How does Human segment the Speech by Prosody?

_Toshie Hatano, Yasuo Horiuchi, Akira Ichikawa_

Graduate School of Science and Technology
Chiba University, Japan

{hatano, hory, ichikawa}@icsd4.tj.chiba-u.ac.jp

Abstract

In this study, we introduced a new model of how a human understands speech in real time and performed a cognitive experiment to investigate the unit for processing and understanding speech. In the model, first humans segment the acoustical signal into some acoustical units, and then the mental lexicon is accessed and searched for the segmented units. For this segmentation, we believe that prosody information must be used. In order to investigate how humans segment acoustical speech using only prosody, we performed an experiment in which participants listened to a pair of segmented speech materials, where each material was divided from the same speech material where the two segmentation positions differed from each other, and judged which material sounded more natural. On the basis of the results of this experiment, it is suggested that humans tend to segment speech based on the accent rules of Japanese, and that the introduced model is supported.

1. Introduction

When we listen to a person’s voice, what is the unit which is used for processing and understanding in our brain? Although the voice is a continuous signal in time and is divided only by pauses, we can detect units which correspond to words or phonemes, and understand them as language.

Regarding this concern, there have been some studies which have developed a mental model, but these models have mainly dealt with the phonological structure [1]. In our study, we discuss the model from a slightly different viewpoint, which considers how we segment units from the acoustical speech signal which is continuous in time, and process these units, and moreover, how we perform this task in real time.

We introduce a new processing model for speech understanding by humans. In this model, first we segment the acoustical signal into a number of acoustical units, then, for the segmented units, we access and search the mental lexicon which is stored as acoustical information. In order to realize these processes in our brain, it is required that we segment the acoustical signal. We believe that prosody information could be used as the basis for this segmentation. In this study, we attempt to investigate how humans segment acoustical speech using only prosody information, and we use Japanese speech for the target language.

It is considered that the pitch of Japanese speech is mainly perceived by $F_0$. The pitch is expressed by the high and low accents. When a content word of Japanese is uttered in isolation, its prosodic features can be fully represented by the accent type, denoted by a high and low pattern of $F_0$ for each mora. Although, for $n$-mora words, 2" high and low combinations are possible, the number of accent types in actual use is strongly limited. In the Tokyo dialect (common Japanese), only $n$ accent types are used, which are usually denoted as “type $i$” accents ($i=0, 1, \ldots, n-1$). Table 1 shows all of the accent types for the cases of $n=4$ and $n=5$. The type $i$ accent is characterized by a rapid downfall in the $F_0$ contour around the end of the $i$-th mora, except for the case of $i=0$ which has apparent downfall.

Applying these rules to Japanese speech, an accentual phrase (AP) becomes the unit of the accent rules. The AP is described by various terms: ‘minor phrase’, ‘accent phrase’, ‘accent unit’, ‘minor phonological phrase’, ‘accentual phrase’, and so on [2]. In this study, we treat the accentual phrase (AP) as the unit into which humans segment and process acoustic information.

In our hypothetical model of speech processing, when the acoustical signal reaches human’s ears, their mental function detects a unit or a break (segmentation) and searches the mental lexicon for the detected unit. The lexicon entry is represented as an acoustical signal. For example, the word “yu-ri-ka-mo-me” has the lexicon entry not of a series of phonemes (discrete symbols) “yu”, “ri”, “ka”, “mo”, “me”, but of the continuous acoustical signal (continuous signal) /yurikamome/. For the unit which is used in searching the mental lexicon, we assume the accentual phrase in the case of Japanese.

It is considered that the accent rules [3] is used for extracting the processing unit of speech, but there is no cognitive experiment for investigating this notion. Therefore, in this study we attempt to examine how the AP could be detected based on the accent rules and could act as the unit for accessing the mental lexicon.

For this purpose, two adjacent APs were used for the experiment, because it is difficult to prove directly the hypothesis that an AP is processed as a unit. As speech material, a speech signal which consists of two adjacent APs is divided into two speech segments. There are two possible approaches to this division; one is to divide the speech signal at the original segmentation point where there is a boundary between two adjacent APs, and the other is to divide the speech signal at a different point. In the experiment, participants listened to two speech

---

1In the experiment, meaningless speech samples were used because we would like to investigate only prosodical features.
Table 1: Binary representation of $F_0$ contours for accent types of 4 mora and 5 mora words.

<table>
<thead>
<tr>
<th>Type 0</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
<th>Type 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-mora</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-mora</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

materials corresponding to two types of division as mentioned above and judged which one is more natural. If the AP could be processed as the unit, the speech material divided at the original boundary was expected to be judged more naturally.

2. Methods

2.1. Participants

Twenty-two undergraduates and graduates who were native speakers of Japanese and who were from 21 to 24 years old participated in the experiment. They received a small reward for participating.

2.2. Materials

Accentual type of materials.

In Japanese, words which consist of 4 morae are the most frequent and are often used with a postpositional particle which has one mora, such as /ga/ or /wa/; therefore, words which consist of 4 and 5 morae were selected for materials.

According to the accent rules of Japanese (Tokyo dialect), there are 4 types for the words of 4 morae and 5 types for words of 5 morae (see Table 1).

Materials for the AP-pairs consist of 4 mora words and 5 mora words. Considering the type of accent, there are 20 AP-pairs which consist of 4 mora words and 5 mora words and 20 AP-pairs which consist of 5 mora words and 4 mora words and therefore the total number of morae of each of these AP-pairs are nine. In the experiment, the AP-pairs are divided into 4 morae + 5 morae or 5 morae + 4 morae, where one satisfies the accent rules of Japanese and the other does not in the usual case, which in rare cases both satisfy the rules.

The rare cases in AP-pairs which consist of 4 morae + 5 morae are \{accent type 1 .. 4\} + accent type 1. If the first mora of the following AP (5 morae) is appended to the preceding AP (4 morae), both patterns will satisfy the accent rules of Japanese.

The other rare cases in AP-pairs which consist of 5 morae + 4 morae are \{accent type 0 .. 4\} + accent type 2. If the final mora of the preceding AP (5 morae) is appended to the following AP (4 morae), both of these patterns will also satisfy the accent rules of Japanese.

Character Strings. In the experiment, we prepared meaningful character strings and meaningless character strings. In order to make AP-pairs which satisfy the above-mentioned designs, meaningful character strings for AP-pairs were carefully selected with reference to “NHK Japanese Pronunciation and Accent Dictionary [3]”. As meaningless character strings, the sequence of /na/ (9 mora) was used because of the clear detection of the boundary between adjacent phonemes and of the $F_0$ contour. When meaningful character strings were selected, we excluded words which have the following specific features;

1. Voiceless vowel
2. Nasal sound
3. Assimilated sound (small “tsu” in Japanese)
4. Long vowel
5. Japanese “n”
6. Japanese word of foreign origin
7. Words which have ambiguous accent types

Recordings. All speech materials (both meaningful and meaningless) were recorded in a soundproof room by two male speakers who were 24 years old and 25 years old and who were native speakers of the Tokyo dialect. For all AP-pairs, the speakers first uttered a meaningful AP-pair twice and then immediately uttered a meaningless AP-pair. They were asked never to put a pause in any AP-pair. When they uttered the meaningless AP-pair /na ... /, they were asked to imitate the prosodic pattern in the utterance of the meaningful AP-pair. All speech materials were recorded on digital audio tape and then they were resampled at a 16 kHz sampling rate with 16 bit accuracy.

Speech Materials. In the experiment, only the utterances of meaningless AP-pairs were used as speech materials. There are 40 patterns of AP-pairs as mentioned above. All patterns have 9 morae and each meaningless utterance is divided into both 4 morae + 5 morae and 5 morae + 4 morae, where one what is equivalent to the segment of two words corresponding to the meaningful AP-pair. First, background noise with duration of 2.5 morae was inserted between the 4th mora and the 5th mora of all recorded meaningless speech materials (40 patterns, Table 2) and therefore the resulting speech material consisted of 4 morae voice + 2.5 morae background noise + 5 morae voice. Secondly, in the same way, the background noise of 2.5 morae was inserted between the 5th mora and the 6th mora and the resulting speech material consisted of 5 morae voice + 2.5 morae background noise + 4 morae voice. Finally, there are 80 patterns of speech materials. The set of speech materials which consist of 80 patterns were prepared for every two speakers.

The feature of materials.

In a pair of speech materials, one (OP, original pattern) has the same boundary as the original meaningful speech and satisfies the accent rules in Japanese, and the other (SP, synthetic pattern) has the different boundary, and therefore it can not satisfy the rules except for some cases. The prior and the following segments of the OP always satisfy the accent rules. In the case of original speech which consisted of 5 + 4 morae, the prior segment of SP which consists of 4 + 5 morae always satisfies

---

2Since there was slight noise even in the soundproof room, this background noise was recorded and was used instead of a pause.

3The duration of a mora (seconds/mora) is calculated as the average duration of each speech sample (9 morae).
the accent rules, while the head part of the following segment of SP could be changed and thus this segment could not satisfy the accent rules. In the case of original speech which consisted of 4 + 5 morae, the final part of the prior segment of SP which consists of 5 + 4 morae could be changed and thus this segment could not satisfy the accent rules, and moreover, the head part of the following segment of SP could also be changed and thus this segment also could not satisfy the accent rules.

2.3. Procedure

The above-mentioned speech materials recorded on digital audio tape were played back for the participants using a DAT player (SONY DTC2000ES) and loud speakers (BOSE VM101) in a silent room. Forty pairs of speech materials are played back randomly and each participant judged which material seemed more natural. Their choices were indicated on a response sheet. There were two sets of speech materials spoken by each of the two speakers, therefore, first one set is played and after a 3 minute rest, the other set is used. This experiment took about 30 minutes per participant.

3. Results

Table 3 shows the result of judgement. The column “Rating” indicates the number of participants who judged it more natural. In each pattern, there are two sets of materials uttered by two speakers and therefore the maximum value of this column is 44.

3.1. Trend of Comparative Judgment

Generally, the response of participants had a tendency to choose 5 + 4 mora patterns. This tendency was a common phenomenon for the meaningless speech materials made from the meaningful speech both with 5 + 4 morae and with 4 + 5 morae. As concerns comparative judgment for the speech patterns made from meaningful voices of 5 + 4 morae, the frequency where participants judged more natural for the speech pattern which satisfies the accent rules of Japanese is 84.6% and the average number of participants who judged the speech pattern more natural was 18.7 (SD=1.4). On the other hand, as concerns comparative judgement for the speech patterns made from meaningful voices of 4 + 5 morae, the frequency where participants judged more natural for the speech pattern which satisfies the accent rules of Japanese is 40.9% and the average number of participants who judged more natural was 8.4 (SD=2.9).

3.2. Influence besides the Accent Rule

On the basis of the results of this experiment, it is suggested that the speech patterns which participants judge to be more natural do not always satisfy the accent rules. When a judgement did not satisfy the accent rules, there might be additional reasons.

First we analyzed the 29 patterns which were judged by the significantly large number of participants. 25 patterns (86%) satisfy the accent rules and the other 4 patterns not. For these 4 patterns, we investigated the change of $F_0$ at the division points of the speech materials. At the division points, there are 4 patterns with respect to the change of the $F_0$ contour, namely, ‘H→L’ , ‘L→H’, ‘H→H’ and ‘L→L’. Surprisingly, the above-mentioned 4 patterns are all type ‘L→H’, and therefore we focused on ‘L→H’ patterns and reanalyzed all of the significant patterns. There were 13 ‘L→H’ patterns among the significant patterns and extraordinarily, all of them were judged as sounding more natural by participants. Considering this result, the ‘L→H’ patterns might have the feature which segment speech signal.

4. Discussion

We assumed that human segment acoustical signal into some acoustical units. The purpose of the experiment is to make it clear that acoustical speech signal can be divided into APs as a unit. If the assumption is true, the result which participants judged would satisfy the accent rules. Mostly the judgement satisfied the rules, but some of them did not. Why some participants judged the pattern which can not satisfy the accent rules? As we described in section 3.2, the pattern with ‘L→H’ tend to be judged more natural, but why?

First we discuss the case of the patterns which were made from 5 + 4 morae. All of these patterns were judged by the significantly large number of participants and they satisfied the accent rules. Behind this result, there might be the feature of Japanese accent. In the Tokyo dialect, the first two morae must be ‘L→H’ or ‘H→L’. The patterns which do not satisfy the accent rules are made by dividing the pattern with 5 + 4 morae into 4 + 5 morae, that is, the following pattern (5 morae) is made by adding 1 mora before the original 4 morae pattern. Therefore the first 3 morae pattern of the following pattern is restricted only 4 pairs; (1) LLH, (2) HLH, (3) LHL and (4) HHL. In these 4 patterns, only (3) can satisfy the accent rules and the others cannot. In the Tokyo dialect, the above restriction about the first two morae is very distinctive and therefore it is concluded that these patterns were not selected.

On the other hand, in the case of the patterns which were made from 4 + 5 morae, the different tendency had been seen. The patterns which do not satisfy the accent rules were made by dividing the pattern with 4 + 5 morae into 5 + 4 morae, that is, the prior pattern (5 morae) is made by adding 1 mora after the original 4 morae pattern and the following pattern (4 morae) is made by deleting 1 mora from the head of the original 5 morae pattern. In the 20 patterns, there were 9 patterns judged by the significantly large number of participants where 5 patterns satisfied the accent rules and 4 patterns did not. These 4 patterns were the type ‘L→H’ as mentioned above. As a result of examining these 4 patterns in detail, it is found that the accent type
Table 3: The judgement of the experiment.

<table>
<thead>
<tr>
<th>No</th>
<th>Pattern</th>
<th>Accent Rule</th>
<th>Rating</th>
<th>Border</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LHHHLLH</td>
<td>5+4</td>
<td>1</td>
<td>O</td>
</tr>
<tr>
<td>2</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>11</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>11</td>
<td>31**</td>
</tr>
<tr>
<td>12</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td>13</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>13</td>
<td>32**</td>
</tr>
<tr>
<td>14</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>14</td>
<td>32**</td>
</tr>
<tr>
<td>15</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>15</td>
<td>32**</td>
</tr>
<tr>
<td>16</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>16</td>
<td>32**</td>
</tr>
<tr>
<td>17</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>17</td>
<td>32**</td>
</tr>
<tr>
<td>18</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>18</td>
<td>32**</td>
</tr>
<tr>
<td>19</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>19</td>
<td>32**</td>
</tr>
<tr>
<td>20</td>
<td>LHHHLLH</td>
<td>4+4</td>
<td>20</td>
<td>32**</td>
</tr>
</tbody>
</table>

\[ \text{Table 3: The judgement of the experiment.} \]

5. Conclusion and Future Work

In future, from the viewpoints of our hypothetical mental model, we have to develop the theory to explain the phenomenon that some people feel the segmentation more natural when the pitch accent changes at the divided point, even if it does not satisfy the accent rules.

6. Acknowledgements

We are greatly indebted to Hisashi Shimizu (a graduate of Chiba University in the class of 2000) for programming the speech information processing in this study.

7. References


