論文 / 著書情報
Article / Book Information

Title	Effects of Long-term Spectrum Variability on Speaker Recognition
Author	SADAOKI FURUI
Journal/Book name	Proc. Joint Meeting of ASA and ASJ, Vol. , No. , pp. NNN28
Issue date	1978,

EFFECTS OF LONG-TERM SPECTRAL VARIABILITY ON SPEAKER RECOGNITION

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

Abstract-One of the most difficult problems in speaker recognition is that the feature parameters frequently vary after a long time interval. We examined this effect on two kinds of speaker recognition; one uses the time pattern of both the fundamental frequency and log-area-ratio parameters and the other uses several kinds of statistical features derived from them. Results of speaker recognition experiments revealed that the long-term variation effects have a great influence on both recognition methods, but are more evident in recognition using statistical parameters. In order to reduce the error rate after a long interval, it is desirable to collect learning samples of each speaker over a long period and measure the weighted distance based on the long-term variability of the feature parameters. When the learning samples are collected over a short period, it is effective to apply spectral equalization using the spectrum averaged over all the voiced portions of the input speech. By this method, an accuracy of 95% can be obtained in speaker verification even after five years using statistical parameters of spoken two words based on learning samples over only 10 days.

I. INTRODUCTION

Speaker recognition has recently received a great deal of attention among speech researchers. One of the most difficult problems in speaker recognition is that intersession variability for a given speaker has a significant effect on the recognition accuracy[1][2]. The author et al. have made clear the long-term variability of both statistical and dynamical speaker dependent features. Statistical features are extracted from longtime averaged spectrum[3] and time averaged characteristics of log area ratios and fundamental frequency derived from voiced portion of spoken words[4][5]. The dynamical characteristics have been analyzed by the use of dynamic programming procedure[6].

In this paper, the effects of the long-term variability of statistical and dynamical features on speaker recognition are compared with each other and the effectiveness of spectral equalization technique as a method to reduce this effect is examined.

II. METHOD OF FEATURE EXTRACTION AND SPEAKER RECOGNITION

Two kinds of Japanese spoken words /namae/ and /baNgo:/ which mean name and number respectively have been uttered repeatedly over a long period of five years by nine male speakers, and are used as test samples. At first, speech wave of each sample is transformed into a time sequence of correlation coefficients by the procedure presented in Fig.1. Spectral equalization is applied to the correlation coefficients, and log area ratios and fundamental frequency are extracted from them.

Two kinds of speaker recognition methods using these parameters are tested; one uses several kinds of statistical features derived from them, and the other uses the time pattern of them, as shown in Figs.1 and 2. Both speaker identification and speaker verification experiments with nine registered speakers are performed.

In order to examine the effect of

Acoustical Society of America and Acoustical Society of Japan Joint Meeting, NNN28

long-term variability on speaker recognition, two kinds of learning samples are collected for each speaker, and the time interval between learning samples and test samples is varied from two or three days to five years. Long-term learning samples are collected over about 10 months, and short-term ones are collected over about 10 days.

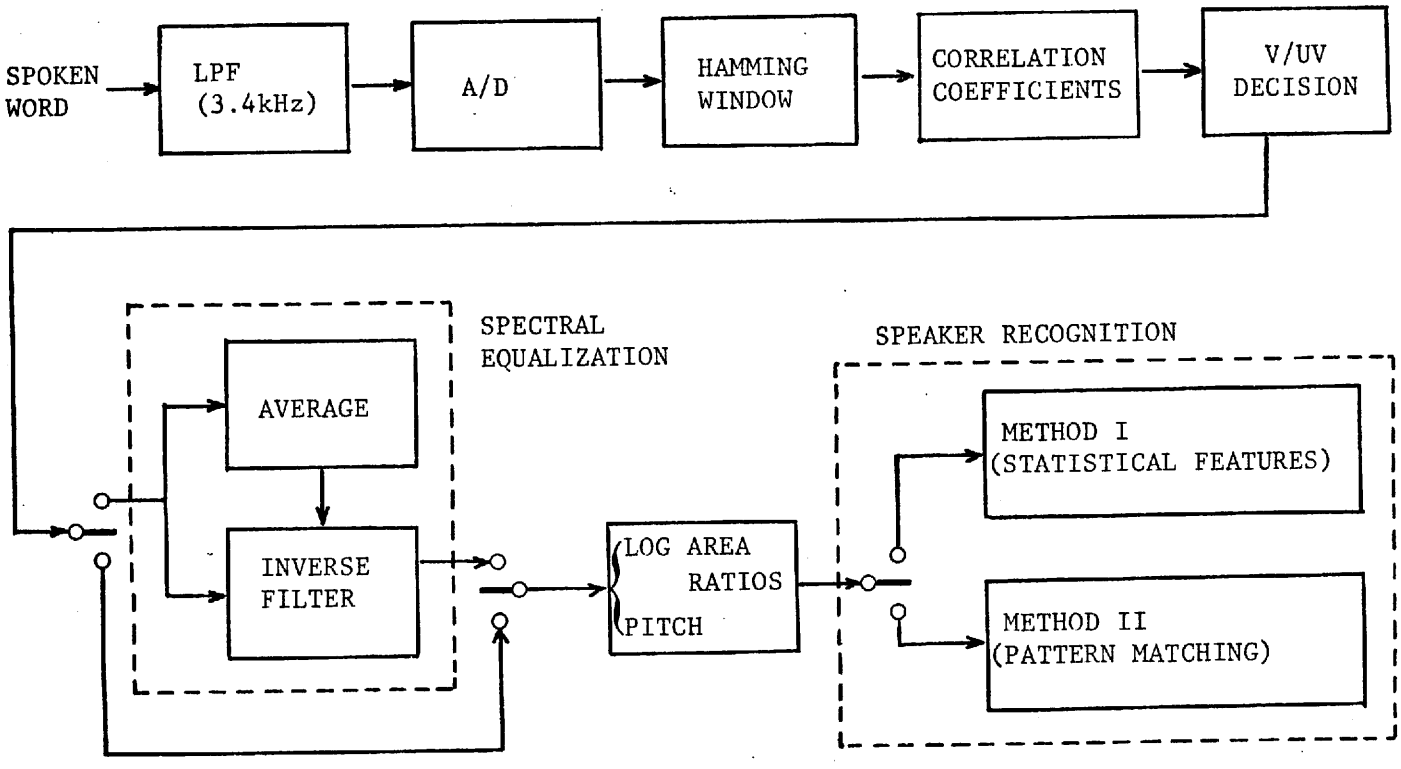

Fig.1 - Block diagram of analysis and speaker recognition.

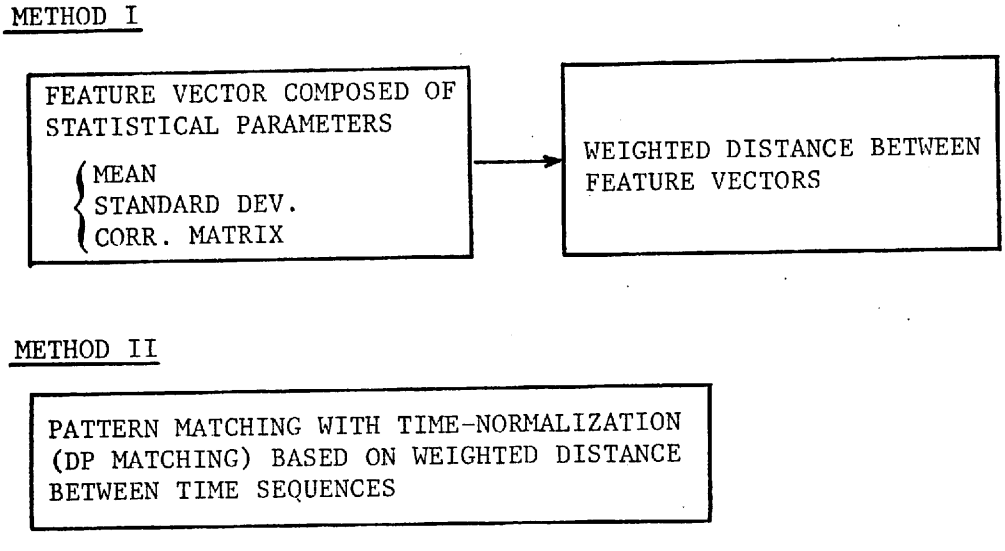

Fig.2 - METHOD I and METHOD II for speaker recognition.

III. RECOGNITION RESULTS

3.1. Comparison of METHOD I and METHOD II

Figure 3 shows the results of speaker verification experiments based on METHOD I and METHOD II for various kinds of time interval between learning and test samples. Both in METHOD I and METHOD II, the recognition accuracy decreases as the interval becomes long, especially when the length of learning period is short.

The accuracy of METHOD II is less sensitive to the length of learning period than that of METHOD I, and when the interval is about three months and the short-term learning samples are used, the accuracy of METHOD II is higher than that of METHOD I.

3.2. Effect of spectral equalization

Figure 4 shows the results of speaker verification based on METHOD I with or without spectral equalization. In the case of short-term learning, the recognition error rate is greatly reduced by the spectral equalization, and it becomes nearly the same value as that obtained in the case of long-term learning.

When the learning samples are collected over a long period, the effectiveness of the spectral equalization can not been seen. The result makes a difference whether the spectral equalization is applied or not, when the interval between learning and test samples becomes one year or more, although it is not presented in Fig.4.

3.3. Effect of words combination

When the feature parameters extracted from different spoken words are combined and used for recognition, the recognition accuracy becomes much higher than that obtained by one word.

Figure 5 shows the results of speaker verification with spectral equalization procedure in which time interval between learning and test samples is varied from three months to about five years. When two words are combined, the error rate is reduced to nearly half of that obtained by one word. The recognition accuracy by two words using long-term learning samples is about 99% and 96% after one year and after five years respectively, and the recognition accuracy by one word is about 96% and 94% on the same conditions. Even when the short-term learning samples are used, an accuracy of about 95% can be obtained after the interval of five years using two words with spectral equalization.

Figure 6 shows the results of speaker identification on the same condition as Fig.5. When the feature parameters of two words are combined, the recognition accuracy of about 93% can be obtained after five years using the long-term learning samples.

IV. CONCLUSION

Based on the above mentioned experimental results, following methods can be proposed to cope with the long-term spectral variability in speaker recognition.

(a) Collect learning samples over a long period.
(b) Renew learning samples continuously.
(c) Apply spectral equalization.
(d) Use dynamical feature parameters.
(e) Combine feature parameters of several words.

When the methods (a), (c) and (e) are applied, 99% and 96% verification accuracy can be obtained after one year and after five years respectively, based on METHOD I using two words.

Speaker recognition experiment based on the combination of the methods (c) and (d) (METHOD II) is now under consideration.

ACKNOWLEDGEMENT

The author wishes especially to acknowledge the guidance provided by Dr. Shuzo Saito, Director of Saito Research Section, during the course of this study, and he also expresses his thanks to other members of the section for constructive discussion.

REFERENCES

[1] B.S.Atal: Proc.IEEE,64,4,p.405,1976
[2] A.E.Rosenberg : ibid.,p.475
[3] S.Furui,F.Itakura and S.Saito: Trans. IECE Jap.,55-A,10,p.550,1972
[4] S.Furui and F.Itakura : ibid.,56-A,11, p.717,1973
[5] S.Furui: ibid.,57-A,12,p.880,1974
[6] S.Saito and S.Furui :4th IJCPR,p.1014, 1978

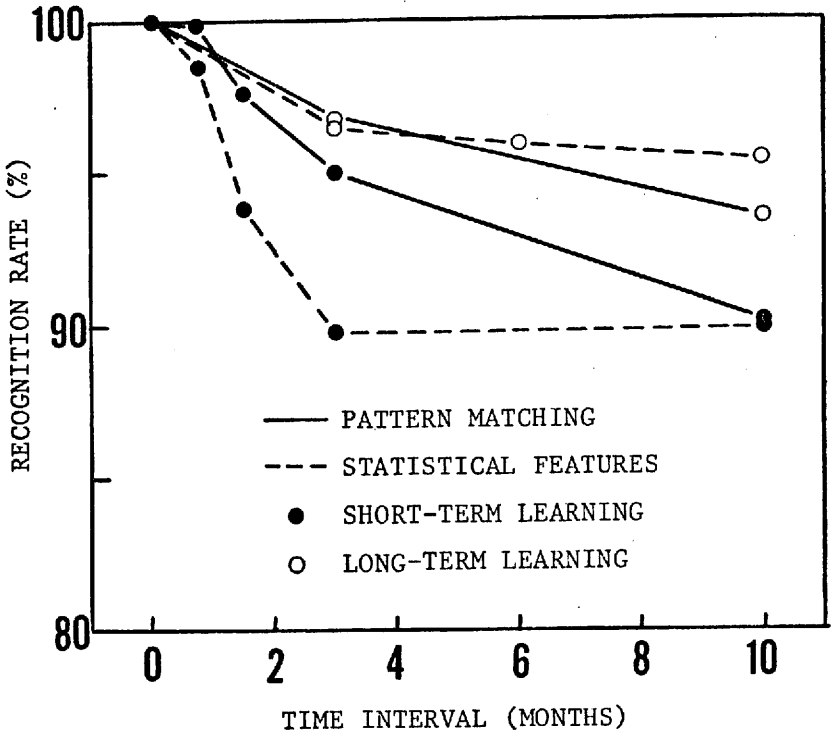

Fig.3 - Comparison between the results of the two kinds of recognition methods; pattern matching and statistical features.

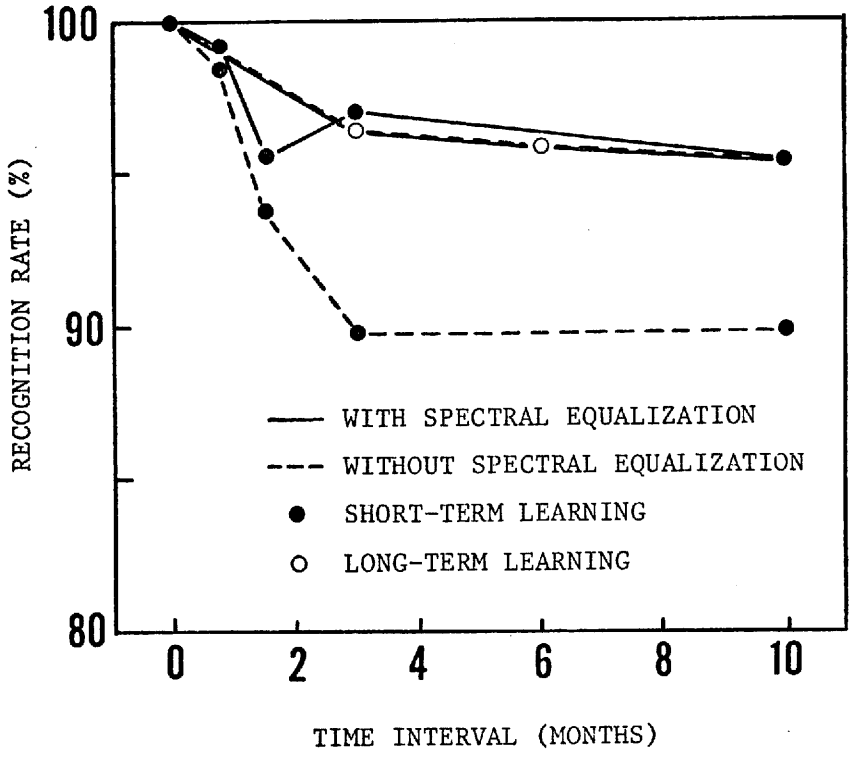

Fig.4 - Effectiveness of spectral equalization on speaker verification.

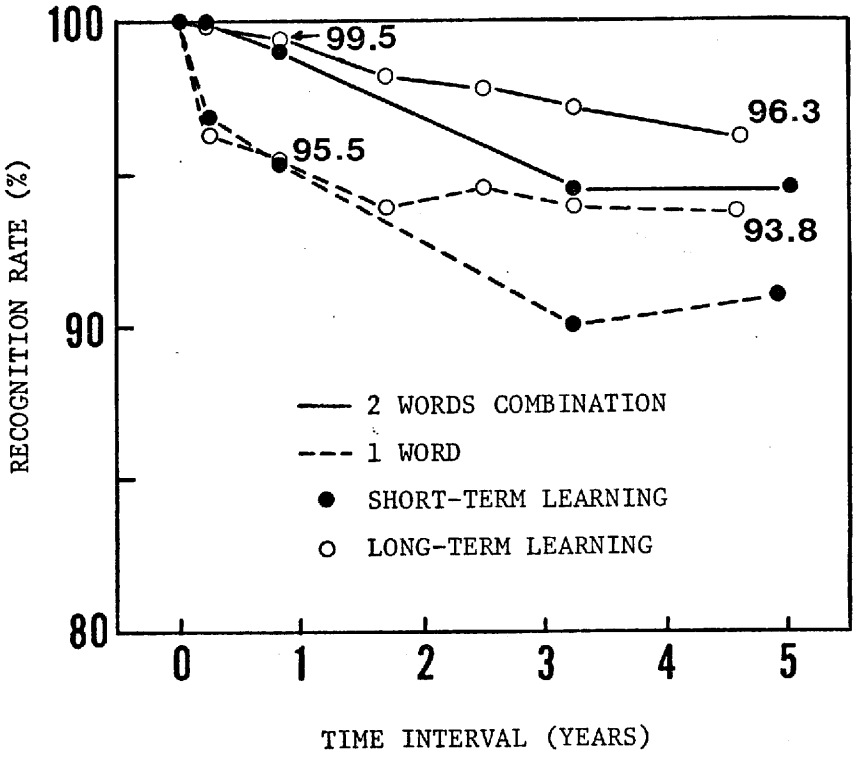

Fig.5 - Results of speaker verification using one word or two words with spectral equalization after long time interval.

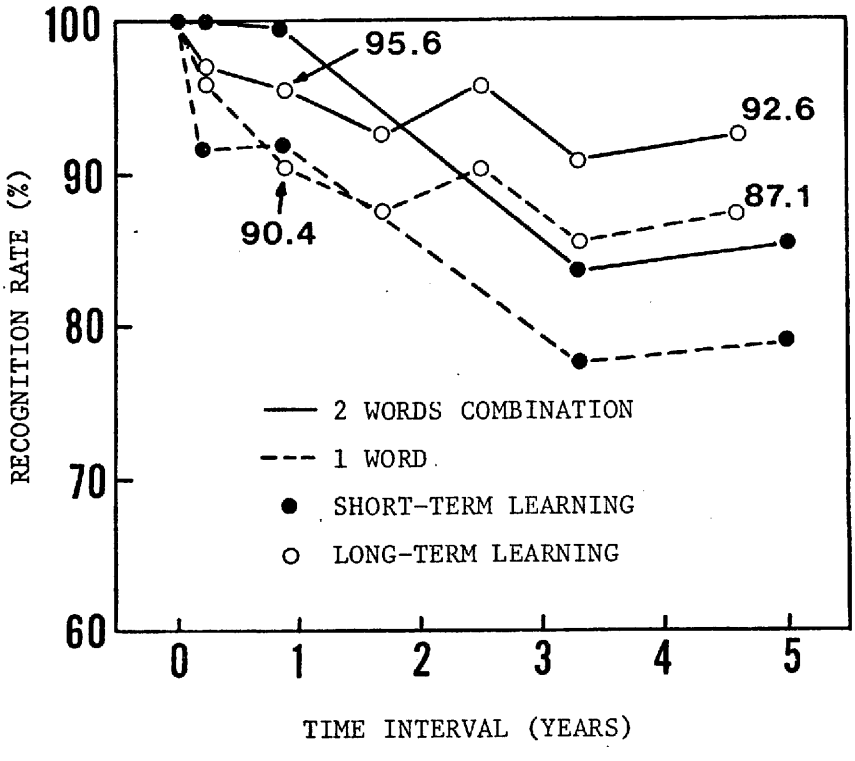

Fig.6 - Results of speaker identification using one word or two words with spectral equalization after long time interval.