

Prosody Based Speech Segmentation

Toshie Hatano (t-hatano@lu-tokyo.ac.jp)

Department of Dynamic Linguistics, Graduate School of Humanities and Sociology,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN

Abstract

Two experiments were conducted to verify whether prosody can be a unit of phonological segmentation. In experiment 1, 24 participants were asked to rate meaningless speech imitating 40 meaningful sound sequences produced by one male speaker. It was found that 94.7% of the selected combinations conformed to Japanese accent rules. Similarly, in experiment 2, 19 participants were asked to rate meaningless speech imitating 76 meaningful sound sequences produced by a different male speaker. 92.8% of the combinations selected conformed to Japanese accent rules. These experiments suggest that native speakers of Japanese can also recognize segment boundaries based on prosody.

Keywords: prosody, segmentation, accent phrase, speech perception.

Introduction

How do we break up a complex sound signal of continuous speech into units that can be processed by our mind? This is the ‘segmentation problem,’ which has been investigated in terms of units such as phonemes, syllables, or morae. But is it really possible to process speech just by breaking it into a large number of small units, given real-time limitations? Previous studies suggest that the recognition of word boundaries is based primarily on phonological features and the transitional probability of phonemes or syllables (statistical information) (e.g., Saffran et al., 1999). We propose to expand the current models by investigating the possibility of segmentation that occurs at the level of prosody. In our model (Hatano et al., 2002), we propose that, from the point of view of real-time speech processing, the first step in the recognition process involves detecting units of continuous acoustic signals (or their boundaries). In the present study we hypothesized that the detected units are equivalent to fundamental frequency (F0) information, i.e., the ‘accent phrase’ (AP) (Fujisaki & Sudo, 1972). We conducted experiments to verify whether APs are perceived as unified wholes, even if no phonological indication facilitating language comprehension is provided.

Experiment 1

Method

Participants 24 (9 males and 15 females) students, Japanese native speakers, ranging from 21 to 42 years old.

Selection of morae and accent type 1 accent phrase (referred to as 1AP below) is a sequence of 4 or 5 morae. By combining two APs, AP pairs consisting of 9 morae were

formed. Thus, 20 types of 4+5 morae AP pairs and 20 types of 5+4 morae AP pairs, 40 types in total, were created. AP pairs can be divided into those that only have one possible structural interpretation according to Japanese accent rules (category A) and those which can be interpreted in two ways (category B)(e.g., for LHHHLHLLL, both 4+5 morae LHHH+LHLLL and 5+4 morae LHHHL+HLLL structures are acceptable).

Sequence formation For this experiment, meaningless speech sequences imitating the prosodic pattern of meaningful speech were created. Meaningless speech sequences were created by repeating mora ‘na,’ chosen due to its clear phoneme boundaries.

Speech recording and material The speaker was a 24 year old Japanese student. Immediately after pronouncing each meaningful sequence twice, he was asked to produce a ‘na’ sequence with the same prosodic pattern. The speech material used in the experiment consisted of 10 training session sequences and 80 experiment sequences (40 AP pair types x 2).

Procedure

The experiment was conducted in small groups in a quiet classroom. One speech rating session lasted about 13 minutes. Speech was reproduced from DAT using loudspeakers.

Results

The results of speech rating are shown in Table 1. Within speech category A, the number of 4+5 morae AP pairs rated as 4+5 morae was 719 (93.6%) and the number of 5+4 morae AP pairs rated as 5+4 morae was 715 (93.1%). Within speech category B, the number of 4+5 morae AP pairs rated as 4+5 morae was 163 (84.9%) and the number of 5+4 morae AP pairs rated as 5+4 morae was 62 (32.3%).

Altogether, 1818 (94.7%) of the combinations selected conformed to Japanese accent rules. 1659 (86.4 %) of the combinations selected were the same as the stimulus.

Thus, the results for category A strongly confirm our hypothesis. For category B however, the expected result was 1) the same tendency as category A, or 2) an evenly distributed result of about 50%; but neither of these tendencies was observed in the experiment. A possible explanation is that the random training samples showed a bias in favor of the 4+5 morae pattern. Therefore, in the next experiment the combination of mora patterns in the training samples was adjusted.

Experiment 2

After correcting the problems that appeared in experiment 1, another experiment was conducted to confirm whether the same results can be obtained with different speakers and for a larger number of mora combinations.

Method

Participants 19 (13 males and 6 females) students, Japanese native speakers, ranging from 19 to 43 years old.

Selection of morae and accent type As in experiment 1, 9 morae combinations were created. Apart from AP pairs used in experiment 1, 3+6 morae combinations were also added. Thus, the combinations consisted of 18 types of 3+6 morae AP pairs, 20 types of 4+5 morae AP pairs, 20 types of 5+4 morae AP pairs, and 18 types of 6+3 morae AP pairs, a total of 76 types. Furthermore, among AP pairs that have two possible interpretations according to accent rules (category B), there were 3 patterns of 3+6 morae, 7 patterns of 4+5 morae, 9 patterns of 5+4 morae, and 5 patterns of 6+3 morae.

Sequence formation The procedure was identical to experiment 1.

Speech recording and Speech material The speaker was a 22 year old male, born and raised in Tokyo. The recording procedure was identical to experiment 1. The speech material used in the experiment consisted of 12 training session sequences with a proportional representation of all AP pair combinations and 152 experiment sequences (76 AP pair types x 2).

Procedure

The experiment was conducted in small groups in a quiet classroom. One speech rating session lasted about 25 minutes. Speech was reproduced in the same way as experiment 1.

Results

The results of speech ratings are shown in Table 2. Within speech category A, the number of 3+6 morae AP pairs rated as 3+6 morae was 517 (90.7%), the number of 4+5 morae AP pairs rated as 4+5 morae was 446 (90.3%), the number of 5+4 morae AP pairs rated as 5+4 morae was 389 (93.1%),

and the number of 6+3 morae AP pairs rated as 6+3 morae was 452 (91.5%). Within speech category B, the number of 3+6 morae AP pairs rated as 3+6 morae was 110 (96.5%), the number of 4+5 morae AP pairs rated as 4+5 morae was 241 (90.6%), the number of 5+4 morae AP pairs rated as 5+4 morae was 308 (90.1%), and the number of 6+3 morae AP pairs rated as 6+3 morae was 117 (61.6%).

Altogether, 2680 (92.8%) of the combinations selected conformed to Japanese accent rules. 2580 (89.3 %) of the combinations selected were the same as the stimulus.

Discussion

The results of both experiments suggest that, even given different speakers and different selection options, over 90% of the participants felt that the place of segmentation corresponded to the accentual structure of speech. Such results were observed for category A, where only one segmentation pattern conformed to accent rules.

This suggests that the participants are consistent in their judgment concerning sequences belonging to category A. For category B, with the improved method in experiment 2, segmentation selection corresponded to the structure of meaningful speech. However, it is possible that in category B, when the first AP is long, it is more difficult to judge the place of segmentation correctly.

The above results suggest that prosody may also serve as a cue that allows native speakers of Japanese to recognize segment boundaries.

References

- Fujisaki, H., & Sudo, H. (1972). A generative model for the prosody of connected speech in Japanese. *1972 Conference on Speech Communication and Processing*, 140-143.
- Hatano, T., Horiuchi, Y., & Ichikawa, A. (2002). Preliminary study about cognitive model of Japanese speech. *Tech. Rep. of IEICE, Vol. 102*, 75-80.
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *COGNITION, 70 (1)*, 27-52.

Table 1: Speech rating results (Exp.1)

Stimulus	Responses		Total
	4+5	5+4	
4+5	882 (91.9)	78 (8.1)	960
5+4	183 (19.1)	777 (80.9)	960
Total	1065	855	1920

(): percentage within stimulus group

Table 2: Speech rating results (Exp.2)

Stimulus	Responses				Total
	3+6	4+5	5+4	6+3	
3+6	627 (91.7)	12 (1.8)	34 (5.0)	11 (1.6)	684
4+5	32 (4.2)	687 (90.4)	35 (4.6)	6 (0.8)	760
5+4	8 (1.1)	19 (2.5)	697 (91.7)	36 (4.7)	760
6+3	29 (4.2)	10 (1.5)	76 (11.1)	569 (83.2)	684
Total	696	728	842	622	2888

(): percentage within stimulus group