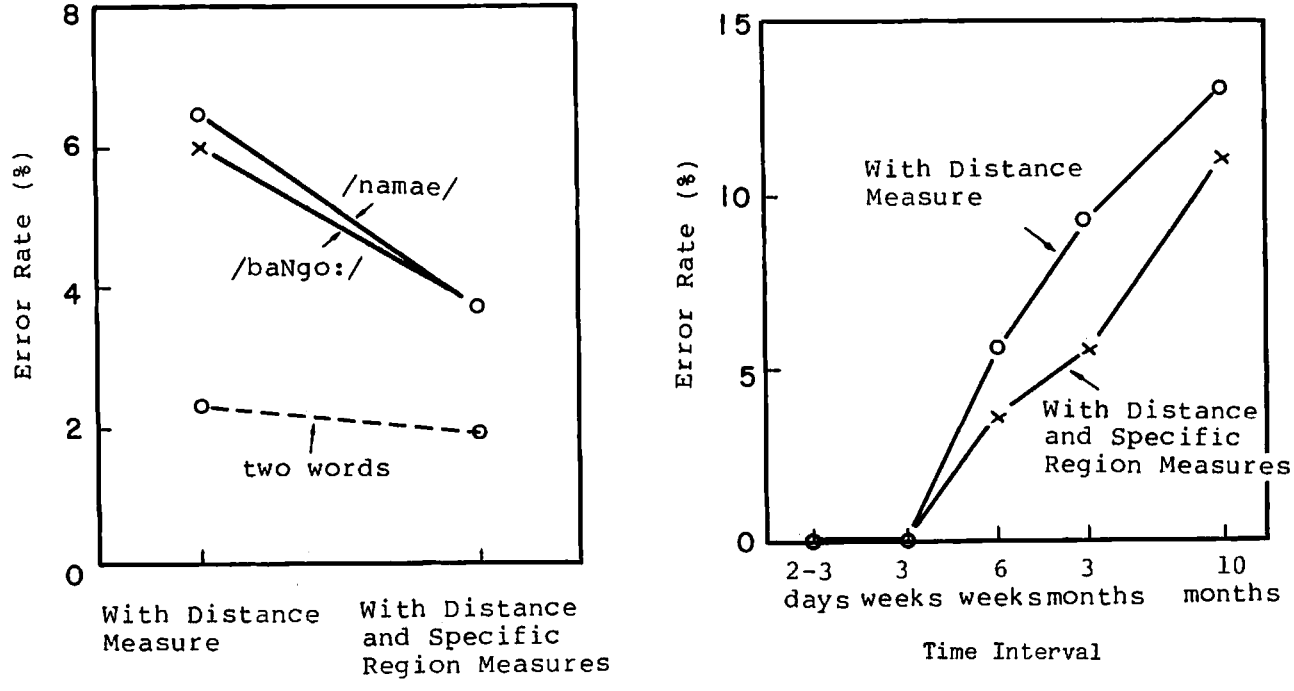

Fig. 5 Error rates of the speaker recognition using distance and specific region measures

samples used for the reference templet were uttered in a few days in the case of the short-term interval: error rates of speaker recognition increased in accordance with the increase of the time interval between the reference templet and speech input of the unknown speaker. It can also be seen that the error rates of the dynamic programming procedure decreased to about one third that of the static characteristics of the speech spectrum. It seems that speaker recognition using the dynamic programming procedure is an efficient means of reducing the temporal variation of speech input, especially in the short-term interval.

The effect of the second measure of the specific region of speech sounds on speaker recognition was tested. The rate of the matched path on the diagonal was calculated for the speech inputs of nine speakers and was used as a supplementary measure for speaker recognition. Results are shown in Fig. 5, and have been compared with those based on the distance measure only. In the case of the long-term interval shown in Fig. 5(a), it appears that the rate of speaker recognition could be improved about 2% by the use of the rate of the matched path on the diagonal for single word speech input. In Fig. 5(b), the results of the speech inputs for the short-term interval are shown. It can be seen that the second measure of dynamic characteristics was also useful in reducing the error derived from the increase of the time interval between the reference templet and speech input.

Conclusion

Summarizing our study:

(1) PARCOR parameter and fundamental frequency were used as feature parameters to analyze speaker information. The dynamic programming procedure was used for matching the reference templet and speech input of an unknown speaker. It was found that this dynamic programming procedure was more efficient than that of the static characteristics of the speech spectrum.

(2) There are several specific regions inherent to each speaker in the dynamic programming matched path, where the rate of the matched path on the diagonal is high. This specific region is useful as a supplementary measure of speaker recognition.

(3) Combining the distance measure and the specific region measure, the error rate of speaker recognition was reduced to about 2% using two test words.

References

(1) S. Furui, F. Itakura and S. Saito, "Talker recognition by the longtime averaged speech spectrum", Trans. IECE Japan, 55-A, p. 550, 1972.

(2) S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds", Trans. IECE Japan, 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.