STUDIA PHONOLOGICA XIII (1979)

A Recognition System of Connected Spoken Words Based on Word Boundary Detection

Sei-ichi NAKAGAWA and Toshiyuki SAKAI

Summary

We have developed a real time spoken words recognition system on a minicomputer.[1),2)] This system is based on the phoneme classification, that is, hierachical linguistic knowledge. To extend this approach to connected words, we developed a ward boundary detection method by using the pitch contour and energy envelope.

In this paper, we discuss the possibility of word boundary detection by the prosodic information from view points of listening and reading. Further we describe the detection algorithm by machine. From experiments on connected digits and city names, the listener could detect correctly the number of words in connected spoken words at the rate of 95%, and reader could locate correctly the word boundary at the rate of 70%～100%.

The detection algorithm was applied to connected spoken digits. The number detected as candidates of word boundary was 1.2～1.7 times of the actual number and the boundary was detected correctly at the rate of 85～95%. The recognition rate of two connected digits was about 83% for unspecific speakers.

I. Introduction

Recently, several systems were developed for the recognition of connected spoken words. Many speech understanding systems also work themselves as the connected spoken word recognition systems. Almost all methods are based on pattern matching.

Sakoe[3)] and Nakatsu et al[4)]. proposed the typical pattern matching method for the recognition of connected words. The principle of their methods is to regard the all possible word sequences as new isolated words. They developed the two level DP-matching algorithm for shortening the calculations.

Bahl et al[5)]. and Lowerre[6)] brought the principle into continuous speech recognition systems by using the model of stochastic process.

Medress et al[7)]. proposed the recognition method of connected words based on the word to word processing. This method recognizes sequentially connected words one by one from left to right by using a tree search technique. Almost all

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

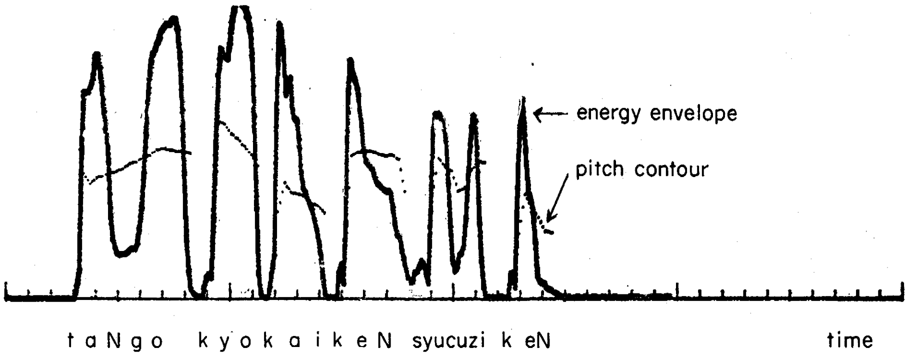


Fig. 1. An example of energy envelope and pitch contour. Utterance: taNgo kyokai kensyucu zi-ken (experiment of word boundary detection)

speech understanding systems based on the hierarchical model have adopted such a sequential method and tree search technique[8].

Rabiner et al[9]. proposed the recognition method on the basis of the word boundary detection of connected digits. However, this algorithm depends on a vocabulary, that is, it is a digit-oriented algorithm. In this paper, we propose a new word boundary detection method, applicable for any word sequence.

We find from the following fact that it is necessary to detect the word boundary in connected words, that is, when we listen the utterance "yamatokoriyama", we can decide whether the utterance consists of two words ("yamato" and "koriyama") or one word. Because the prosodic information makes each word understood as one, well-arranged utterance. In this paper, we used the pitch contour and energy envelope as prosodic information. The role of these parameters was ascertained by perceptual experiments and also the possibility of word detection was investigated by visiual observations. From these results, we developed the algorithm of word boundary detection on a machine, and it was embedded to our isolated word recognition system.

The pitch frequency was extracted every 10 ms interval by the pitch extractor consisting of three low pass filters. The output was smoothed by the five point median algorithm[10]. Although the pitch extractor has the error of about 5 Hz/sec, it is not so significant, because only the comparative change of pitch contour is important for this study. The energy level was calculated from the output values of the 20-channel 1/4-octave filter bank, that is, the root mean square of their outputs. Fig. 1 illustrates the pitch contour and energy envelope of the utterance "tango kyokai kensyucu zikken" (experiment of word boundary detection).

II. WORD BOUNDARY DETECTION IN CONNECTED WORDS BY LISTENING

We investigated through the listening test how the prosodic information brings the role of word detection. On studying the perception of prosody, we often want to generate the speech which contains only prosodic information.

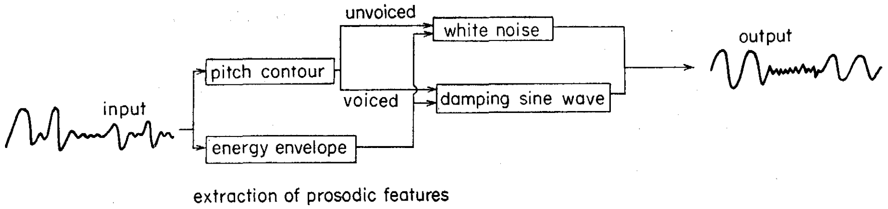


Fig. 2. A speech analysis-synthesis method by pitch contour and energy envelope.

Liberman et al[11]. and Nakatani et al[12]. analyzed "ma ma ..." nonsense phrases that mimicked actual English phrases in the study of prosodic phenomena. However, it is difficult for a speaker to mimick actual sentence by the concatenation of only [ma]. Streeter[13] controlled the prosodic features by using a linear predictive coding analysis—synthesis procedure in the study of phrase boundary perception. We also use an analysis—synthesis procedure.

Fig. 2 shows the block diagram of speech synthesis method for the listening test. The output speech is synthesized by the damping sine wave, $E\cdot\left(1-\frac{t}{2T}\right)\cdot\sin\frac{2\pi t}{T}$, and white noise. The former corresponds to the voiced sounds of input. E and T are synchronized by the energy level and pitch period of input speech, respectively, every 10 ms interval. The white noise corresponds to the unvoiced sounds of input. Thus, the output speech has not phonemic information, but only prosodic information. The sound of damping sine wave is similar to [u] or [m]. This synthesis system can also hold either the pitch contour or energy envelope at a constant value. By this function, we can investigate which of the two parameters is a significant main effect on the word detection.

Experiments and Results

Two speakers uttered 50 digit sequences and 50 city name sequences, respectively. The digit sequences consisted of 25 two connected digits and 25 three connected digits. The half of these were sequences consisting of only voiced sounds. The city name sequences consisted of 10 isolated, 20 two connected and 20 three connected city names. The half of these also consisted of only voiced sounds.

These utterances were analyzed in terms of pitch and energy and synthesized. The subjects, who were 7 adult males, decided how many words there were in the utterance. They were told that utterances would contain more less than three words. Table 1 shows the experimental results, which were decided by majority.

Table 1. Results of word boundary detection by perception

prosodic features	pitch & energy	pitch only	energy only
digit	95%		
city name	95%	92%	92%

The subjects could decide correctly the number of words at the rate of 95%. When only one of two parameters was preserved, the rate was 92%. Therefore we may think that these is no significant difference on the role of word detection between the pitch contour and energy envelope. However, when the utterance consisted of only voiced sounds, the rate was improved to 97.6%. From this fact, we found the pitch contour is more significant than energy envelope on the word boundary detection in Japanese.

III. Word boundary detection in connected words by reading

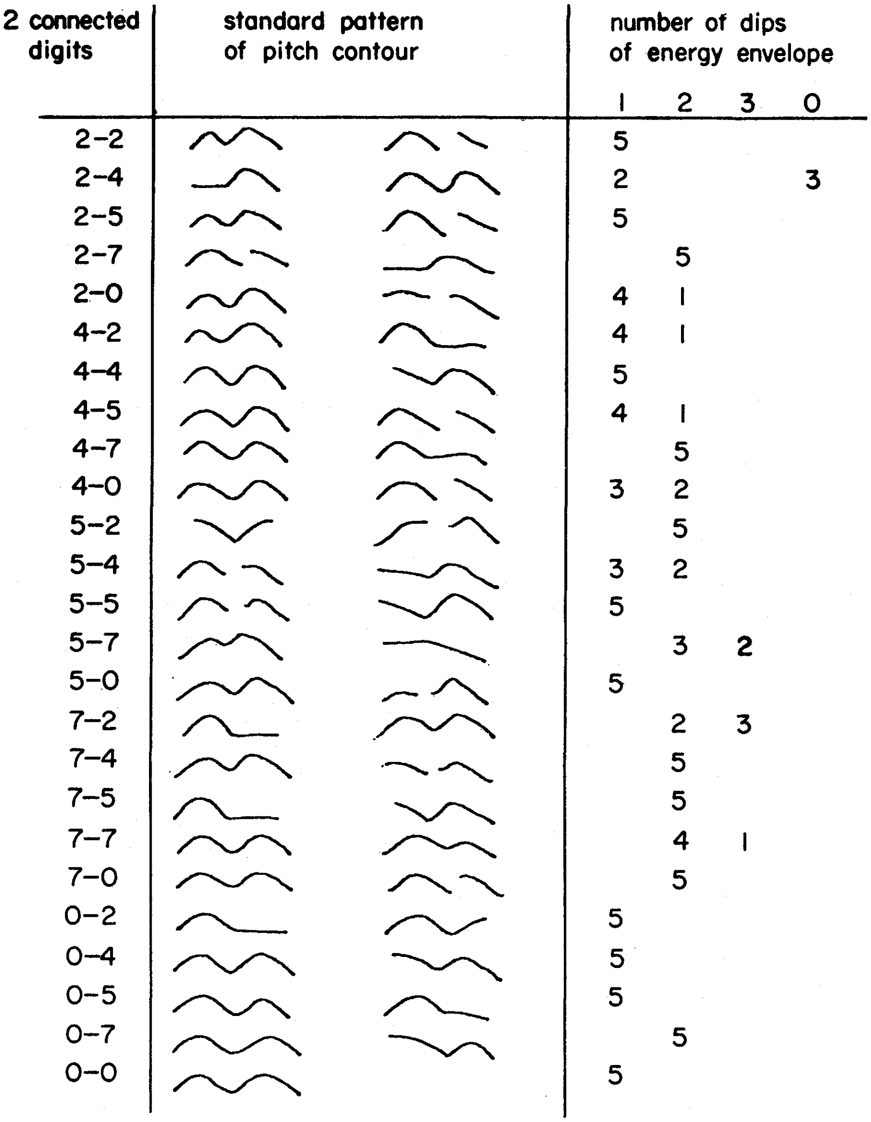
We observed the pitch contour and energy envelope of connected digits spoken by five male adults (in Kinki dialect). Table 2 shows the ways of pronunciation of digits. Figs. 3 and 4 summarize the standard pitch patterns and the number of dips of energy levels. From these observations, we found that almost all word boundaries had the dips of pitch contour and energy envelope. However, a few boundaries do not have the dip. We can approximately say that all boundaries are the points of flexion of pitch contour in Japanese. These phenomena of two or three connected digits are modelled by Fig. 5.

According to these models, we detected the word boundaries in connected digits and connected city names by visual observations. The energy envelope was used to locate the boundaries precisely. The digits consisted of two or three digits and the city names consisted of isolated, two or three city names. The three subjects were experienced in the prosodic study at least one year. The stimulus materials were the same as the previous experiments for listening. Table 3 shows the results of word boundary detection by visual observations.

The column (a) shows the number detected as candidates of word boundary. The total number in real utterances was 75 for connected digits and 60 for connected city names. The column (b) shows the number detected correctly. The number of the columns (c) and (d) corresponds to the number of the shift of only one phoneme and syllable, respectively. Generally speaking, the dip of pitch contour sometimes shifts to the right direction of real word boundary. In this case, this sift error will be corrected by using the energy envelope. However if the two dips of energy envelope are close each other, the word boundary will shift one phoneme or syllable.

Table 2. Pronunciation of Japanese digits

digit	pronunciation	digit	pronunciation
1	ici	6	roku
2	ni	7	nana
3	saN	8	haci
4	yoN	9	kyu
5	go	0	rei

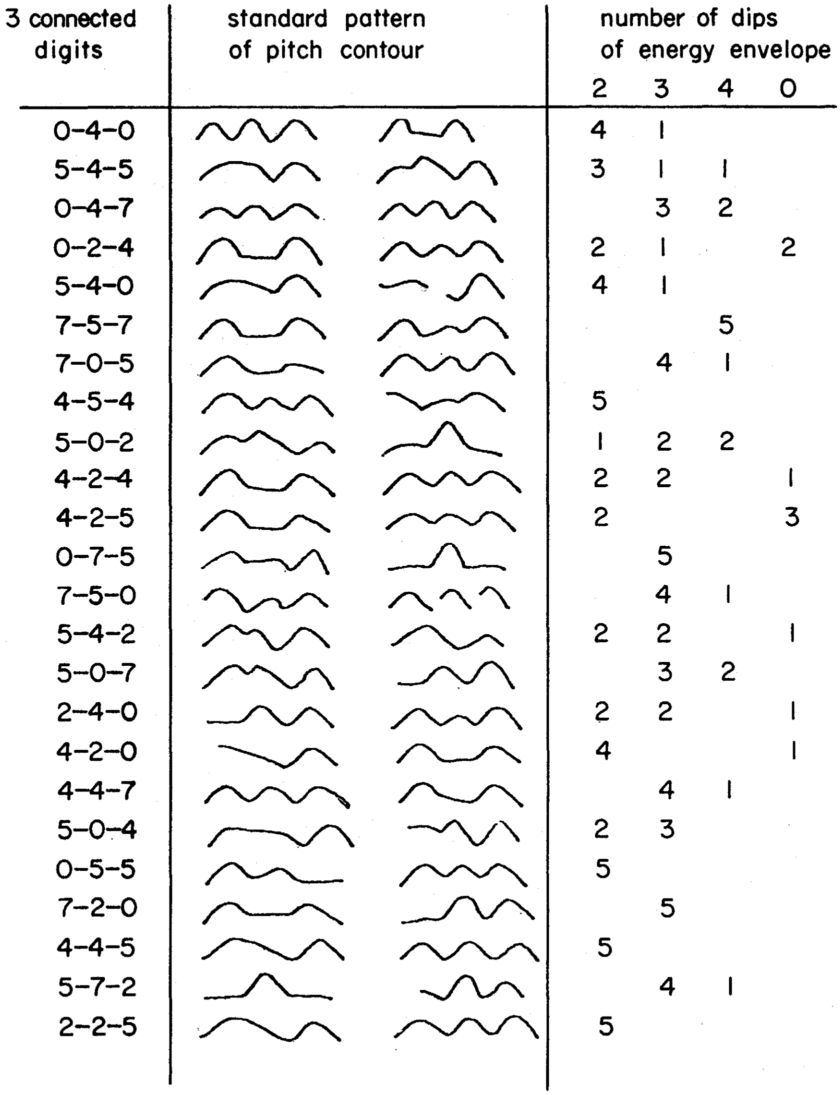


2 connected digits	standard pattern of pitch contour	number of dips of energy envelope 1	2	3	0
2-2		5			
2-4		2			3
2-5		5			
2-7			5		
2-0		4	1		
4-2		4	1		
4-4		5			
4-5		4	1		
4-7			5		
4-0		3	2		
5-2			5		
5-4		3	2		
5-5		5			
5-7			3	2	
5-0		5			
7-2			2	3	
7-4			5		
7-5			5		
7-7			4	1	
7-0			5		
0-2		5			
0-4		5			
0-5		5			
0-7			5		
0-0		5			

Fig. 3. Energy envelope pattern and pitch contour pattern of 2 connected digits which consist of only voiced sounds.

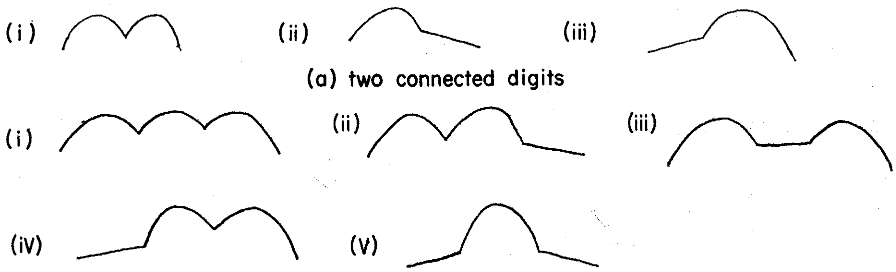
Such shift errors might be the permissible range at the word recognition procedure.

For connected digits, almost all word boundaries were detected exactly. For city names, about 70% were detected correctly and the rate of shift errors was about 25%. The shift error of one syllable was caused by the absence of pitch contour and irregular of energy envelope in unvoiced sounds or devocalized vowels. The parenthesis of Table 3 shows the results of connected words which consist of only voiced sounds. The total number of word boundaries in real utterances was 30. The correct rate was improved for voiced sounds.



3 connected digits	standard pattern of pitch contour	number of dips of energy envelope			
		2	3	4	0
0-4-0		4	1		
5-4-5		3	1	1	
0-4-7			3	2	
0-2-4		2	1		2
5-4-0		4	1		
7-5-7				5	
7-0-5			4	1	
4-5-4		5			
5-0-2		1	2	2	
4-2-4		2	2		1
4-2-5		2			3
0-7-5			5		
7-5-0			4	1	
5-4-2		2	2		1
5-0-7			3	2	
2-4-0		2	2		1
4-2-0		4			1
4-4-7			4	1	
5-0-4		2	3		
0-5-5		5			
7-2-0			5		
4-4-5		5			
5-7-2			4	1	
2-2-5		5			

Fig. 4. Energy envelope pattern and pitch contour pattern of 3 connected digits which consist of only voiced sounds.



(a) two connected digits

(b) three connected digits

Fig. 5. Models of pitch contours of two and three connected digits.

Table 3. Results of word boundary detection by visual observations.

subject	digit				city name			
	a	b	c	d	a	b	c	d
HN	75	75	0	0	62 (30)	44 (25)	2 (1)	12 (3)
TU	64	61	2	1	58 (28)	38 (20)	5 (3)	9 (3)
SN	72	67	5	0	59 (30)	40 (21)	4 (3)	12 (4)

IV. Word boundary detection in connected words by machine

The word boundary of connected spoken words is detected by the procedure as shown in Fig. 6. First, the system decides the preliminary candidates of the word boundary by the use of the pitch contour. Secondaly, it decides the final candidates by the use of the valley of the energy envelope. The pitch pattern of spoken word is like the figure of "Λ". Therefore we can detect the word boundary by finding the pattern in the pitch contour. The energy envelope shows a valley at the word

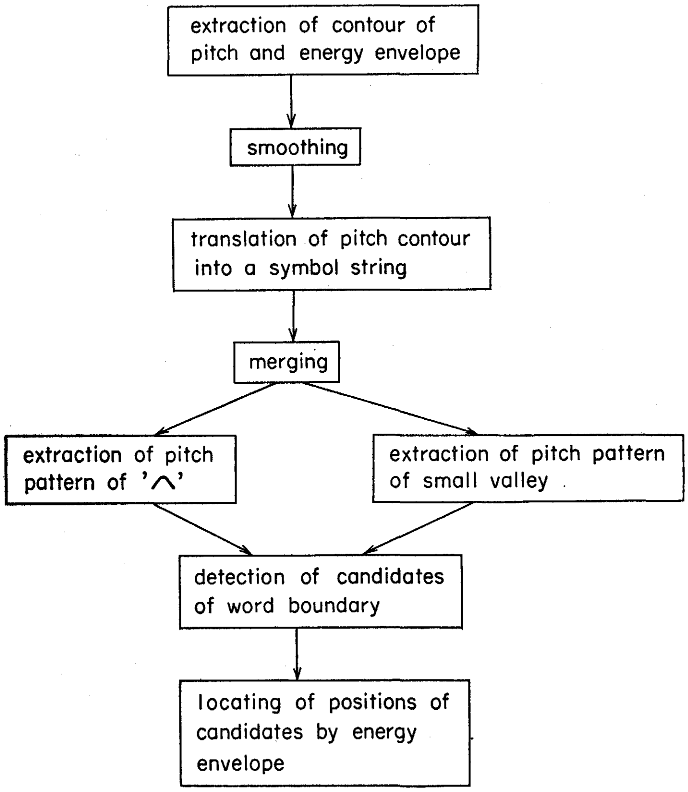

extraction of contour of pitch and energy envelope
smoothing
translation of pitch contour into a symbol string
merging
extraction of pitch pattern of 'Λ'
extraction of pitch pattern of small valley
detection of candidates of word boundary
locating of positions of candidates by energy envelope

Fig. 6. Word boundary detection procedure.

boundary, so we can locate the word boundary more accurately.

(a) Detection of candidates by pitch contour.

The extracted pitch period is transformed into a fundamental frequency (F_0) every 10 ms intervals. This output sequence is smoothed by a 5-point median smoothing method[10]. To treat a global pitch pattern, the smoothed sequence is sampled at every 30 ms interval (down sampling). Further, this new sequence, F0(1), F0(2), ..., is transformed into a 5-value symbol sequence by the slope of pitch contour as follows.

If $\alpha \leqq F0(i) - F0(i\text{-}1)$, then $F0(i) \rightarrow +$
If $0 \leqq F0(i) - F0(i\text{-}1) < \alpha$, then $F0(i) \rightarrow 0^+$
If $\beta \leqq F0(i) - F0(i\text{-}1) < 0$, then $F0(i) \rightarrow 0^-$
If $F0(i) - F0(i\text{-}1) < \beta$, then $F0(i) \rightarrow -$
Otherwise, $F0(i) \rightarrow \Lambda$

That is, if either F0(i) or F0(i-1) is not defined or zero, F0(i) is transformed into a null symbol. This corresponds to unvoiced sounds or silence. For convenience' sake, the symbol "0" represents either '0^+' or '0^-'. If there are three or more successive symbols of '0^+' (or '0^-'), they are replaced by the symbol of '+' (or '—'). If the successive symbols are same, they are merged. The rules of merging process are shown in Fig. 7. The output of this process consists of a sequence of three tuples; symbol, duration and mean value of incliniation of pitch contour.

If there exists a valley of pitch pattern, it is considered that there is a boundary, because the pitch contour becomes a dip at almost all the boundaries. However this dip does not sometimes appear. Therefore, we adopted a strategy which finds the pitch pattern of 'Λ', not 'V'. Fig. 8 shows the rewriting rules. 'W' corresponds to the pitch pattern of a word. The condition of duration assumes that the vocabulary is ten digits. Only this term depends on a vocabulary. Both ends of a word pitch pattern 'W' are regarded as candidates of word boundary. However, these rules overlook the small dip of pitch pattern. Thus, the system also adopts

1) $+\ \underset{1}{0}\ + \longrightarrow +$

2) $-\ \underset{1}{0}\ - \longrightarrow -$

3) $0\ \underset{1}{-}\ 0 \longrightarrow 0^-$

4) $\Lambda\ \underbrace{\underset{1}{0}\ +} \longrightarrow \Lambda\ +$

5) $-\ \underbrace{\underset{1}{0}\ \Lambda} \longrightarrow -\Lambda$

6) $\Lambda\ \underbrace{\underset{1}{0}\ \underset{1}{\Lambda}\ +}\ 0 \longrightarrow \Lambda + 0$

7) $\Lambda\ \underbrace{\underset{1}{-}\ \underset{1}{\Lambda}\ +}\ 0 \longrightarrow \Lambda + 0$

8) $0^+\ 0^- \longrightarrow 0^-$

Fig. 7. Merging rules for a symbol sequence. The unit of length is 30 ms.

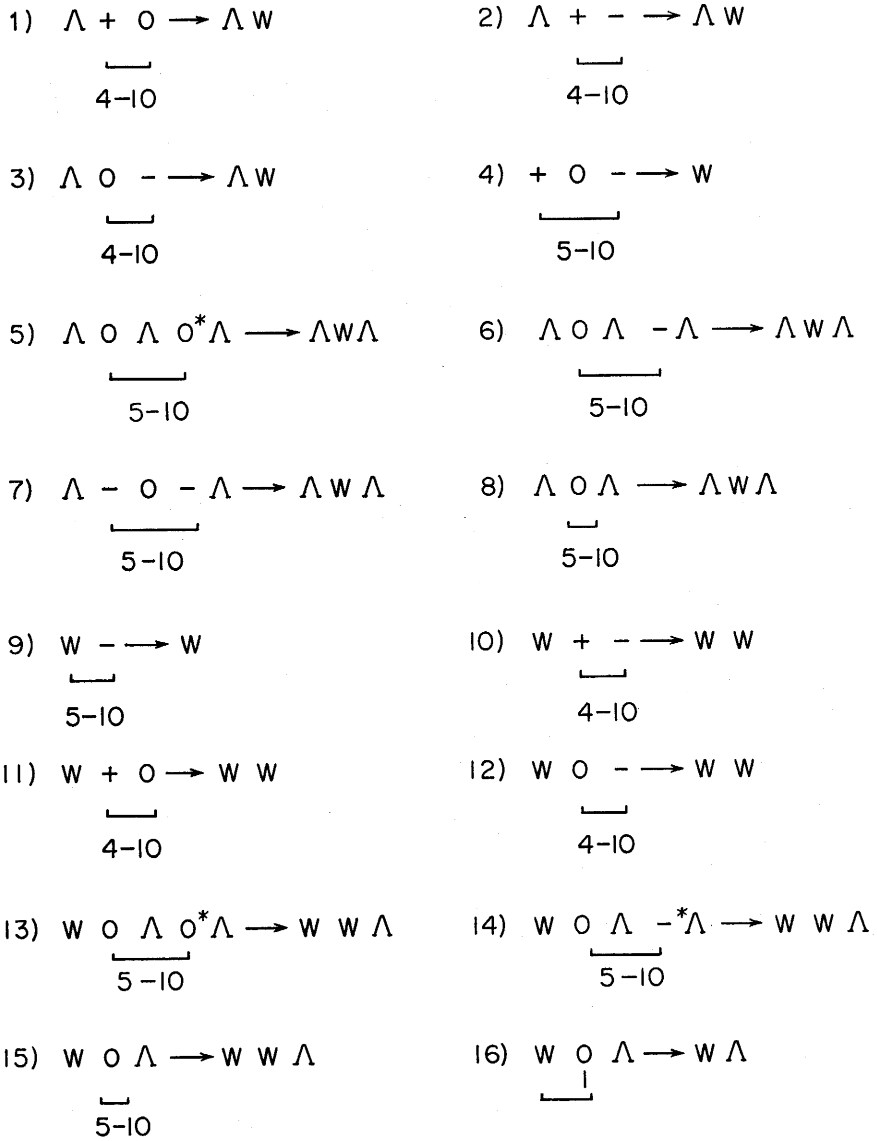


Fig. 8. Rewriting rules which find the pitch patterns of words. The unit of length is 30 ms. The symbol '*' denotes that there should be the difference of pitch freqeuncy more than +10 Hz between the end point of preceding voiced sound and the front point of following voiced sound.

the another rule; a small dip between '0^-' and '0^+' is also regarded as a candidate of a word boundary.

(b) Location of candidates by energy envelope.

The pitch contour usually sifts to a right direction (delay) in comparison with the intention of a speaker, or linguistical acoustic phenomenon. On the other hand, the energy envelope corresponds to that phenomenon. Thus, the energy envelope is used to locate the word boundary more accurately, since the envelope usually shows a valley at the word boundary. But it shows delicate changes in the unvoiced consonant when a word begins at an unvoiced consonant, so the system must consider the case. The search range of valley is from the point preceding 150 ms of candidate to the point following 90 ms. If there is a valley, the point is regarded as a final candidate of word boundary.

V. A RECOGNITION METHOD OF CONNECTED SPOKEN WORDS

The recognition method is based on the isolated word recognition system [1, 2]. Fistly, we explain in breif our isolated word recognition method. The input utterance is classified into a segment string, each of which consists of the first candidate of phonemes, second candidate, reliability and duration. The classified segment (phoneme) string is matched with a phoneme string in a lexical entry by using the similarity between a segment and an element in the entry (we call this word matching). The best matching of all possible associations, which is calculated by a dyanmic programming technique, is regarded as the likelihood of that word. The word which has the largest likelihood in all words is regarded as the input word.

(a) Extraction of word sequence

For convenience' sake, we restrict that the number of words in connected words is no more than 3. Therefore the system assumes that an input utterance contains one, two, or three words. Let B_s and B_f be the beginning point and ending point of 'speech' of an input, respectively. Now, let n be the detected number of candidates of word boundary and also B_1, ..., B_n be their points in connected speech, respectively. The input utterance is divided into several parts according to n as follows, each of which may correspond to a word.

(i) $n=0$

The input is regarded as consisting of one word, that is,

B_s—B_f

(ii) $n=1$

The input is regarded as consisting of one or two words, that is,

B_s—B_f
or B_s—B_1, B_1-B_f

(iii) $n=2$

The input is regarded as consisting of two or three words, that is,

B_s—B_1, B_1—B_f
or B_s—B_2, B_2—B_f
or B_s—B_1, B_1—B_2, B_2—B_f (if B_2—$B_1 \geqq 200$ ms)

(iv) $n=3$

The input is regarded as consisting of two or three words, that is,

B_s—B_1, B_1—B_f

or B_s—B_2, B_2—B_f

or B_s—B_3, B_3—B_f

or B_s—B_1, B_1—B_2, B_2—B_f (if B_2—$B_1 \geqq 200$ ms)

or B_s—B_1, B_1—B_3, B_3—B_f (if B_3—$B_1 \geqq 200$ ms)

or B_s—B_2, B_2—B_3, B_3—B_f (if B_3—$B_2 \geqq 200$ ms)

(v) $m=4$

The input is regarded as consisting of three words, that is,

B_s—B_1, B_1—B_3, B_3—B_f

or B_s—B_2, B_2—B_3, B_3—B_f (if B_3—$B_2 \geqq 200$ ms)

or B_s—B_2, B_2—B_4, B_4—B_f

(b) Modification of classified phoneme string

The portion of a word is extracted by the above procedure, then the classified phoneme string which corresponds that portion is extracted by the following procedure.

(i) If the shortest duration between the extracted word boundary and classified phoneme boundaries is shorter than 30 ms, the phoneme boundary is regarded as a word boundary.

(ii) If the shortest duration is longer than 30 ms, a classified phoneme is divided into two segments at the point of extracted word boundary.

Thus, the system can extract a classified phoneme string corresponding to one word. However, the word boundary detection procedure as mentioned in the previous section often yields a shift of one phoneme caused by the symbol of 'Λ', that is, unvoiced sound. Therefore the extracted phoneme string is modified as follows. Let F_1, F_2, $\cdots$ $F_s \cdots F_t \cdots F_n$, S_1, S_2, $\cdots S_s \cdots S_t \cdots S_n$, and P_1, P_2, $\cdots P_s \cdots P_t \cdots P_n$ be the first candidate of phonemes, the second candidate and reliability of first candidate, respectively, where $P=0$ denotes that there is no difference of reliability between the first candidate and the second candidate. When the phoneme string from s to t corresponds to a word,

(i) If F_s is a vowel and F_{s-1} is an unvoiced consonant,

$S_s \rightarrow F_{s-1}$, $P_s \rightarrow 0$

(ii) If F_s is an unvoiced consonant,

$S_s \rightarrow F_{s+1}$, $P_s \rightarrow 0$

(iii) If F_t is a vowel and F_{t+1} is an unvoiced consonant,

$S_t \rightarrow F_{t+1}$, $P_t \rightarrow 0$

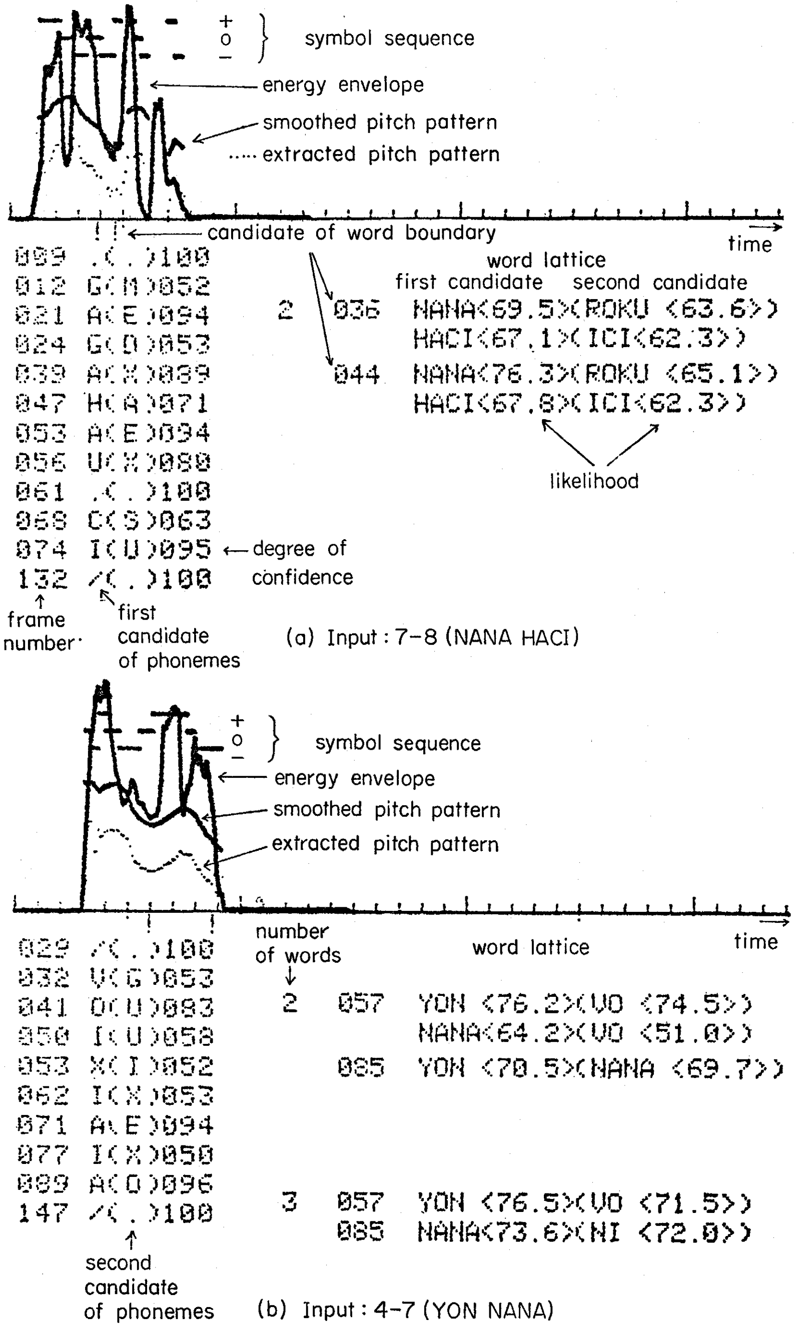
(iv) If F_t is an unvoiced consonant,

$S_t \rightarrow F_{t-1}$, $P_t \rightarrow 0$

This new modified phoneme string is transformed to the word by the word matching procedure.

(c) Evaluation of word lattice

By the method mentioned above, several wrod sequences are generated. The system should select the best word sequence as the recognition result. We adopted



```
009 .( . )100
012 G( M )052
021 A( E )094          2   036  NANA<69.5>(ROKU <63.6>)
024 G( O )053                   HACI<67.1>(ICI<62.3>)
039 A( X )039              044  NANA<76.3>(ROKU <65.1>)
047 H( A )071                   HACI<67.8>(ICI<62.3>)
053 A( E )094
056 U( X )080
061 .( . )100
068 C( S )063
074 I( U )095
132 /( . )100
```

(a) Input: 7–8 (NANA HACI)

```
029 /( . )100
032 V( G )053
041 O( U )083      2   057  YON <76.2>(VO <74.5>)
050 I( U )058               NANA<64.2>(VO <51.0>)
053 X( I )052           085  YON <70.5>(NANA <69.7>)
062 I( X )053
071 A( E )094
077 I( X )050
089 A( O )096
147 /( . )100      3   057  YON <76.5>(VO <71.5>)
                        085  NANA<73.6>(NI <72.0>)
```

(b) Input: 4–7 (YON NANA)

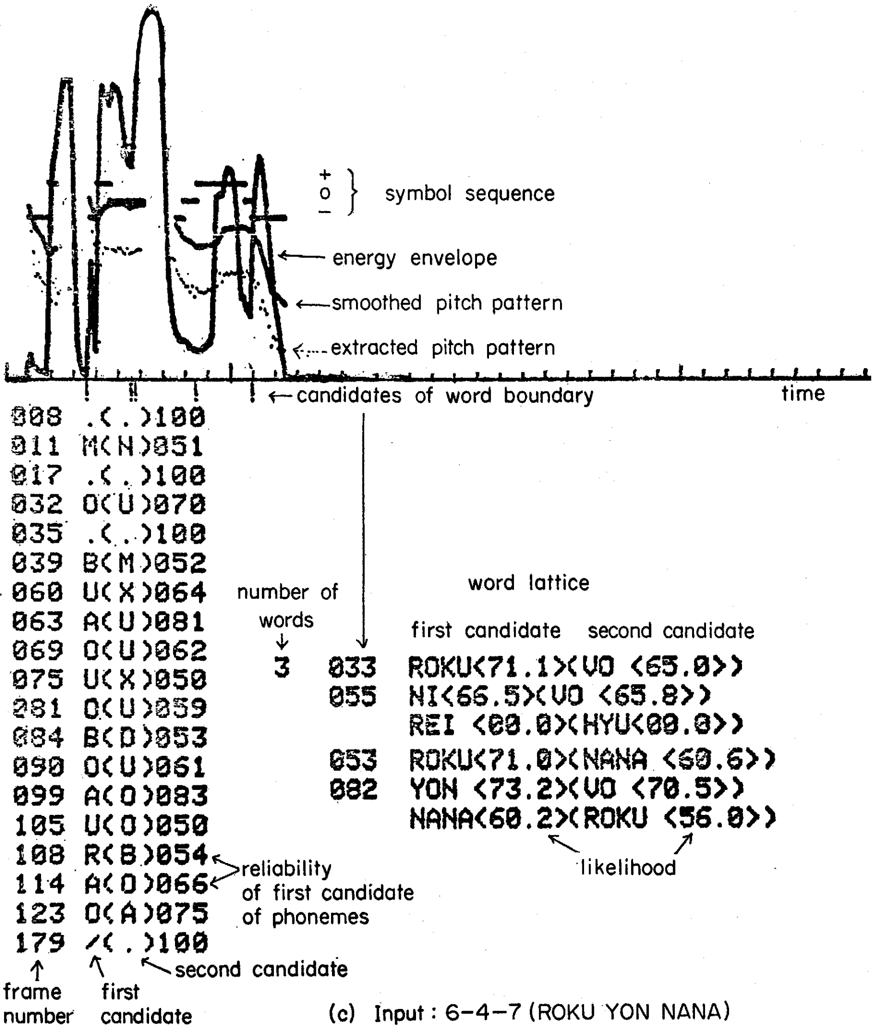
Fig. 9. Examples of recognition process.
X: the syllabic nasal, V: 'η'

the following criterion for the selection. To take preference the longer word sequence than the shorter word sequence, the system adds 1.0 to the average score for two connected words, and 2.0 for three connected words. From all of such new average scores, the system selects the word sequence with the highest average score and it is regarded as the input utterance.

VI. Experimental results

We applied our system to the speech recognition of connected digits. The reference spectral patterns of phonemes and similarity matrix between phonemes were calculated from VCV contexts which were included in 2450 words spoken by other 10 male adults. The experiments described here were not used any learning procedure of speaker differences.

We experimented on the recognition of connected digits spoken by four male adults. Each speaker uttered 10 isolated-digits (one time per digit), 100 two-connected-digits (all pair) and 100 isolated three-connected-digits (selection from

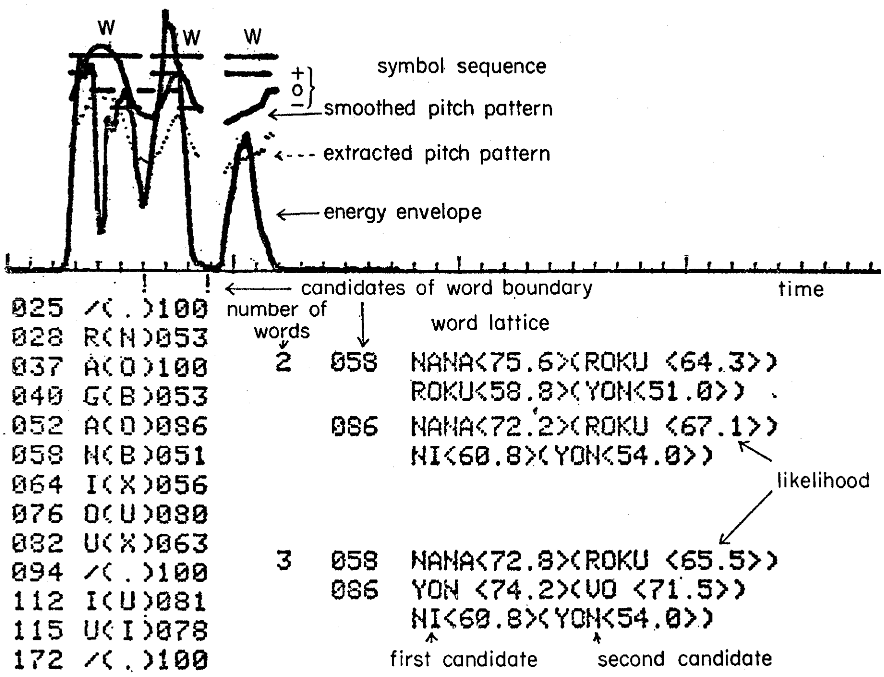


```
008 .(.)100
011 M(N)051
017 .(.)100
032 O(U)070
035 .(..)100
039 B(M)052
060 U(X)064
063 A(U)081
069 O(U)062
075 U(X)050
281 O(U)059
084 B(D)053
090 O(U)061
099 A(O)083
105 U(O)050
108 R(B)054
114 A(O)066
123 O(A)075
179 /(.)100
```

```
3   033   ROKU<71.1>(VO <65.0>)
    055   NI<66.5>(VO <65.8>)
          REI <00.0>(HYU<00.0>)
    053   ROKU<71.0>(NANA <60.6>)
    082   YON <73.2>(VO <70.5>)
          NANA<60.2>(ROKU <56.0>)
```

(c) Input: 6–4–7 (ROKU YON NANA)

Fig. 9.



```
025 /( . )100
028 R( N )053
037 A( O )100    2   058   NANA<75.6>(ROKU <64.3>)
040 G( B )053              ROKU<58.8>(YON<51.0>)
052 A( O )086        086   NANA<72.2>(ROKU <67.1>)
058 N( B )051              NI<60.8>(YON<54.0>)
064 I( X )056
076 O( U )080
082 U( X )063
094 /( . )100    3   058   NANA<72.8>(ROKU <65.5>)
112 I( U )081        086   YON <74.2>(VO <71.5>)
115 U( I )078              NI<60.8>(YON<54.0>)
172 /( . )100
```

(d) Input : 7-4-2 (NANA YON NI)

Fig. 9.

1000 possible strings at random), respectively. The number of digits in an input utterance was restricted up to 3. Fig. 9 illustrates some examples of recognition process. The results are tabulated in Table 4.

The column (a) shows the number detected as candidates of word boundary. The total number in real utterances is 0 for isolated digits, 100 for two-connected digits and 200 for three-connected digits, respectively. The column (b) shows the number detected correctly. The number in parentheses contains the sift of only one phoneme in a classified phoneme string. Such a sift may be correctly processed by the modification of a classified phoneme string described in section 5(b). The column (c) shows the recognition rate.

The number detected as candidates of word boundary was 1.2∼1.7 times of the

Table 4. Recognition results of connected digits

Sspeaker	isolated		2 connected			3 connected		
	a	c	a	b	c	a	b	c
NK	3	100%	168	92(95)	88%	265	177(181)	76%
IN	4	90	167	97(98)	88	276	184(191)	79
MS	2	100	137	88(91)	70	240	170(177)	63
MZ	2	90	139	80(95)	87	271	190(194)	84
average	2.8	95	153	92(95)	83	263	180(186)	76

real number and the boundary was detected correctly at the rate of 85~95%. In the case of connected digits consisting of only voiced sounds, the rate ranged from 90~99%. From this result, we found it is necessary to take unvoiced sounds into consideration. Although the recognition rate of isolated digits was about 95% for unspecific speakers, the rate of two connected digits decreased to about 83%.

VII. Conclusion

One of the roles of prosodic information is to make one word or one phrase understood as one well-arranged utterance. We investigated this theory from view points of listening and reading. On the basis of these investigations, we proposed a new recognition method of connected spoken wrods, which was based on the detection of word boundary. It was performed by using the contours of pitch pattern and energy envelope. The system could detect the word boundaries in two and three connected digits at the rate of more than 90%. This system could recognize isolated spoken digits at the rate of 95% for unspecific speakers. However, the recognition rate decreased to about 83% for two connected digits. If the pitch contour is extracted more precisely, the detection of word boundary will be improved.

Acknowledgment

The authors wish to thank Mr. M. Morimoto for his cooperation of perceptual experiments and Mr. H. Nakase for the development of the connected word recognition method.

References

1) T. Sakai and S. Nakagawa: On Line, Real Time Spoken Words Recognition System with Learning Capability of the Speaker Differences, Studia Phonologica, X, p. 46~59 (1976).

2) S. Nakagawa and T. Sakai: A Pre-Matching Method for a Real Time Spoken Word Recognition System and a Learning Procedure of Speaker Differences, Studia Phonologica, XII, p. 39~58 (1978).

3) H. Sakoe: Recognition of Continuously Spoken Words Based on two Level DP-Matching, Report of the 1975 Spring Meeting of ASJ (1975).

4) R. Nakatsu and M. Kohda: Speech Recognition of Connected Words, Proceedings of the 4-th IJCPR, P. 1009~1011 (1978).

5) L. R. Bahl et al.: Preliminary Results on the Performance of a System for the Automatic Recognition of Continuous Speech, Conference Record of 1976–ICASSP, p. 425~429 (1976).

6) B. T. Lowerre: The Harpy Speech Recognition System, PhD. thesis, Carnegie-Mellon University (1976).

7) M. F. Medress et al.: A System for the Recognition of Spoken Connected Word Sequences, Conference Record of 1977–ICASSP, p. 468~473 (1977).

8) T. Sakai and S. Nakagawa: Continuous Speech Understanding System LITHAN, Studia Phonologica, IX, p. 45~63 (1975).

9) L. R. Rabiner and M. R. Sambur: Some Preliminary Experiments in the Recognition of Connected Digits, IEEE Trans. Vol. ASSP–24, No. 2, p. 170~182 (1976).

10) L. R. Rabiner, M. R. Sambur and C. E. Schmidt: Application of a Nonlinear Smoothing Algo-

rithm to Speech Processing, IEEE Trans. Vol. ASSP–23, No. 6, p. 552～557 (1975).

11) M. Y. Liberman and L. A. Streeter: Use of Nonsense-Syllable Mimicry in the Study of Prosodic Phenomena, JASA, Vol. 63, No. 1, p. 231～233 (1978).

12) L. H. Nakatani: Hearing "Words" without Words: Prosodic Cues for Word Perception, JASA, Vol. 63, No. 1, p. 234～245 (1978).

13) L. A. Streeter: Acoustic Determinants of Phrase Boundary Perception, JASA, Vol. 64, No. 6, p. 1582～1592 (1978).

(Aug. 31, 1979, received)