

論文 / 著書情報
Article / Book Information

論題(和文)	
Title	Effect of Long-term Variations on Talker Recognition by Averaged Speech Spectrm
著者(和文)	古井貞熙
Author	SADAOKI FURUI, Fumitada Itakura, Shuzo Saito
出典(和文)	, Vol. 28, No. 8, pp. 414-415
Journal/Book name	J. Acoust. Soc. Jpn., Vol. 28, No. 8, pp. 414-415
発行日 / Issue date	1972, 7

寄 書

日本音響学会誌 28 卷 8 号 (1972)

Effect of Long-term Variations on Talker Recognition
by Averaged Speech Spectrum*Sadaoki FURUI, Fumitada ITAKURA and Shuzo SAITO**
(Musashino Electrical Communication Laboratory, N. T. T.)

Long-term variations of speech spectral energy patterns are studied from the viewpoint of talker recognition. The averaged frequency spectrum of speech uttered by each of 9 talkers is investigated during a period of more than 18 months, and the effect of its long-term variations on a talker recognition is examined.

Speech signal is recorded in a sound-proof room, and is bandlimited to 3.4 kHz, sampled every 1.25×10^{-4} seconds, and converted into a 11-bits value. The averaged frequency spectrum of a talker is obtained by analyzing an utterance of about 10 seconds. In order to treat overall and detailed structures of the spectral pattern separately, the averaged frequency spectrum is converted into the cepstrum, the Fourier transform of the logarithmic power spectrum.

The recognition of the talker identity is made on the basis of various distances in a multidimensional space of the cepstrum as defined in the following formulae.

1) Weighted distance between the unknown sample and the mean of reference samples of each talker.

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

where C : cepstrum of the unknown sample (N -dimensional vector)

D_{rj} : cepstrum of j -th reference sample of r -th talker

$\bar{D}_r = \frac{1}{k} \sum_{j=1}^k D_{rj}$: mean of reference samples by r -th talker

k : the number of reference samples of individual talker

V_r : covariance matrix of cepstrum for r -th talker

Equation (1) implies that the matrix V_r be nonsingular, and this requires the condition $k \geq N$ to be satisfied. Since it is very difficult to satisfy this condition when N is large, the following matrices are used in place of V_r .

$$[\text{Method I}] V_{Ir} = \begin{pmatrix} V_{r11} & & 0 \\ & V_{r22} & \\ 0 & & \ddots \\ & & & V_{rNN} \end{pmatrix} \quad (2)$$

$V_{r ii}$: diagonal elements of the matrix V_r .

$$[\text{Method II}] V_{II} = \frac{1}{m} \sum_{r=1}^m V_{Ir} \quad (3)$$

m : the number of talkers

$$[\text{Method III}] V_{III} = \frac{1}{m} \sum_{r=1}^m V_r \quad (4)$$

2) Nearest-neighbor method [Method IV]

$$S(C, \{D_{rj}\}) = \min_j \|C - D_{rj}\| \quad (5)$$

3) Squared Euclidian distance between the unknown sample and the mean of the reference samples [Unweighted]

$$S(C, \{D_{rj}\}) = \|C - \bar{D}_r\| \quad (6)$$

In the case of talker identification, the unknown talker is identified on the basis of proximity to reference samples founded on either one of these distances, while in the case of talker verification a two-class recognition is made as to the talker's purported identity. In the talker identification experiment, a population of 9 male talkers are used, and in the verification experiment 8 or 26 casual impostors are pitted against 9 true talkers. Dimension of the cepstrum N is selected from 21 to 64 for each recognition method.

Two kinds of talker recognition experiments are

* 長時間平均スペクトルによる話者認識における経時変化の効果

** 古井真昭, 板倉文忠, 齊藤収三 (日本電信電話公社武蔵野電気通信研究所)

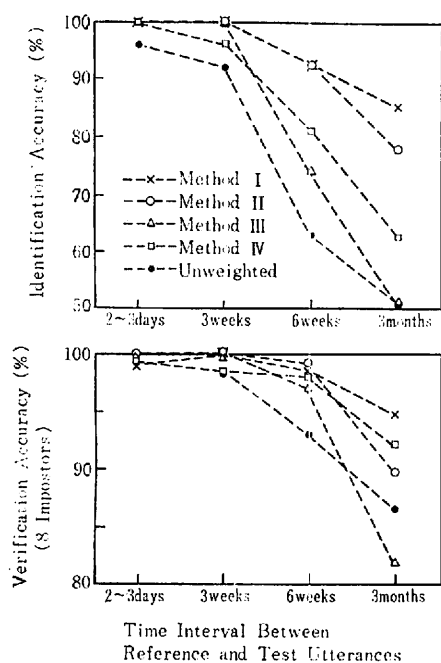


Fig. 1 Identification and verification accuracy. Reference utterances are collected over a short period (about ten days).

-- x -- Method I
 -- o -- Method II
 -- Δ -- Method III
 -- □ -- Method IV
 -- • -- Unweighted

made to examine the effect of the number of reference samples and the period over which these references are collected, and to see the effect of the time interval between reference and test samples. In the first experiment, the reference samples are collected in three or four sessions over a period of about ten days, and the time interval between the last reference sample and the test sample is varied from several days to three months. In the second experiment, the reference samples are collected in one to six sessions at intervals of about three months and the time interval between the last reference sample and the test sample is three months.

Results of talker recognition for 9 talkers are shown in Figs. 1 and 2. From these results, the following conclusion is obtained for the case where the reference samples are collected at one sitting or

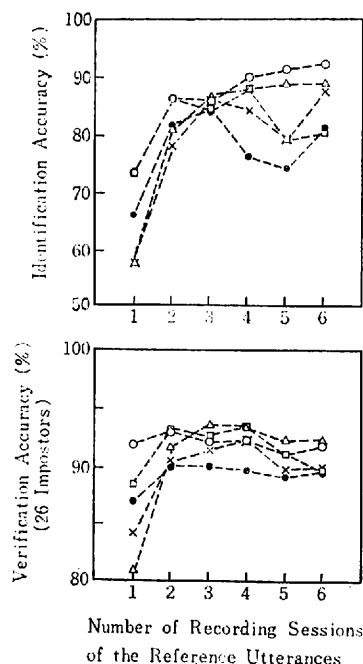


Fig. 2 Identification and verification accuracy. Reference utterances are collected over a long period, and the time interval between reference and test utterances is three months.

-- x -- Method I
 -- o -- Method II
 -- Δ -- Method III
 -- □ -- Method IV
 -- • -- Unweighted

over a short period of time. When the time interval between the reference and test samples is short, the accuracy of talker recognition is very high, but when the time interval becomes longer, the accuracy drops even by the use of weighted distances. When the reference samples are collected over a period of about a year, however, very high accuracy is obtained for test samples taken three months after the last reference sample. In such a case an accuracy of 93% is obtained in separating a talker from impostors, and 90% for the identification.

The differences of the recognition rate among individual talkers is fairly significant, and verification errors between talkers are highly correlated with identification errors. Among the four methods of weighting, Method II and III lead to better results.