STUDIA PHONOLOGICA X (1976)

On-Line, Real-Time Spoken Words Recognition System with Learning Capability of the Speaker Differences

Toshiyuki SAKAI and Sei-ichi NAKAGAWA

SUMMARY

We have developed the LISTEN (LIMITED SPOKEN TEXT ENCODER) system which automatically recognizes spoken words in isolation for a limited vocabulary. This system is a subpart of the LITHAN (LISTEN-THINK-ANSWER) speech understanding system[1,2].

This makes a great feature of the recognition in real time on a mini-computer. Owing to this development, it became capable of trying the various experiments on many speech data. There are other two features in this system: One is to learn the speaker differences by preliminary uttered vowels. The other is that the system is composed of two stages, i.e., phoneme recognition and word recognition. In the latter stage, the effect of coarticulation is taken into account.

The system performance obtained the recognition rate of 98.0% on experiments of spoken digits that were uttered by 40 male adults. And also the system obtained the rate of 98.4% on preliminary learning by some spoken digits. When no learning procedure, however, the rate decreased to 95.8%.

I. INTRODUCTION

In an automatic word recognition, if all of the input pattern is regarded as a point in the pattern space, the recognition can avoid the problem of coarticulation, and also it can use the linguistic information through lexicons, i.e., the redundancy of a natural language. Therefore, we have become capable of recognizing spoken words for a limited vocabulary in the case of limited speakers.

There are two main problems which make phoneme recognition in continuous speech do difficult. One of these is that of coarticulation and the other is that of the speaker differences. We have adopted rewriting rules in phoneme recognition precess and word recognition process on the problem of coarticulation, the learning approach on that of the speaker differences.

There are other difficult problems on phoneme recognition such as acoustic signal variations causing by the rate of speech, speaker's emotion and diarect.

Toshiyuki SAKAI (坂井利之): Professor, Department of Information Science, Kyoto University.
Sei-ichi NAKAGAWA (中川聖一): Assistant, Department of Information Science, Kyoto Univeristy.

But the most important problem might be to extract the best acoustic features from speech wave. We employed the spectrum analysis by a filter bank. Ichikawa et al. reported that the short time spectrum was one of the best acoustic features for speech recognition in the current art of speech analysis[3].

On a matching algorithm between an input pattern and a reference (In generally, these patterns are a time series of feature parameters.), the authors believe that the matching method using dynamic programming (DP) is one of the best algorithms for an automatic word recognition. The DP matching algorithms may be divided into many kinds of various types in related to the level of matched patterns[1]. All DP matching algorithms used make good use of the properties of speech sound such as the continuity and the regular order on time.

We propose a new method of word recognition on the basis of a DP matching between a recognized phoneme string and phoneme string given by a lexical entry in a word dictionary. This method has the following merits:

1. fast algorithm for the recognition.
2. normalization of the influence of coarticulation.
3. adoption of the phoneme similarity matrix associated with the confusion matrix of phoneme recognition.
4. automatic construction of the effective word dictionary.
5. adoption of dynamic programming with some matching restrictions.
6. capability of word spotting or connected words recognition.

II. Outline of System Organization

Fig. 1 indicates the block diagram of LISTEN (LImited Spoken Text ENcoder), is composed of the preliminary learning stage of the speaker differences

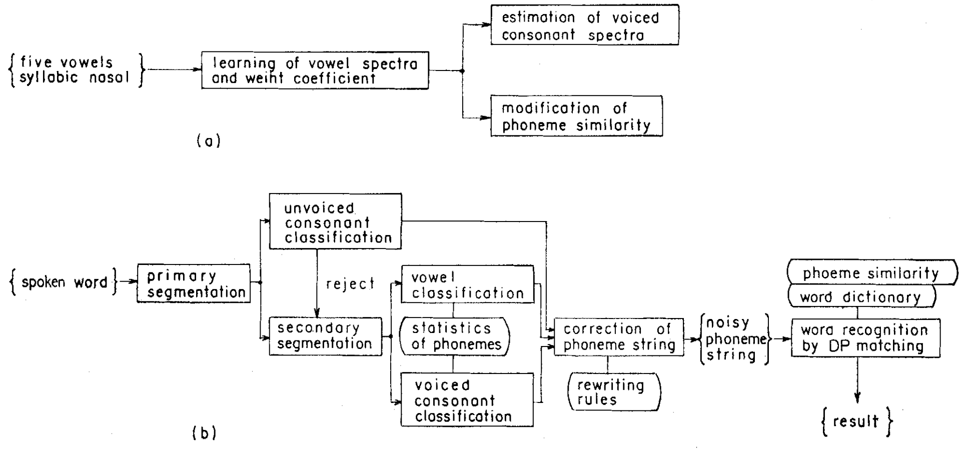
five vowels
syllabic nasal
learning of vowel spectra and weiht coefficient
estimation of voiced consonant spectra
modification of phoneme similarity
(a)
spoken word
primary segmentation
unvoiced consonant classification
reject
secondary segmentation
vowel classification
statistics of phonemes
voiced consonant classification
correction of phoneme string
rewriting rules
noisy phoneme string
phoeme similarity
word dictionary
word recognition by DP matching
result
(b)

Fig. 1. Configuration of the LISTEN spoken words recognition system.
(a) preliminary learning of the speaker diderences.
(b) spoken words recognition.

and the stage of spoken words recognition in real time. The latter stage consists of phoneme classification component and word recognition (translation) one from a phoneme string to a word.

First, a new user of the system is requested to utter clearly five vowels (/a/, /i/, /u/, /e/, /o/) and the syllabic nasal /N/ for the preliminary learning of his own voice. Since the spectral variation of the syllabic nasal is very large inter-speakers, it is learned as well as vowels. Note that it is uttered as /uN/, because we cannot utter it in isolation. The learned spectrum of the syllabic nasal is used for the estimation of voiced consonant spectra.

Second, he can utter any word in a vocabulary after the learning. In order to recognize in real time, the system employs a simple algorithm in the part of the phoneme classification component. Final, the recognized phoneme string is passed to the stage of word recognition.

This system was implemented on a mini-computer (MELCOM-70, cycle time= 0.8 μs, core memory=24 Kw). The algorithms were programmed in an assembly language. The program consisted of about 4.5 K words (4500 steps) and the work area was about 5.5 K words. The matching time between a phoneme string and a lexical entry in the word dictionary was required about 8 ms. After an utterance of a word, the system can illustrate the recognition result on a graphic display within 100 ms for a small vocabulary.

III. PHONEME RECOGNITION

III-1. Outline of Phoneme Recognition

Speech signal is analyzed by a 20-channel 1/4-octave filter bank. It is passed into a pre-emphasis circuit with a slope of 6-dB per octave below 1600 Hz and fed into the 20-channel filter bank. Its output is rectified, smoothed and sampled at every 10 ms intervals, thus yielding a short time spectrum of 20 dimensions. The filters cover the frequency spectrum 200 Hz to 6400 Hz.

The aim of phoneme recognition is to segment input speech into a unit of phoneme (We call this unit a segment, that is, by a segment we mean a portion of the utterance which is hypothesized to be a single phoneme.), and assigns one of the phoneme categories to the unit. We classify Japanese phonemes into following categories.

1. vowel /a/, /i/, /u/, /e/, /o/
2. semi-vowel /y/, /w/
3. nasal /m/, /n/, /ŋ/
4. voiced plosive /b/, /d/, /g/
5. liquid-like /r/
6. voiced fricative /z/, /d_z/
7. voiceless plosive /p/, /t/, /k/
8. voiceless fricative /s/, /ʃ/
9. affricate /c/
10. aspirated /h/
11. syllabic nasal /N/
12. silence /./

The silence corresponds to the closure of unvoiced plosive consonants or a choked sound. LISTEN treats /z/ and /d_z/ or /s/ and /ʃ/ as the same phoneme.

LISTEN does not classify voiceless plosive group, therefore, we simply denote this group as /p/.

The primary segmentation[4] is performed for analyzed speech signal, that is, it classifies the input (a sequence of short time spectra) into one of silence, voiceless-nonfricative, voiceless-nonplosive and a voiced group based on energy and deviation around the low or high frequency of spectrum at every 10ms. (We call this spectrum unit as one frame hereafter.) If a part of a sequence of recognized phonemes is composed of the same phonemes, they will be combined. On the other hand, if it is irregular, it will be smoothed by using rewriting rules or phonological rules. The segment classified as a voiceless group is further classified into one of the detailed group corresponding to each phoneme on the basis of the segment duration, the presence of silence in preceding segment and spectral change, etc.

In the processing of vowels, the system computes a distance measure between each frame in voiced parts and each of six reference patterns: five vowels and the syllabic nasal (the syllabic nasal is treated like the five vowels in the following process). The cityblock distance is used as a distance measure. The calculation of this distance need not be multiplicative. Two nearest neighbors are selected as candidate phonemes and ordered by application of the corresponding linear discriminant functions. The voiced consonant parts, which are detected by the change of spectrum or power, are recognized on the basis of Euclidean distance. These phoneme recognition procedures can be achieved for each frame within the sampling interval, 10 ms.

A recognized phoneme string in voiced parts is smoothed and merged by rewriting rules. An output of phoneme classification process is a sequence of segments, each consisting of 4 tuples: the first candidate of phonemes, the second candidate, degree of confidence of the first candidate and the segment duration. Finally, the recognized phoneme string is passed to the stage of word recognition.

III-2. Preliminary Learning of the Speaker Differences

On the vowel spectrum learning, the system uses the three successive frames around the one with the largest energy and for the syllabic nasal, the successive three frames in the portion preceding by 50 ms from the end of the utterance (/uN/). Fig. 2 illustrates how the learning samples are extracted from the utterance.

It involves risks to use the spectra of isolated vowels for the recognition of the vowels in continuous speech since the spectral difference between isolated vowels and the vowels in-words may be large. Now, let y_i $({}_1y_i, {}_2y_i, \dots, {}_{20}y_i)$ be the spectrum common to all speakers for the isolated vowel i and y_i^s be the spectrum of the speaker s. And also, let $\bar{y}_i$ and $\bar{y}_i^s$ be the spectrum in-words, respectively. We can know previously only y_i and $\bar{y}_i$. Mentioned above, y_i^s is obtained from the utterances of isolated vowels. We want to know $\bar{y}_i^s$ by using these known spectra. The system estimates $\bar{y}_i^s$ by the following equation

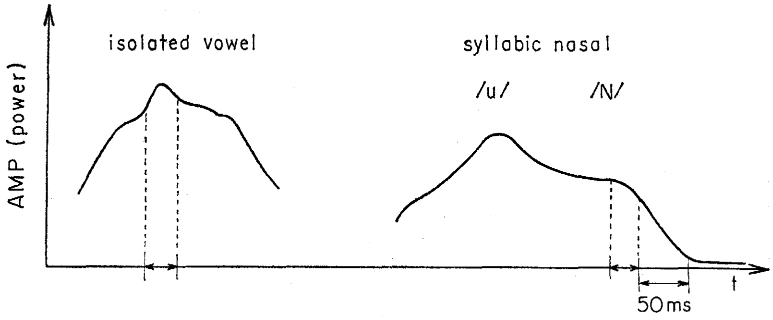

Fig. 2. Portions of extracted samples for preliminary learning.

Table 1. Estimation error of vowel's spectrum in-words.

α	$\varepsilon_1=\Sigma\mid\ \mid$	$\varepsilon_2=\Sigma(\ )^2$	note
0	97.8	15.88	isolated vowel's spectrum for each speaker
1	60.3	6.77	
2	53.7	5.43	
3	52.3	5.07	
4	52.0	4.96	
5	52.0	4.93	
6	52.1	4.93	
7	52.2	4.94	
8	52.4	4.95	
9	52.5	4.97	
10	52.9	4.97	
∞	54.5	5.28	common vowel's spectrum in-words for all speakers

($\hat{y}_i{}^s$ denotes the estimation of $\bar{y}_i{}^s$).

$$\hat{y}_i{}^s=(\alpha\bar{y}_i+y_i{}^s)/(\alpha+1)$$

Table 1 shows the estimation errors which are evaluated by the following.

$$\varepsilon_1=\sum_s\sum_i\sum_{k=1}^{20}|{}_k\bar{y}_i{}^s-{}_k\hat{y}_i{}^s| \qquad \varepsilon_2=\sum_s\sum_i\sum_{k=1}^{20}({}_k\bar{y}_i{}^s-{}_k\hat{y}_i{}^s)^2$$

In this table, $\alpha=0$ corresponds to the pre-registration of isolated vowels, and also $\alpha=\infty$ corresponds to the usage of the vowels common to all speakers which are extracted from the connected speech. In order to evaluate the effect of the learning stated above, we recognized the vowels in digits spoken by 10 male adults by using linear discriminant functions.

The learning of linear discriminant functions, which discriminate a category in the vector space, can be also considered as mentioned above. Let $\bar{w}_{ij}$ be the weighting vector of the linear discriminant function distinguishing i from j. Then, the estimation of $\bar{w}_{ij}$ for the speaker s is given by the following.

$$\hat{w}_{ij}{}^s=(\alpha\bar{w}_{ij}+\bar{y}_i{}^s-\bar{y}_j{}^s)/(\alpha+1)$$

Experimental results are shown in Table 2. From this table, we find that the best value of α is 2, and further this method has almost the same ability as the case of using the individual spectrum for each speaker.

The best value of α is different between the estimation of spectra and the

Table 2. Recognition results of vowels in digits (number of errors). Number of samples /a/=200, /i/=100, /u/=50, /e/=50, /o/=150.

α	a	i	u	e	o	total	note
0	31	24	10	4	55	124	isolated vowel's spectrum
2	20	6	5	1	18	50	learning by isolated vowels
4	30	10	5	2	25	72	learning by isolated vowels
∞	44	6	6	2	9	67	common vowel's spectrum
	27	4	2	1	10	44	spectrum in-words for each speaker

estimation of weight. We think that this is caused by non-linear relation between the two estimation and by the usage of different speech materials.

Since the speaker differences of voiced consonant spectra are larger than that of vowels, a successful system must normalize or learn the differences. In order to learn them speedy, we use the learned spectra of vowels.

Let $\bar{y}_l$ be the spectrum common to all speakers for the phoneme l and $\bar{y}_l^s$ be that of the speaker s. The system attempts to estimate the spectrum of the voiced consonant l for the speaker s, by using $\bar{y}_i$ and $\bar{y}_i^s$ (i: vowel) as follows.

$$\hat{y}_l^s = \bar{y}_l + K_{li}(\bar{y}_i^s - \bar{y}_i)$$

Where K_{li} is estimated as the (20×20) diagonal matrix expressing the relation between l and i. In practice, the system adopts the previously learned spectrum $\hat{y}_i^s$ instead of the spectrum $\bar{y}_i^s$ because the system cannot know $\bar{y}_i^s$. K_{li} for all pairs of phonemes were obtained from the speech materials of 245 words con-

Table 3. Estimation error of voiced consonant's spectrum for all speakers.

(a) classwise estimation error of speaker

estimation method \ speaker	YS	NK	MT	AR	total
Common spectrum for all speakers	420	454	517	505	1896
Estimation by spectrum of /a/	476	419	526	480	1901
Estimation by spectrum of /i/	393	385	518	482	1776
Estimation by spectrum of /u/	404	387	489	473	1753
Estimation by spectrum of /e/	408	589	513	569	2079
Estimation by spectrum of /o/	406	437	477	499	1819
Estimation by spectrum of /N/	383	366	496	501	1746
personally adjusted spectrum*	272	240	378	379	1269

* corresponds to the spectrum of known speaker.

(b) classwise estimation error of voiced consonant

method	m	n	$\tilde{g}$	b	d	g	r	z
common	214	184	289	233	136	405	200	236
a	204	163	282	236	129	455	195	239
i	189	163	256	238	118	411	188	216
u	181	159	240	226	125	422	188	214
e	249	212	320	255	131	499	188	226
o	198	160	255	254	132	437	186	227
N	153	127	258	230	131	425	192	228
individual	116	92	164	173	88	306	162	168

taining VCV contexts, spoken by six male adults, based on the minimum squared error criterion. Other four male speakers were regarded as unknown speakers. Experimental results are shown in Table 3, at every unknown speaker and every voiced consonant. From these results, we find that the best estimation is derived from the spectrum of /u/ or /N/. Strictly speaking, the best one depends on the kind of voiced consonants. Table 4 shows the recognition results of voiced con sonants by using /u/ and /N/ for the estimation, on the basis of Euclidian distance. These give us an interesting suggestion that the estimation for unknown speakers is possible by using their vowel spectra. The system have used the estimation by /N/.

Table 4. Recognition results of voiced consonants by estimated spectrum (error number).

estimation method	speaker	m	n	$\tilde{g}$	b	d	g	r	z	total
common spectrum for all speakers	known	57	89	101	101	49	77	61	20	555
	unknown	58	85	81	57	32	54	42	11	420
estimation by spectrum of /u/	known	57	84	95	85	46	63	59	22	511
	unknown	47	87	74	62	34	49	37	18	408
estimation by spectrum of /N/	known	53	74	94	90	49	67	61	20	508
	unknown	50	62	75	60	32	50	39	17	385
personally adjusted spectrum	known	36	46	74	61	46	52	51	15	381
	unknown*	21	31	48	38	22	39	32	14	245

* corresponds to known speaker.

IV. WORD RECOGNITION

IV-1. Similarity Matrix and Word Dictionary

Since the phoneme recognition is performed by using the statistics of the spectrum for each phoneme, if the system makes mistakes in the phoneme recognition, we can consider that these errors have been caused by the fact that statistics calculated from an uttered phoneme are very similar to those of the misrecognized phoneme. These errors are generally divided into three kinds, that is, a) substitution error; b) insertion error; c) omission error.

Word matching is defined fundamentally as a process to make a one-to-one correspondence between each phoneme of a recognized phoneme string and each phoneme of an entry in the word dictionary. To evaluate a degree of matching between two phonemes, we introduce a concept of the similarity between the two phonemes. We obtained the phoneme similarity S(i, j) for all pairs of phonemes (i, j) by Bhattacharyya distance between the spectrum distributions of the two phonemes[2,4]. Therefore we can evaluate the degree of the matching between two phoneme strings with the help of the similarity matrix. Then, this distance was converted into the value from 0 to 100 by the linear transformation. But if a phoneme i or j was a voiceless consonant, the similarity S(i, j) was decided on the basis of the result of the phoneme recognition.

In order to reduce the matching time and to save the computer storage, each word in the dictionary contains only one description of that word as a string of phoneme symbols. Some phonemes in a word are often influenced by phoneme environments. In consequence of this influence, these phonemes are omitted or misrecognized as other phonemes. Therefore, we introduce a sub-phoneme 'k' in addition to a main-phoneme 'I' and denote this description in the dictionary by I/k(c). This notation means that the phoneme 'I' can be replaced by the phoneme 'k', where 'c' means the weight of the sub-phoneme 'k' ($0 \leq c \leq 1.0$). Table 5 shows the examples of lexical entries. The phoneme with a plus symbol (+) in the table indicates to be able to associate with one or two segments in the recognized phoneme (or segment) string. These descriptions for given words are automatically constructed by the constructing rules of the word dictionary.

Table 5. Entries in the Word Dictionary.

word	symbol	phonemic representation	duration	
			Dmax	Dmin
ichi	1	i ./c(1.0) c $\overset{+-}{\text{i}}$/c(1.0)	350ms	100ms
ni	2	n i	300	100
san	3	s a N	550	200
yon	4	y/$\tilde{\text{g}}$(0.95) o N	450	150
go	5	g o	300	100
roku	6	r o ./k(0.95) k $\overset{+-}{\text{u}}$/*(1.0)	450	100
nana	7	n a/N(0.85) n/a(0.85) $\overset{+}{\text{a}}$/N(0.85)	550	200
hachi	8	h a/N(0.85) ./c(1.0) c $\overset{+-}{\text{i}}$/c(1.0)	500	150
kyu	9	//c(0.95) k/c(0.95) y/u(0.95) u	500	200
rei	0	r/p(0.85) e i/e(0.95)	400	100

IV-2. Matching Algorithm by Dynamic Programming

Let J be the first candidate of the j-th recognized segment, l the second candidate and p the reliability of the first candidate ($0 \leq p \leq 1.0$). Let I be the i-th main-phoneme of a given lexical entry, k the sub-phoneme and c the weight of the sub-phoneme. Then, if that segment in a recognized string associates with the i-th element of this entry, the similarity is defined as the following equation:

$$S(I, k, c;\ J, l, p) = \max \begin{cases} S(I, J) \\ c \times S(k, J) \\ p \times S(I, J) + (1-p) \times S(I, l) \\ c \times p \times S(k, J) + c \times (1-p) \times S(k, l) \end{cases}$$

We simply denote S(I, k, c; J, l, p) by $S_0(i, j)$. We introduced the following restrictions with respect to the matching between a recognized phoneme string and a phoneme string of an entry.

1. Except for a phoneme marked with a puls symbol, a vowel and the syllabic nasal in an entry are associated with phonemes of three or less in a

recognized phoneme string.

2. A consonant in an entry is associated with phonemes of two or less.

3. Three successive phonemes in an entry is not associated with only one phoneme in a recognized string.

4. Except for an elongated vowel, when the total duration of three successive segments in a recognized string is beyomd 250ms, a vowel in an entry need not be associated with these phonemes.

5. When the duration of one segment is not beyond 100 ms, an elongated vowel in an entry need not be associated with only this segment.

6. If a word matching is performed beyond the range of a given duration by a lexicon, the matching score is decreased.

Fig. 3 illustrates some examples of these restrictions. The evaluation score of matching is calculated by the average of the similarity for all phonemes in an entry. The likelihood for a given word is defined as the highest score of all possible associations. This can be obtained efficiently by the method of dynamic programming.

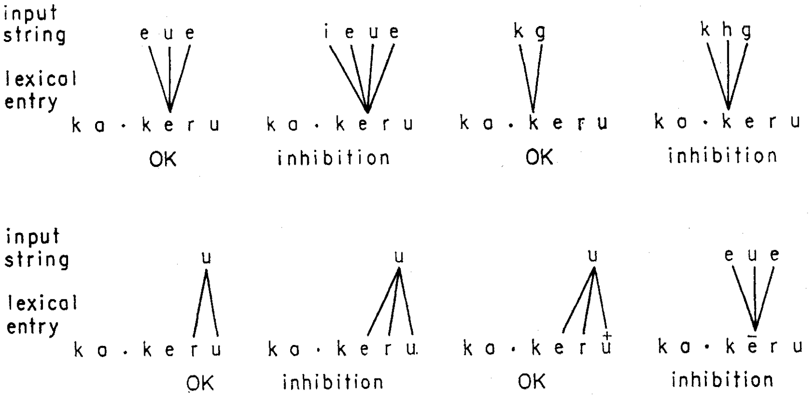

Fig. 3. Examples of restrictions on the matching.

Let L(i, j) be the highest cumulative score up to the i-th element in a lexical entry and the j-th recognized segment. When the i-th element of the entry is a vowel, L(i, j) is calculated with the following equation:

$$L(i,j)=\max\begin{cases}L_1(i,j)=L^*(i-1,j)+S_0(i,j)\\ L_2(i,j)=L(i-1,j-1)+S_0(i,j)\\ L_3(i,j)=L^*(i-1,j-1)+[S_0(i,j-1)+S_0(i,j)]/2\\ L_4(i,j)=L(i-1,j-2)+[S_0(i,j-1)+S_0(i,j)]/2\\ L_5(i,j)=L^*(i-1,j-2)+[S_0(i,j-2)+S_0(i,j-1)+S_0(i,j)]/3\\ L_6(i,j)=L(i-1,j-3)+[S_0(i,j-2)+S_0(i,j-1)+S_0(i,j)]/3\end{cases}$$

where $L^*(i,j)=\max\{L_2(i,j),\ L_3(i,j),\ L_4(i,j),\ L_5(i,j),\ L_6(i,j)\}$ (see restriction 3). The boundary (or initial) conditions are the following:

$$L(1,1)=S_0(1,1)$$

$$L(1, 2)=[S_0(1, 1)+S_0(1, 2)]/2$$

$$L(1, 3)=[S_0(1, 1)+S_0(1, 2)+S_0(1, 3)]/3$$

In the case of a consonant, $L(i,j)$ is $\max\{L_1(i,j), L_2(i,j), L_3(i,j), L_4(i,j)\}$. $L_1, L_2, \ldots, L_6$ corresponds to kinds of possible routes, that is, $r_1, r_2, \ldots, r_6$ in Fig. 4, respectively. If the length of the entry and the recognized string is i_0 and j_0, respectively, the likelihood is obtained as $L(i_0, j_0)/i_0$.

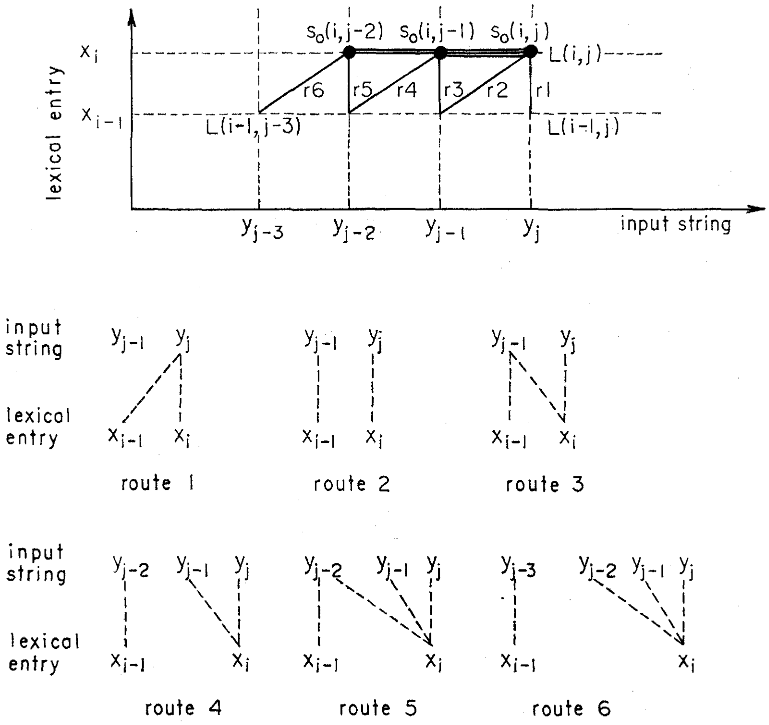

Fig. 4. Kinds of possible matching routes on lattice plane.
route 1~6 for vowels and the syllabic nasal.
route 1~4 for other phonemes.

IV-3. Normalization of Coarticulation

If we use the matching algorithm mentioned above, the similarity between an element in a lexical entry x_i, and three successive segments in a recognized string $\{y_{j-1}, y_j, y_{j+1}\}$ is calculated by the following:

$$S(x_i;\ y_{j-1}, y_j, y_{j+1})=[S_0(i, j-1)+S_0(i, j)+S_0(i, j+1)]/3$$

Now let us consider a matching example such as illustrated in Fig. 5. In this example, the similarity obtained from above equation for the case of (a) is the same similarity as the case of (b). But we will consider obviously that the association of (a) is more natural than that of (b). From this point of view, we try to improve the matching algorithm.

In Fig. 5, if we can regard the association between x_i and $\{y_{j-1}, y_j, y_{j+1}\}$ as a valid association, we could assume that $y_{j-1}(y_j)$ is a transient segment between

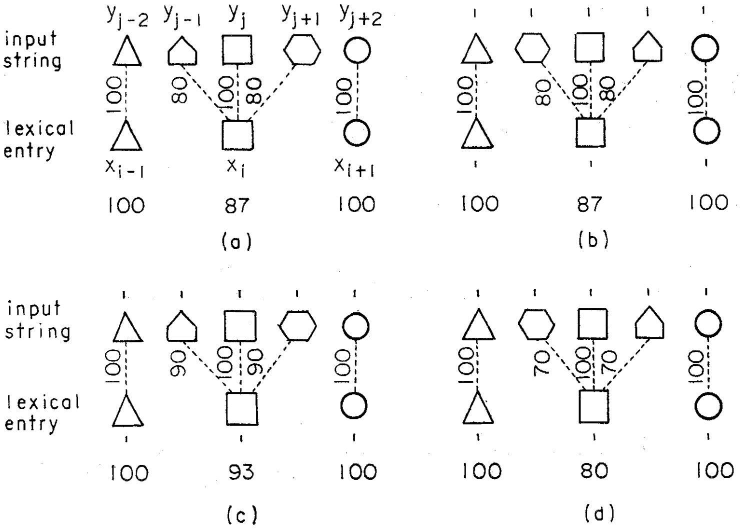

Fig. 5. Graphic model of normalization of coarticulation. (a) and (b): before normalization, (c) and (d): after normalization. Where, we assume that S(△, ⋀)=S(○, ⋁)=80, S(△, □)=S(○, □)=60, S(△, ○)=40, k=0.5.

x_{i-1} and x_i(x_i and x_{i+1}). In this case, the following inequalities might be satisfied, because the transient segment represents an intermediate phoneme between two successive phonemes in the entry (or uttered phonemes).

$$S_0(i-1, j-1) > S_0'(i-1, i)$$
$$S_0(i+1, j+1) > S_0'(i+1, i)$$

where, a and b in $S_0'(a, b)$ denote the a-th and b-th elements in the entry. Therefore, we consider that the normalization of coarticulation can be performed by the modification of the similarity. This modification is defined as the following:

$$\tilde{S}_0(i, j-1) = S_0(i, j-1) + k[S_0(i-1, j-1) - S_0'(i-1, i)]$$
$$\tilde{S}_0(i, j+1) = S_0(i, j+1) + k[S_0(i+1, j+1) - S_0'(i+1, i)]$$

After all, the association between x_i and $\{y_{j-1}, y_j, y_{j+1}\}$ is evaluated by the following equation.

$$S(x_i;\ y_{j-1}, y_j, y_{j+1}) = [\tilde{S}_0(i, j-1) + S_0(i, j) + \tilde{S}_0(i, j+1)]/3$$

Moreover, when an element x_i in a lexical entry associates with two successive segments $\{y_{j-1}, y_j\}$ in a recognized string, the association is evaluated by the following equation.

$$S(x_i;\ y_{j-1}, y_j) = \{S_0(i, j-1) + k[S_0(i-1, j-1) - S_0'(i-1, i)]$$
$$+ S_0(i, j) + k[S_0(i+1, j) - S_0'(i+1, i)]\}/2$$

We hope that this procedure can be applied to the recognition of general context-sensitive patterns.

V. Experimental Results

The similarity matrix and reference patterns were calculated from [vowel - semivowel - vowel] and [vowel - voiced consonant - vowel] contexts which were included in 2450 words spoken by 10 male adults.

First, we examined the performance of the matching procedure of the normalization of coarticulation. A set of three successive vowels was used for this experiment as speech materials, because this kind of phoneme strings might be remarkably influenced by coarticulation. Ten meaningless words were selected at random, and uttered five times by each of 10 male adults. The results are shown in Table 6. The recognition rate without the normalization was 93.6%. On the other hand, the normalization improved the rate to 95.6%.

Table 6. Confusion matrix of speech recognition of three successive vowels.

(a) without normalization of coarticulation

in \ out	uie	uoi	oia	iei	oue	aeo	ueo	uai	ioi	eia
uie	50									
uoi		48			1			1		
oia	1		46		3					
iei				49					1	
oue					50					
aeo						49				
ueo							50			
uai		10			1	1		38		
ioi		2		11					37	
eia										49

(b) with normalization of coarticulation

in \ out	uie	uoi	oia	iei	oue	aeo	ueo	uai	ioi	eia
uie	50									
uoi		48			1			1		
oia			48		2					
iei				49					1	
oue					50					
aeo						49				
ueo							50			
uai		3			1			46		
ioi		2		11					37	
eia										49

Next, we made experiments to recognize isolated spoken digits. Test materials consisted of 2000 digits, each of ten digits was spoken five times by each of 40 male speakers. Ten of them were subjects for the system design, although the designing materials were not the digits. The age of speakers ranged from 21 to 26. It should be noticed that all the experimental results mentioned below are based on 'forced decision'. The three kinds of experiments were carried out.

Experiment a......All the spoken digits were recognized by the common reference patterns (spectra), that is, without learning the speaker differences.

Experiment b......They were recognized by the personal reference patterns which were obtained by the preliminary learning using previously uttered vowels and syllabic nasal in isolation.

Experiment c......They were recognized by the personal reference patterns which were obtained by the preliminary learning using some digits. The spectrum of /a/ was learned by /a_1/ or /a_2/ in 'na_1na_2 [seven]', /i/: 'ni (two)', /e/: 'rei [zero]',

Table 7. Confusion matrices for 2000 digits of 40 male speakers.

(a) without learning: by common reference patterns for 10 male speakers.

in \ out	1	2	3	4	5	6	7	8	9	0
1	199									
2	1	197								2
3	2		175	1					22	
4				167	25				6	2
5				1	199					
6						196				
7					10	1	189			
8	1					6		192		
9		1							199	
0										200

(b) preliminary learning by isolated vowel's spectrum.

in \ out	1	2	3	4	5	6	7	8	9	0
1	199									
2		195		1						4
3	2		194		1			1	2	
4				175	21		1		2	1
5					200					
6						196				
7					2		198			
8						1		198		
9		1							199	
0										200

(c) preliminary learning by some spoken digits.

in \ out	1	2	3	4	5	6	7	8	9	0
1	199									
2		199		1						
3	3		193				2		2	
4				188	10				2	
5				2	198					
6						196				
7					4		196			
8	1					5		193		
9									200	
0										200

Table 8. Relation between the number of speakers and the correct rate of digit's recognition.

a: without learning.

b: preliminary learning by isolated vowel's spectrum.

c: preliminary learning by some spoken digits.

	a	b	c
100%	15	19	22
98	9	11	9
96	1	5	4
94	4	2	3
92	3	1	0
90	3	1	2
less 90	5	1	0

/o/: 'go [five]', /N/: 'yoN [four]', respectively. The spectrum of /u/, however, was learned by an isolated vowel /u/.

Table 7 shows these experimental results. The recognition rates for *Experiment a, b and c* were about 95.8%, 98.0% and 98.4%, respectively. The relation between the number of speakers and the recognition rate is summarized in Table 8. In special, the preliminary learning by isolated vowels had the effect of normalizing the spectrum of a speaker whose voice contrasted in a striking way with the reference voice.

Acknowledgement

The authors wish to thank Mr. T. Ukita and S. Hayashi for their cooperation and various conveniences.

References

1. T. Sakai and S. Nakagawa: "Speech Understanding System LITHAN and Its Evaluation", Technical Report of the Professional Group on Speech of Acoustic Soc. of Japan, S75-30, Nov. 1975.
2. T. Sakai and S. Nakagawa: "Continuous Speech Understanding System LITHAN", Studia Phonologica IX, p. 45, (1975). "Speech Understanding System of Simple Japanese Sentences in a Task Domain", IECEJ Trans. Vol. E-60, p. 13, (1977).
3. A. Ichikawa, Y. Nakano and K. Nakata: "Evaluation of Various Parameter Sets in Spoken Digits Recognition", IEEE Trans. vol. AU-21, p. 202, (1973).
4. T. Sakai and S. Nakagawa: "A Classification Method of Spoken Words in Continuous Speech for Many Speakers", Jour of Information Processing Soc. of Japan, Vol. 17, p. 650, (1976).

(Aug. 31, 1976, received)