# TEXT-INDEPENDENT SPEAKER IDENTIFICATION BASED ON PIECEWISE CANONICAL DISCRIMINANT ANALYSIS

Hiroshi Matsumoto

Department of Electronic Engineering,
Faculty of Engineering, Shinshu Univ.
Nagano, 380, Japan

Tadamoto Nimura

Department of Electronic Engineering,
Faculty of Engineering, Tohoku Univ.
Sendai, 980, Japan

ABSTRACT This paper describes a method for text-independent speaker identification. In this method, in order to utilize phoneme-dependent personal information in addition to personal information common to all phonemes, multiple personal factor spaces are constructed by applying canonical discriminant analysis to the predetermined subspaces in the observation space. The decision is based on a liklihood measure derived from a posteriori probabilities in all the factor spaces. Using the 21-dimensional observation vectors obtained from every 40 msec voiced segments, the methods of construction of the subspaces and others were examined. An identification accuracy comparable to human listeners was achieved.

## I. INTRODUCTION

In the past few years, a great deal of research has been directed toward automatic speaker recognition. In most researches, the test utterances have been restricted to the same phrase with reference utterance. However, it is a common experience that we can often identify the familiar voices with high accuracy from a brief utterance irrespective of the text being spoken.

The personal information utilized by human listener can be broadly calssified into the two groups: the phoneme-independent and the phoneme-dependent information. In most studies on text-independent speaker recognition reported so far, the recognition scheme has utilyzed the former information which is extracted from a variety of parameters through several statistical methods, such as average operation [1] or principal component analysis [2], [3]. However, these procedures generally has needed relatively long utterances to perform successful speaker recognition. On the other hand, only a few attempts to intend to utilize the latter personal information as well as the former has been reported in past [4].

In this paper, in order to acomplish a successful speaker identification even with a brief utterance like human listener, we will examine a method to utilize the both types of personal information, in which canonical discriminant analysis is applied to the predetermined subspaces in the observation space.

## II. BACKGROUND

If we consider both the speaker and the phoneme factor as the static information contained in voiced speech, an observation vector $\mathbf{x}$ can be expressed by a simple model associated with a two-way analysis of variance,

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{x}_s + \mathbf{x}_p + \mathbf{x}_{sp} + \mathbf{e} \quad . \qquad (1)$$

where $\bar{\mathbf{x}}$ is the mean vector over all observation vectors; $\mathbf{x}_s$ is the main factor of speaker which contains phoneme-independent personal information; $\mathbf{x}_p$ is a main factor of phoneme; $\mathbf{x}_{sp}$ is a interaction term between speaker and phoneme which contains phoneme-dependent personal information; and $\mathbf{e}$ is a residual term which involves variation due to changing emotions, states of health, and so on. Actually, according to a preliminary experiment based on multivariate analysis of variance, the interaction term $\mathbf{x}_{sp}$ was found to be not so large as the main factor $\mathbf{x}_s$ but still highly significant statistically.

In previous studies, in order to extract only phoneme-independent personal information from $\mathbf{x}_s$, $\mathbf{x}_p$ and $\mathbf{x}_{sp}$ have been treated as a part of the residual term in a linear model of one-way classification. Consequently, the identification accuracies were not satisfactory for a brief utterance, partly because it is necessary to average the parameter to cancel the variation caused by $\mathbf{x}_p$ which accounts for the largest part of a total variance of $\mathbf{x}$, partly because the phoneme-dependent term $\mathbf{x}_{sp}$ was not utilized.

Then, if each factor in Eq.1 can be exactly separated from an observation vector, we will be able to achieve an identification accuracy as high as in text-dependent situations. However, in actual text-independent situations, it is next to impossible because phoneme identification in connected speech is very difficult due to coartidulation effects.

Therefore, if we roughly divide the observation space into several subspaces $\Omega^k(k=1,..,q)$ which might include the similar phoneme-dependent personal information, $(\mathbf{x} + \mathbf{x}_p)$ and $(\mathbf{x}_s + \mathbf{x}_{sp})$ in Eq.1 can be considered roughly as the mean vector $\bar{\mathbf{x}}^k$ and the speaker factor $\mathbf{x}_s^k$, respectively, for the observation vectors $\mathbf{x}^k$ belonging to the subspace $\Omega^k$. Consequently, Eq.1 may be replaced by the q linear models associated with one-way analysis of variance,

$$\mathbf{x}^k = \bar{\mathbf{x}}^k + \mathbf{x}_s^k + \mathbf{e}^k \quad . \qquad (\mathbf{x}^k \in \Omega^k) \qquad (2)$$

Then, as illustrated in Fig.2, if the multiple personal factor spaces $\Theta^k$ (k=1,..,q) are derived by

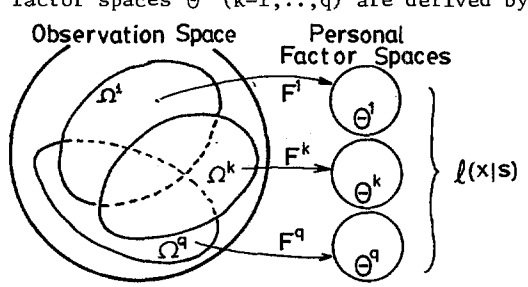


Fig. 1. Illustration of "Piecewise Canonical Discriminant Analysis.

CH1285-6/78/0000-0291$00.74@1978IEEE

applying canonical discriminant analysis to each subspace $\Omega^k$, it may be possible not only to utilize phoneme-dependent personal factor and also to reduce the influence of variation due to phoneme factor. The important points in this idea are: (1) how to construct a set of subspaces; (2) how to combine the multiple measures of similarity between an input test vector and the references in each $\Theta^k$ into a single measure. We will call this method "Piece-Wise Canonical Discriminant Analysis (PCDA)," and examine a concrete procedure in the subsequent sections.

## III. IDENTIFICATION PROCEDURE

### A. Learning Procedure

Suppose the q subspaces $\Omega^k$(k=1,..,q) has been defined in the r-dimensional observation space in advance. We first estimate the mean vector of the $s$th speaker $\hat{u}_s^k$ (s=1,..,p), an interspeaker covariance matrix $\hat{\Sigma}_B^k$, and an intraspeaker covariance matrix $\hat{\Sigma}_w^k$ from the input observation vectors belonging to $\Omega^k$. Then, to estimate the mapping matrix $\hat{F}^k$ from $\Omega^k$ to a personal factor space $\Theta^k$, we wish to maxmize an interspeaker variance Trace$\{\hat{F}^{k'}\hat{\Sigma}_B^k\hat{F}^k\}$, under the constraint that an intraspeaker variance matrix $\hat{F}^{k'}\hat{\Sigma}_w^k\hat{F}^k$ in $\Theta^k$, is equal to an identity matrix $I$. Consequently, the $i$th row vector of $\hat{F}^{k'}$ is given by the corresponding eigenvector from the characteristic equation

$$\{\hat{\Sigma}_B^k - \lambda^k\hat{\Sigma}_D^k\}\cdot f^k = 0 \quad . \tag{3}$$

The above procedure is carried out for each of the q subspaces. The identification scheme is constructed in terms of the q sets of $\hat{u}_s^k$(s=1,..,p) and $\hat{F}^k$.

### B. Decision Procedure

In order to derive a single measure of similarity between a test vector $x$ and the multiple references of the $s$th speaker, if $x$ is closest to the prototype point (typically, the center of subspace) of $\Omega^k$, we define a liklihood measure of the $s$th speaker as

$$\ell(x|s) = \frac{1}{\sqrt{2\pi}}\cdot\exp\{-\frac{1}{2}\cdot\|\hat{F}^k(x-\hat{u}_s^k)\|^2\} \quad . \tag{4}$$

and assigned $x$ to the speaker giving the largest $\ell(x|s)$.

When multiple observation vectors $\{x_1, x_2, .., x_n\}$ are obtained from an unknown utterance, we define the two types of liklihood measure as follow;

$$L(x_1, x_2, \ldots, x_n|s) = \sum_i \ell(x_i|s), \tag{5}$$

$$L(x_1, x_2, \ldots, x_n|s) = \prod_i \{\ell(x_i|s) + \ell_0\}. \tag{6}$$

Eq.5 implies the average probability, and Eq.6 implies the probability that all the test vectors belong simulteneously to the $s$th speaker under the condition that $x_1, \ldots, x_n$ are statistically independent. The $\ell_0$ in Eq.6 is a small constant to avoid the influence of a few small $\ell(x|s)$ to the whole $L$ and will be examined experimentally.

## III. DATA BASE

### A. Speech Materials

In this study, the two bodies of speech materials were involved. The first set of speech, which was used for construction of a phoneme factor space, consisted of six repetitions of the five Japanese vowels (/i/, /e/, /a/, /o/, and /u/) and three repetitions of the three nasals (/m/, /n/, and /ŋ/) embeded in the 15 cvcv disyllables which were all the combinations of the five Japanese vowels and the three nasals, for example, /mama/. The second body of materials, which was used for speaker identification experiments, consisted of the four reading sentences about general weather condition. The first three sentences were used as design sets and the last sentence provided test utterances.

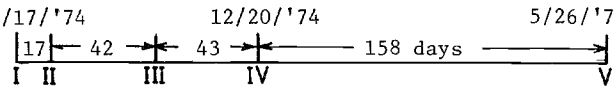


Fig. 2. The schedule for the five recording sessions.

These speech materials were spoken by 10 male speakers ranged in age from 22 to 38 years, none of whom had a noticeable speech defect. These two bodies of materials were recorded in an unechoic chamber in the same five sessions, which spread over a period of nine months as illustrated in Fig.2.

### B. Observation Vector and Analysis

In this study, the 20-dimensional cepstral components $c_1, .., c_{20}$ and an average pitch $F_0$ of a short segment of voiced speech were chosen as the components of an observation vector. The duration of segment was chosen to be 40 ms long, so as to avoid containing several different phoneme events in it.

In analysis, the speech signals were digitized into 11-bit binary numbers at sampling frequency of 10 KHz after low pass filtering at 4.8 KHz. Then, the voiced portions in utterance were automatically detected by an algorithm based on amplitude and duration of silent portions and poses, and, after check by listening, the errors in detection were modified manually based on a waveform displayed on a CRT. The cepstral analysis was conducted every 20 ms with 40 ms Hamming window, and a first-order backward difference for pre-emphasis by means of 512-point FFT algorithm.

In the case of nasals, two frames were manually excerpted from each disyllable. The four sentences yielded the average frames of 104, 100, 85, and 177, respectively, for each speaker.

## V. CONSTRUCTION OF SUBSPACES

For the purpose of reducing the variation caused by phoneme as well as extracting the phoneme-dependent personal information, it may be resonable to classify the input vectors on the basis of the vectors mapped onto the phoneme factor space in stead of the observation space itself. Although we have a great choice of subspace, we restricted the subspaces to the eight spheres in the phoneme factor space, of which the centers correspond to the centroids of the phonemes /i/, /e/, /a/, /o/, /u/, /m/, /n/, and /ŋ/, and, then, we examined the optimum radius.

First, we derived the phoneme factor space by means of canonical discriminant analysis from the 300 observation vectors for each of the eight phonemes, which were recorded in the same session with the design set. The identification accuracy of a single test vector belonging to each phoneme region as a function of the radius of each subspace was examined. In this experiment, only the speech data recorded in the session I were used, and the design set consisted of the first two sentences. Since $F_0$ is considered Phoneme-independent, in this paticular section, the effect of subdivision of the observation space was examined using only the 20 cepstral compornents.

The results are shown in Fig.3. It is found that the identification accuracy for every phoneme region is improved at small radius compared with that

for the infinite radius, which means an usual canonical discriminant analysis (CDA). Above all, the accuracy for the regions of /i/, /e/, /m/, and /n/ is improved at smaller radius than others. In paticular, the improvement of accuracy for the region /i/ is reached to 20%. This suggests that the phoneme regions in which the place of articulation is front will contain somewhat different personal information from other regions.

On the basis of the above results, we determined the radius of the subspaces corresponding to /i/, /e/, /a/, /o/, /u/, /m/, /n/, and / / to the radiuses of 5, 6, 6, 7, 6, 5, 5, and 6, respectively, which gave the highest accuracy for each phoneme region. Since the mean distance among the centroids of the five vowels was 7.9, these subspaces are extensively overlapped each other. The overall identification accuracy for a single unknown vector was 66% for CDA, while that increased to 74% for PCDA using the above eight subspaces. Then, although it may not be always optimum, we will use the set of subspaces determined above in subsequent experiments.

## VI. IDENTIFICATION EXPERIMENTS

### A. Identification based on Multiple Test Vectors.

In this experiments, the same speech data with the Sec.V were used. In order to prepare the test utterances of various durations, the sequences of contiguous speech frames of various numbers were excerpted from the sequence of test vectors obtained from the 4th sentence of each speaker by shifting the begining two frames by two frames.

Initially, it was found that the highest identification accuracy can be achieved when the value of in Eq.6 is on the order of $10^{-6}$. Then, the results of identification experiments based on both Eq.5 and 6 for the two values of $\ell_0$, 0 and $10^{-6}$, is showen in Fig.4. It is clear that Eq.6 is superior to Eq.5, especially for short speech, and is very effective for long speech. Consequently, the identification accuracy based on Eq.6 with $\ell_0=10^{-6}$ dramatically increased to 90% for a duration of 0.1 s and reached to 100% for a duration of 1.4 s. Therefore, the following experiments will be based on Eq.6 with $\ell_0=10^{-6}$.

### B. Examination of the Number of Training Samples.

In this experiments, the speech data in the session I were used again. Fig.5 whows the results of identification experiments based on both PCDA and CDA using the first three sentences. The identification accuracy for CDA gradually increases with the number

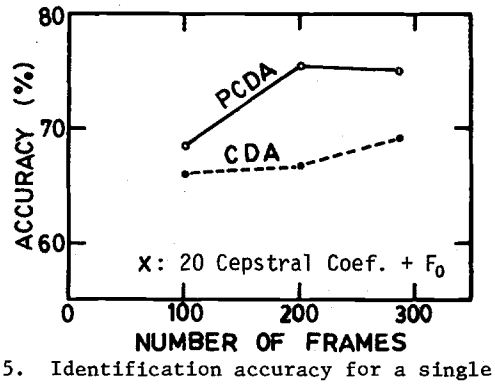


Fig. 5. Identification accuracy for a single test vector as a function of the number of observation vectors from the design set for each speaker.

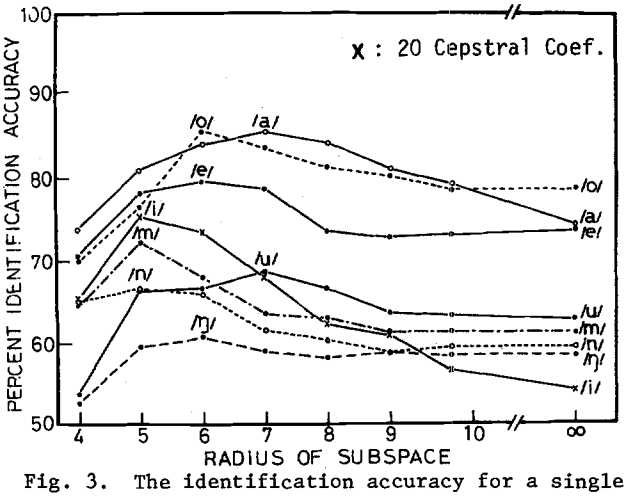


Fig. 3. The identification accuracy for a single test vector belonging to each phoneme region as a function of the radius of corresponding subspace.

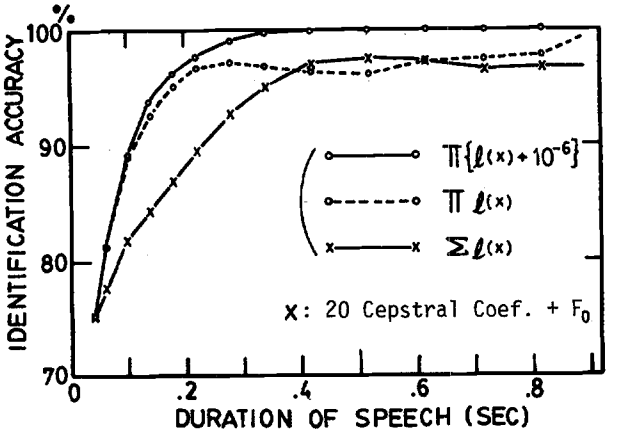


Fig. 4. Comparison of the two types of liklihood measures for identification accuracy.

of design vectors, while that for PCDA rapidly increases and tends to reach a limit beyond 200 fames. This result implies PCDA can estimate the personal informations from fewer speech samples than CDA, because the intraspeaker variation caused by phoneme is smaller in the subspaces than in the whole space. Therefore, we can conclude an adequate duration of learning speech is about s for each speaker.

### C. The Effects of Inter-Session Variation

First, the test utterances in each of the five sessions were identified based on PCDA using the first two sentences in the session I. As shown Fig.6, the effect of inter-session variance is so strong as to reduce the identification accuracy for the test utterances recorded on the different sessions from the design set by 20% to 30% depending on duration or session. The systematic effects of the time interval, however, are not observed for our speech materials, except for less accuracy for session IV.

Then, when the further speech data for other sessions were combined to the design set, an identification accuracy was found to increase gradually. The experimental results based on both PCDA and CDA using the design sentence in the first three session

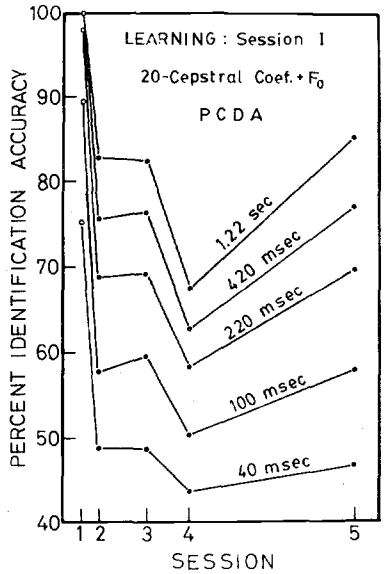


Fig. 6. Effects of time interval between recordings of the design and test speech on the identification accuracy.

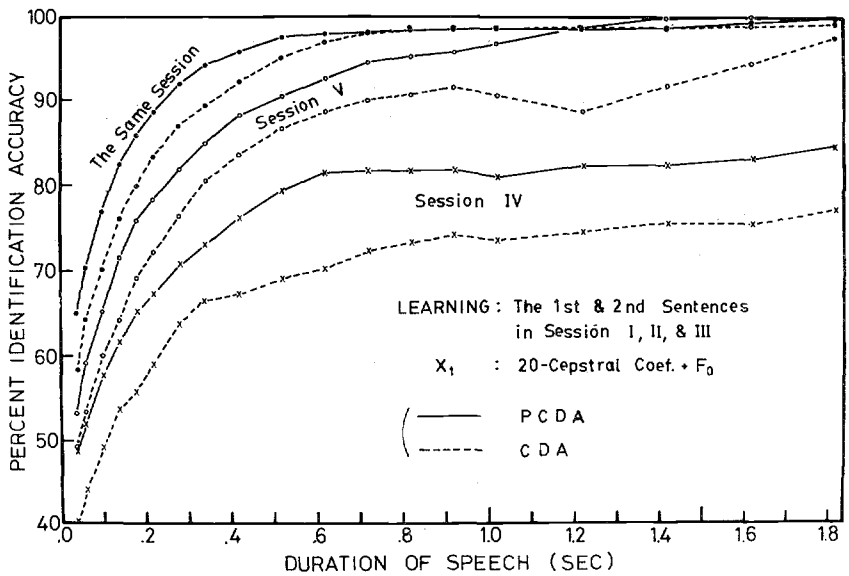


Fig. 7. The identification accuracy as a function of speech duration based on PCDA and CDA, where the design set consists of the first two sentences in the first three sessions.

were shown in Fig. 7. The improvement of identification accuracy by PCDA is found to be larger in the case of lower accuracy, that is, for shorter duration and for the test samples in the session IV and V. Then, for the PCDA, the identification accuracy for the same session with the design set, which is averaged over the results of the first three sessions, is increased to 90% for a duration of 0.25 s , and exceeds aproximatly 98% for 0.5 s, and reaches to 100% for 1.8 s. Furthermore, the results for the test samples of the session V, in spite of the interval of a half year between recordings of the test and design samples, shows a high identification accuracy, which increase from 53% for 40 ms to 90% for o.5 s, and reached to 100% for 1.4 s.

These identification scores in comparison with performance of human listener reported by Briker and Pruzansky are shown in Table I. Although it is impossible to compare exactly because of the differences in speakers, materials and others, the method proposed, PCDA, achieved the high accuracy comparable to the best listener even for a brief utterance.

So far, the pitch has been contained as a compornent of an observation vector. Finally, an additional experiment in which only the 20 cepstrum compornents were used was conducted. As a result, the identification accuracy was found to decrease by only a few percent for the case of lower accuracy but to be unchanged for the case of higher accuracy. Therefore, we can conclude that the identification accuracy attained so far were owing mostly to the spectral information.

## VII. CONCLUSION

It was found that phoneme-dependent personal information can be utilized in text-independent speaker kdentification by applying CDA piecewesely to the observation space. In the result, the proposed method could attain a high accuracy comparable to humman listener.

Table I. Comparison of the identification accuracy in the text-independent speaker identification based on PCDA with the performance by human listener [5].

| | duration (msec) | Accuracy in % Human listener Max. | Human listener Mean | PCDA Same | PCDA 5th |
|---|---|---|---|---|---|
| Sentences | 2400 | 100 | 98 | 100 | 100 |
| Disyllables | 446 | 98 | 87 | 96 | 89 |
| Monosyllables | 498 | 94 | 81 | 97 | 90 |
| CV excerpts | 117 | 83 | 63 | 79 | 73 |
| Vowel excerpts | 117 | 75 | 56 | | |

ACKNOWLEDGEMENT

The authors would like to thank Prof. K. Kido and Assistant Prof. S. Hiki for their valuable cooporation and for use of the computer facilities.

REFERENCE

[1] S. Furui, et al, "Talker recognition by longtime averaged speech spectrum," Electron. Commun. Jap., vol. 55A, pp. 54-61, Oct. 1972.

[2] B.S.Atal, "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., vol. 55, pp. 1304-1312, June 1972.

[3] M.R.Sambur, "Speaker recognition using orthogonal linear prediction," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 238-239, August 1976.

[4] R.L.Kashyap, "Speaker recognition from an unknown utterance and speaker-speech interaction," IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-24, pp.481-488, December 1976.

[5] P.D.Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Amer., vol.40, pp.1441-1449, June 1966.