STUDIA PHONOLOGICA XXI (1987)

Correlation Analysis of Speaker Differences in Vowels, Consonants and Spoken Digits

Seiichi NAKAGAWA

Abstract

The speaker differences are divided into two kinds. One is inter-group differences—speaker differences in age and sex. The other is intra-group differences—speaker differences in the same generation and sex. The former is the physical differences of the apparatus (most of hardware differences). The latter is further divided into two types. They are the minute differences of articulators (part of hardware differences) and the differences of linguistic environments or articulation (software differences).

In this paper, we investigate the property of intra-group speaker differences (adult males) caused by hardware-factor and software-factor through correlation analyses between voices on speaker-factor.

From experiments we find the following: 1) the speaker differences caused by physical differences in vowels, consonants and spoken digits are not random (structural), 2) the speaker differences of articulation in spoken digits are also not random, 3) the speaker differences of a spoken digit in which a vowel may be uttered by the manner of devocalization are influenced by the differences of artiuclation, and 4) the correlation between voiced consonants except /z/ is large, but that between a voiced consonant and an unvoiced consonant except plosive consonants is small.

1. Introduction

For automatic recognition of continuous speech, we must solve very difficult problems such as segmentation, coarticulation, speaker differences, word juncture, prosody and so on. In this paper, we consider the problem of speaker differences.

The speaker differences are divided into two kinds. One is inter-group differences—speaker differences within age and sex.[1] The other is intra-group differences—speaker differences in the same generation and sex. The former is the physical differences of the apparatus (most of hardware differences). The latter is further divided into two types. The first one is the minute differences of articu-

Seiichi NAKAGAWA (中川聖一): Associate Professor, Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, 440 Japan.

The author has done the research under the direction of Dr. Toshiyuki Sakai, Professor of Information Science, Kyoto University.

lators (part of hardware differences) and the second one is the differences of linguistic environments or articulation (software differences).

However, it is very difficult to separate the speaker differences into the differences caused by hardware-factor and software-factor, and also we think that there is no uniform normalization technique for both differences of inter-group and intra-group, or both differences caused by hardware-factor and software-factor.

For Japanese vowels, Matsumoto et al. investigated vocal individualities among different vowels through listening tests of speaker verification.[2] They obtained the correct rate around 60% by using the stimulus of a pair of different vowels. This implies that there exist vowel-independent vocal individualities in vowel spectra. From different view points, Sakai and Tabata showed this fact through a variance analysis of vowel spectra.[3] Shikano and Sugamura stated that the speaker differences in spoken words are mostly caused by the differences of articulation, but not spectral differences of each phoneme.[4]

In this paper, we investigate the property of intra-group speaker differences caused by hardware-factor and software-factor through correlation analyses between different phonemes (or words) on speaker-factor.

If there exists a uniform speaker normalization technique or relationship among speakers, we call that the speaker difference is not random, that is, *structural*, where "*structural*" means that the relationship of speaker differences is analogous between two phonemes or two words. If we use speech materials of adult males, adult females and children, we will obtain the result that the speaker difference is structural,[1] because the vocal tract shape (or length) is different doubtlessly among them. Therefore we use speech materials of only adult males.

2. Speaker Differences in Vowels and Spoken Words[5]

Table 1 shows the kinds of speaker differences in spoken words. It is con-

Table 1. Speaker differences

kind	phenomenon	countermeasure
physical difference of apparatus	distortion of spectrum (shift of spectrum)	shift of formant frequency (frequency warping) normalization of vocal tract length clustering of speakers (grouping of speakers) multiple templates statistical/probabilistic model
difference of linguistic environment or articulation	duration extent of coarticulation or word juncture nasalization/vocalization/devocalization/palatalization	dynamic time warping model of coarticulation multiple templates (or lexicons) statistical/probabilistic model phonological rules (rewriting rules)

sidered that the most part of speaker differences in vowels is caused by the physical differences of the apparatus. On the other hand, we can consider that the speaker differences in spoken words are caused by both the differences of apparatus and articulation. If the correlation of speaker differences between any vowel and a spoken word is large, we can consider that the speaker differences in the word are caused by the differences of apparatus, but not articulation. If the correlation is small, we can consider conversely that the speaker differences in the word are caused by the differences of articulation, but not by apparatus.*

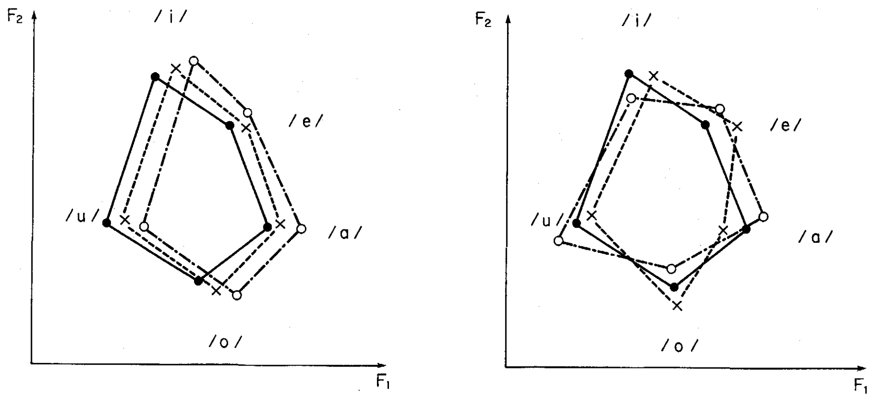
Fig. 1 illustrates the relationship of correlation between vowels on speaker-factor. In the case of (a) in Fig. 1, vowels uttered by three different speakers deeply relate to each other, respectively. This implies that the speaker differences are caused by physical differences and they are structural. On the other hand, in the case of (b), vowels uttered by a specified speaker do not relate to each other. This implies that the speaker differences are caused by differences of articulation.

However, it is very difficult to calculate the correlation directly. Therefore we obtain it from the correlation between the distance matrices which indicate the distance measures among speakers for vowels or spoken words.

Let d_{ij}^k be the distance between phoneme (or spoken word) k of speaker i and phoneme (or spoken word) k of speaker j. Let r^{mn} be the correlation of speaker differences between phoneme m and phoneme n. We define r^{mn} as follows:

$$r^{mn}=\frac{\sum_{i,j}(d_{ij}^m-d^m)\cdot(d_{ij}^n-d^n)}{\{\sum_{i,j}(d_{ij}^m-d^m)^2\cdot\sum_{i,j}(d_{ij}^n-d^n)^2\}^{1/2}},$$

where d^m denotes the average distance of d_{ij}^m for all i and j $(i\neq j)$. The distance between vowels is calculated by Chebychev norm (absolute value norm; city-block distance) between local frames in stationary parts. On the other hand, the distance



(a) large correlation (b) small corrrelation

Fig. 1. Graphic representation of correlation between vowels on speaker-factor (F_1–F_2 plane)
●——● : speaker A, ×······× : speaker B, ○-·-○ : speaker C

* We can consider generally there is no correlation between the speaker difference of aparatus and that of articulation.

Table 2. Examples of distance matrices among speakers for /m/ and /n/

(a) /m/ (1–20 denote speakers)

2	168																		
3	257	244																	
4	244	203	230																
5	148	135	267	218															
6	148	137	239	200	118														
7	185	168	267	216	162	145													
8	243	194	247	228	209	218	210												
9	257	222	241	255	262	260	284	262											
10	171	168	223	192	185	165	201	229	236										
11	319	270	283	261	300	305	276	213	275	263									
12	244	183	266	233	202	203	203	182	263	214	229								
13	210	186	188	226	236	217	240	219	211	197	275	225							
14	232	207	263	246	251	239	263	239	222	207	291	246	202						
15	248	255	220	201	262	226	224	250	262	224	296	266	230	260					
16	220	169	191	211	234	212	246	214	208	181	270	210	157	187	245				
17	221	196	245	225	206	208	213	196	262	210	220	215	208	241	245	209			
18	237	207	222	209	241	210	219	205	217	178	214	196	195	190	222	186	210		
19	222	203	216	210	216	179	233	242	256	188	294	240	192	223	236	184	211	199	
20	282	233	234	176	255	240	254	210	243	226	228	209	219	232	224	212	231	198	224
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

between consonants or spoken words is calculated by dynamic time warping based on a dynamic programming technique. Therefore, we should notice that the speaker differences on duration will be normalized.

Table 2 shows examples of distance matrices which indicate distance measures among 20 male speakers for nasal consonants (/m/ and /n/). In this case, the correlation between $d_{ij}^{/m/}$ and $d_{ij}^{/n/}$ becomes about 0.80 (see Section 4).

We can test the null hypothesis (no correlation between d_{ij}^{m} and d_{ij}^{n}; $r^{mn}=0$) since the following value

$$t_0=\frac{r\sqrt{N-2}}{1-r^2}$$

is distributed approximately as t-distribution with $N-2$ degrees of freedom, where r shows the correlation between m and n obtained from test samples.[14] When the number of samples, N, is ${}_{20}C_2=190$, if r is larger than 0.19, the null hypothesis would be rejected with 0.01 significant level.

3. Experimental Results of Vowels and Spoken Digits

Speech materials are five Japanese vowels and ten digits uttered by 20 adult males and they are analyzed by a 20 channel 1/4 octave filter bank and sampled at

Table 2. (continned)

(b) /n/ (1–20 denote speakers)

2	180																		
3	267	247																	
4	223	181	241																
5	165	186	295	220															
6	139	176	231	202	155														
7	184	193	254	199	192	171													
8	227	198	253	206	198	211	203												
9	236	195	234	235	272	235	269	250											
10	193	178	227	215	229	164	225	227	188										
11	305	213	254	268	287	283	255	198	244	242									
12	207	157	233	203	210	167	184	172	229	182	222								
13	209	198	218	223	259	214	267	243	164	201	262	221							
14	215	197	250	243	261	219	258	240	174	194	258	226	188						
15	231	231	224	148	253	221	223	239	247	233	280	229	231	249					
16	212	179	204	196	254	212	254	238	167	166	250	186	163	185	223				
17	262	193	244	208	242	251	235	190	246	261	194	214	239	254	225	230			
18	241	197	216	224	258	215	235	207	204	190	198	169	210	221	240	157	210		
19	235	211	231	165	232	202	234	240	206	198	290	224	198	234	208	175	228	214	
20	274	214	240	167	274	242	233	222	218	212	230	205	220	245	204	198	230	206	215
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

every 10 ms. The output sample's power is normalized, that is, each output sample, $X(=X_1, X_2, \ldots, X_{20})$, is transformed as follows:

$$X_i \rightarrow \frac{X_i}{(\sum_{k=1}^{20} X_k^2)^{1/2}}, \text{ where } (\sum_{k=1}^{20} X_k^2)^{1/2} \text{ means the power}$$

Table 3 shows the correlation of speaker differences between a pair of vowels, between a pair of digits and between a vowel and a digit. From Table 3, we can conclude as the following:

1) The correlation of speaker differences between any pair of vowels is large relatively. This shows that the speaker differences in vowel spectra are structural and there exists a speaker normalization method which is common to all vowels and speakers (intra-group), or an estimation method of the spectrum of a vowel from spectra of other vowels for a specified speaker.[6)]* And also, this implies that we can select an optimal set of multiple templates by speaker-clustering for speaker independent speech recognition.[7,8,15)]
2) The correlation of speaker differences between spoken digits is large com-

* Notice that we only insist on "there exists a speaker normalization technique". The technique is not evident and the problem to find it is beyond the scope of this paper.[6)~9),13)]

Table 3. Correlation of speaker differences between two different voices

(a) between vowels

	a	i	u	e
i	0.40			
u	0.34	0.37		
e	0.34	0.20	0.35	
o	0.67	0.24	0.38	0.48

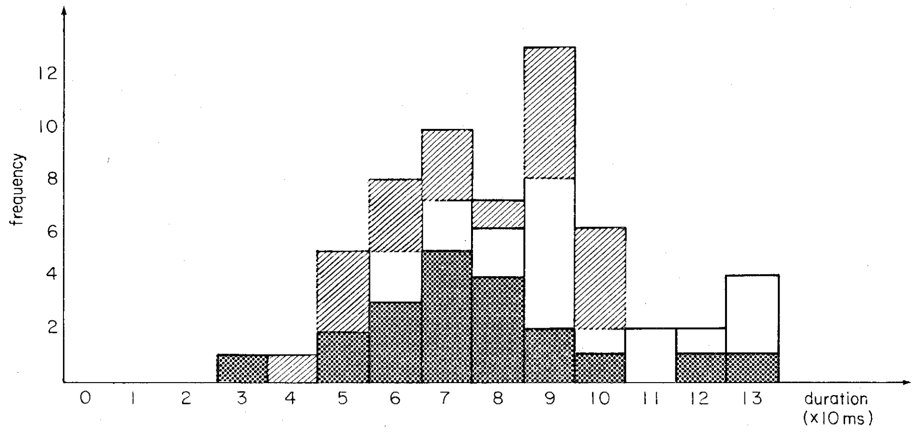
(b) between spoken digits (1: "ici")

	1 (ici)	2 (ni)	3 (san)	4 (yon)	5 (go)	6 (roku)	7 (nana)	8 (haci)	9 (kyu)
2 (ni)	0.11								
3 (san)	-0.08	0.34							
4 (yon)	0.07	0.48	0.61						
5 (go)	-0.03	0.29	0.26	0.48					
6 (roku)	0.26	-0.07	0.18	0.15	0.01				
7 (nana)	0.01	0.37	0.60	0.49	0.29	0.07			
8 (haci)	0.47	0.12	0.23	0.26	0.07	0.39	0.24		
9 (kyu)	0.13	0.32	0.19	0.28	0.33	0.02	0.32	0.18	
0 (rei)	-0.05	0.33	0.16	0.28	0.40	-0.16	0.24	-0.11	0.28

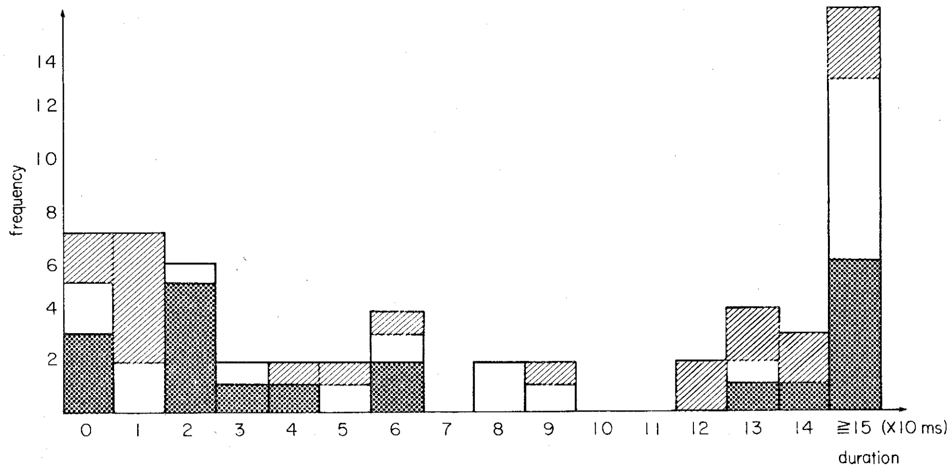
(c) between vowel and digit

	1 (ici)	2 (ni)	3 (san)	4 (yon)	5 (go)	6 (roku)	7 (nana)	8 (haci)	9 (kyu)	0 (rei)
a	-0.18	0.19	0.37	0.22	0.23	-0.04	0.34	-0.04	0.05	0.22
i	0.07	0.38	0.07	0.23	0.30	0.02	0.15	-0.04	0.29	0.31
u	-0.03	0.26	0.30	0.30	0.16	0.14	0.12	0.19	0.34	0.15
e	-0.01	0.37	0.29	0.29	0.24	-0.15	0.22	0.08	0.13	0.32
o	-0.05	0.18	0.48	0.30	0.24	0.06	0.25	0.27	0.15	0.09

paratively except for few pairs. The digits can be divided into two groups, that is, one is (1, 6, 8) and the other is (2, 3, 4, 5, 7, 9, 0). These correspond to a group with devocalized vowel and a group without one, respectively. The correlation of speaker differences between digits except for few pairs is larger than the correlation of speaker differences between a vowel and a digit. These facts show the speaker differences of articulation are also structural. However, since the spoken digit of the group (1, 6, 8; /ici/, /roku/, /haci/) has a silence part at the front of plosive consonant, the small correlation of speaker differences



(a) Duration of silence part at the front of plosive consonant



(b) Duration of devocalized vowel

Fig. 2. Distribution of duration of silence part and devocalized vowel
■ : /ici/, □ : /roku/, ▨ : /haci/

between this group and another group might be caused by the differences of the silence duration. Therefore we investigated the duration of silence parts and devocalized vowels. Fig. 2 illustrates the distribution. We find from this figure that the correlation is not influenced by the speaker differences of silence parts. (Such small difference of silence duration could be normalized by DP matching.)

3) The correlation of speaker differences between a vowel and a digit group (2, 3, 4, 5, 7, 9, 0) is large. This shows that the spectra of such digits could be estimated by the vowel spectra for a given speaker.[9]

4) The speaker differences of a spoken digits in which a vowel may be uttered by the manner of devocalization are much influenced by the differences of articulation. Therefore we must use every discretion in dealing with the

construction of reference pattern for such a word, on speaker independent word recognition.

4. Experimental Results of Consonants

Speech materials are 62 Japanese monosyllables, each of which consists of a consonant and a following vowel, uttered by 20 adult males (who are different speakers from the above). They are analyzed by a 20 channel 1/4 octave filter bank and sampled at every 10 ms. The power of output sample is normalized as mentioned above. All consonant parts are extracted manually by the display of spectra. The boundary was decided as a point of about 30 ms forward from the boundary between a consonant and a following vowel, because the phonetic information of consonant is also included in a transient part.[10]

The distance between consonants of different speakers is defined as the average distance between consonants with the different following vowel. For example, the distance on /*m*/ between speaker *i* and speaker *j*, d_{ij}^{m}, is defined as $(d_{ij}^{ma}+d_{ij}^{mi}+d_{ij}^{mu}+d_{ij}^{me}+d_{ij}^{mo})/5$, where "*ma*" means /*m*/ in syllable /*ma*/.

Table 4 shows the correlation of speaker differences between consonants. From Table 4, we can conclude as the following:

1) The correlation of speaker differences between consonants with the same manner of articulation, such as (m, n), (b, d, g), (z, s, c), (p, t, k), is large. We guess, but not clear, this is related to the facts that the distances between consonants with the same manner of articulation are small for a specified speaker[11] and that the distance between speakers for /*m*/ or /*n*/ is large[3]

Table 4. Correlation of speaker differences between consonants

	y	w	m	n	b	d	g	r	z	s	c	h	p	t
w	0.45													
m	0.46	0.47												
n	0.50	0.47	0.80											
b	0.48	0.33	0.34	0.43										
d	0.21	0.27	0.20	0.43	0.47									
g	0.34	0.37	0.34	0.45	0.45	0.40								
r	0.37	0.26	0.38	0.51	0.42	0.43	0.25							
z	0.13	0.12	0.16	0.24	0.14	0.07	0.35	0.13						
s	0.19	0.08	0.07	0.14	0.25	0.22	0.20	0.20	0.43					
c	0.34	0.16	0.13	0.17	0.27	0.19	0.14	0.34	0.41	0.51				
h	0.30	0.23	0.15	0.28	0.26	0.31	0.12	0.31	0.04	0.07	0.27			
p	0.32	0.32	0.22	0.36	0.39	0.48	0.20	0.36	0.17	0.15	0.47	0.37		
t	0.06	0.11	0.03	0.26	0.30	0.58	0.05	0.39	0.03	0.17	0.25	0.44	0.51	
k	0.37	0.16	0.01	0.24	0.37	0.33	0.30	0.21	0.16	0.25	0.47	0.41	0.55	0.29

(nasal consonants are significant phonemes for speaker identification.[11)12)]*

2) The correlation of speaker differences between voiced consonants except /z/ is large relatively, but the correlation of speaker differences between a voiced consonant and an unvoiced consonant except plosive consonants is small.

From these results, we can guess that the speaker differences of consonants are caused by the physical differences of apparatus and articulation differences of source, and that there is no correlation between a glottal source and a turbulent noise source for a specified speaker.

5. CONCLUSION

We investigated the structure of speaker differences in vowels, consonants and spoken digits through correlation analyses between voices on speaker-factor. From experiments we found the following:

1) The speaker differences caused by physical differences in vowels, consonants and spoken digits are structural.
2) The speaker differences of a spoken digit in which a vowel may be uttered by the manner of devocalization are much influenced by the differences of articulation.
3) The speaker differences of articulation in spoken digits are also structural.
4) The correlation of speaker differences between voiced consonants except /z/ is large, but that between a voiced consonant and an unvoiced consonant except plosive consonants is small.

In other words, these facts show: 1) the spectra of spoken digits could be estimated by the vowel spectra for a specified speaker, 2) we must use every discretion in dealing with the construction of reference pattern for a spoken word with devocalized vowels on speaker independent word recognition, and 3) there is no correlation between a glottal source and a turbulent noise source for a specified speaker.

Although we simply investigated the average correlation for every consonant, it is necessary to investigate in detail the correlation for every following vowel. However, for this purpose, we must use several materials for each monosyllable and speaker, because the spectra of these utterances are random variables. This is an open problem.

Further, we did not compare the size of the speaker differences of apparatus with that of articulation. This is also an open problem.

* For example, if the distance between phonemes /*a*/ and /*b*/ for any speaker is constant, the larger the speaker difference for /*a*/ or /*b*/ is, the larger the correlation r^{ab} becomes.

References

1) S. Nakagawa, H. Shirakata, M. Yamao and T. Sakai, "Considerations on speaker grouping by sex and age for automatic speech recognition," Trans. Elect. Comm. Engrs. Jpn., 63-D, 12, 1002–1009 (1980, in Japanese), or "Differences in feature parameters of Japanese vowels with sex and age" Studia Phonologica, XIV, 35–52 (1980).
2) H. Matsumoto, T. Sone and T. Nimura, "A study on vowel individualities of the vowels—vocal individualities among different vowels," Report of the 1968 Autumn Meeting, The Acoust. Soc. Jpn. 147–148 (in Japanese).
3) T. Sakai and K. Tabata, "Multivariate statistical analysis of VCV syllables," Trans. Inst. Elect. Comm. Engrs. Jpn., 56-D, 1, 63–70 (1973, in Japanese).
4) K. Shikano and N. Sugamura, "On the effect of reference patterns for spoken word recognition," Report of the 1981 Autumn Meeting, the Acoust. Soc. Jpn., 1, 111–112 'in Japanese).
5) S. Nakagawa, "Correlation of speaker differences in vowel spectra and spoken word spectra," Report of the 1982 Spring Meeting, the Acoust. Soc. Jpn., 1, 23–24 (in Japanese).
6) M. Kohda and S. Saito, "Mechanical recognition of spoken digits by incomplete learning samples," Report of the 1972 Spring Meeting, the Acoust. Soc. Jpn., 387–388 (in Japanese).
7) T. Nakamoto and S. Nakagawa, "Speaker-independent spoken word recognition for a large vocabulary based on isolated syllable recognition," Trans. Committee on Speech ReseOrch, the Acoust. Soc. Jpn., S82-20, 153–160 (1982, in Japanese).
8) M. Kinoshita, R. Mizoguchi and O. Kakusho, "Design of phoneme templates for unspecified speakers and their unsupervised learning," Report of the 1982 Autumn Meeting, the Acoust.
9) S. Nakagawa and T. Sakai, "A real time spoken word recognition system with various learning capabilities of the speaker differences," Trans. Inst. Elect. Comm. Engrs. Jpn., 61-D, 6, 395–402 (1978, in Japanese), or "A pre-matching method for a real time spoken word recognition system and a learning procedure of speaker differences" Studia Phonologica, XII, 39-58 (1978).
10) S. Nakagawa and T. Nakamoto, "Speaker-independent large vocabulary word recognition based on syllable by syllable input," Trans. Inst. Elect. Comm. Engrs. Jpn., 65-D, 12, 1558–1565 (1982, in Japanese).
11) S. Nakagawa and T. Sakai, "Feature analyses of Japanese phonetic spectra and considerations on speech recognition and speaker identification," J. Acoust. Soc. Jpn., 35, 3, 111–117 (1979, in Japanese), or "Some properties of Japanese sounds through perceptual experiments and spectral analyses," Studia phonologica, XI, 48–64 (1977).
12) S. Nakagawa and M. Sakamoto, "Evaluation FFT cepstrum and LPC cepstrum for speech and speaker recognition," Trans. Inst. Elect. Comm. Engrs. Jpn., 66–A, 12, 1199–1206 (1983, in Japanese).
13) S. Furui, "A training procedure for isolated word recognition systems," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-28, 2, 129–136 (1980).
14) M. Shiotani and C. Asano, *Multivariate Analysis* (Kyoritsu-Shuppan Co., Tokyo, 1967, in Japanese), p. 194.
15) R. J. Golibersuch, "Automatic prediction of linear frequency warp for speech recognition," Proceedings of Int. Conf. on Acous. Speech & Signal Process., 769–772 (1982).

(Aug. 31, 1987, received)