ESTIMATING SYNTACTIC STRUCTURE FROM PROSODIC FEATURES IN JAPANESE SPEECH

Tomoko Ohsuga  Masafumi Nishida  Yasuo Horiuchi  Akira Ichikawa
Graduate School of Science and Technology
Chiba University, Japan
(osuga@graduate.chiba-u.jp)

ABSTRACT
In this study, we introduce a method of estimating the syntactic tree structure of Japanese speech on the basis of the F0 contour and the time duration. We introduce a method of estimating the syntactic structure including the following phrase by using the local prosodic features of the first and final part of the leading phrase. This method involves discriminant analysis which is a statistical method based on a large amount of training data. We applied the method to the ATR 503 speech database, and performed discrimination experiments. The results indicated an estimation accuracy of 84% for the branching judgment of each sequence of three leaves. In addition, the accuracy of discrimination saturated when using only the features up to the head part of the second phrase. We consider this result to be fairly good for the difficult task of estimating a syntactic structure that includes a future part on the basis of using only local prosodic features in the past, and also consider prosodic information to be very effective in real-time communication with speech.

1. INTRODUCTION

Prosodic information is very useful for communicating efficiently with listeners in real time. Prosody, which is peculiar to speech, provides many functions, for example, assistance in understanding linguistic information, clarification of the distinction of semantic boundary, and expression of speaker’s intention and emotion. Although syntactic information is mainly represented as phonemic information, we believe that prosody is also effective in enabling the understanding of the syntactic structure of utterances, particularly for read speech.

In previous studies, syntactic information included in prosody was utilized in tasks such as syntactic analysis, the judgment of dependence relations and the detection of syntactic boundaries [1, 2, 3]. The results of those studies have suggested that by using prosodic information effectively, it is possible to improve the accuracy of speech recognition and the semantic understanding of utterances, although in most of those studies, linguistic knowledge of the speech data was also used in addition to prosodic information.

Other researchers have attempted to infer parsing trees only from prosody [4, 5], on the basis of which we previously proposed a method for estimating syntactic structures in Japanese speech using only prosody [6, 7]. In that method, we defined the parameters for the fundamental frequency (F0) contour and the pause duration for the generation of a tree structure, and adopted the statistical discriminant analysis using a large amount of data in order to determine these parameters more objectively.

In this study, to improve the above-mentioned method, we take a new approach from another viewpoint. We have formed the hypothesis that we can infer a syntactic relation with the following phrase by listening only to the leading phrase of speech, and we consider that speech may also include much information in local parts of the leading phrase. Therefore, we define new parameters which express the local features of the top part and the final part of the leading phrase. If the estimation accuracy of the new method is high, our hypothesis will be supported and it will confirm that prosody is very effective in real-time communication.

2. METHOD OF ESTIMATION

We have previously introduced a method for estimating the syntactic structure based on Komatsu et al.’s method [6, 7]. This previous method generates a tree structure from prosody (we call this a prosodic tree), which is close to a syntactic tree. In the following sections, we describe this method simply, and present some new points which are specific to this study.

2.1. Generating a prosodic tree

Initially, we decide on a Prosodic Unit (PU) which can act as a leaf of a prosodic tree. A PU is divided by two local minimal points of an F0 contour or pause [6]. In this study, we use an Accentual Phrase (AP) as the minimum PU because a PU almost corresponds to an AP which represents a tonally delimited unit in Japanese [8]. Therefore, we use an AP to express a PU, hereafter.

A prosodic tree generally consists of several subtrees which have three leaves. We categorize these subtrees into two types: the left-branching tree and the right-branching tree (see Fig. 1).

(a) Left-branching tree   (b) Right-branching tree

Fig. 1. Two types of subtrees.

By identifying the subtree type by means of discriminant analysis, a pair of PUs at a low branch of the subtree is combined into one new PU. By combining PUs repeatedly from left to right, the entire tree structure is gradually formed [6, 7]. This method is effective in that serial processing and more objective estimation by
applying the statistical analysis to a large corpus of speech data are possible.

2.2. Parameters

2.2.1. Global parameters in the previous method

In the previous study, we defined four parameters for the F0 contour and pause duration as variables of the discriminant function. The F0 contour of each AP is represented by a single approximately straight line based on the least squares method. F0 values are treated on a logarithmic scale to allow for variations among speakers.

The four parameters are as follows.

In Fig. 2, we denote the parameters between the first and the second APs by $g_1, l_1, d_1$ and $p_1$, and those between the second and the third APs by $g_2, l_2, d_2$ and $p_2$.

2.2.2. Local parameters in the new method

The above parameters other than $p$ are related to the entire AP, so it is conceivable that these parameters express more global information. On the other hand, humans can understand utterances serially in real time, so it may be considered that speech also includes much information in local parts of the leading phrase. Therefore, we define new parameters which express the local features.

We select two regions as the local parts, one is the first mora of the AP and the other is the last mora of the AP, for the following reasons. First, we took into account a tendency for the mora (or syllable) duration to become longer in front of a strong phrase boundary (this phenomenon is known phonologically as final lengthening). Next, we are also aware of a tendency for the onset position of the phrase to become higher when it follows a strong boundary. Last, we formed a hypothesis that the mora is the unit that governs the F0 contour.

An approximate straight line for the F0 contour is calculated in each vowel part of the first and the last morae, because consonants usually include microprosody. F0 values are also treated on a logarithmic scale.

New local parameters in the first mora of the AP are as follows.

In Fig. 3, we denote the parameters for the first mora of the AP by $s^b, h^b, t^b$.

The parameters for the last mora of the AP are as follows.

We performed an experiment to evaluate the new method using the local parameters, and compared the results with those obtained with the previous method using the global parameters. In addition, we examined the case of using both types of parameters.
3.1. Speech data

Similar to our previous study, for the purpose of investigating the fundamental relationship between prosody and syntactic structure, we used read speech data in this experiment.

We used the ATR Phonetically Balanced Sentences for the assessment of the above-mentioned method. This corpus includes 503 sentences extracted at random from, for example, newspapers, magazines, novels, letters, and textbooks, and thus almost all are declarative sentences and include no emphasis on individual words or phrases. Each sentence was read carefully by ten Japanese professional speakers (six male speakers and four female speakers). The number of sentences we used was 50 for each speaker, for a total of $50 \times 10 = 500$.

This corpus also includes other data such as the F0 value, syntactic structure, phonemic boundaries, and AP boundaries. However, the AP information only exists for one speaker in the database, and thus the data for the nine remaining speakers were obtained by listening.

3.2. Evaluation of the method

We considered the syntactic trees provided by the corpus to be ideal trees, and the discriminant function was obtained using these trees.

A few sentences were removed from the database because they consisted of less than three APs, which is insufficient to generate a subtree. Therefore, the total number of sentences used was 496, and the number of APs was 3267.

We used 10 sentences from each speaker (100 sentences in total) as a test set and the remainder as learning data for the discriminant analysis, and changed the test set in turn (open test). However, sentences were also excluded from each experiment when one or more local parameters of the subtree could not be obtained.

We performed discrimination experiments only at the lowest subtree level in order to evaluate the prosodic features of the local part. Table 1 shows the total number of subtrees composed of three APs, which were used in the following experiments.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Closed test</th>
<th>Open test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local (only first mora)</td>
<td>69.1%</td>
<td>68.1%</td>
</tr>
<tr>
<td>Local (only last mora)</td>
<td>81.8%</td>
<td>80.9%</td>
</tr>
<tr>
<td>Local (both)</td>
<td>85.0%</td>
<td>83.8%</td>
</tr>
<tr>
<td>Global</td>
<td>84.7%</td>
<td>83.9%</td>
</tr>
<tr>
<td>Local &amp; Global</td>
<td>87.3%</td>
<td>86.8%</td>
</tr>
</tbody>
</table>

3.3. Condition of using parameters

We experimented with discrimination for each subtree under various conditions. As shown below, we increased the region of parameters from the head of the subtree, in the following order.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) 1st AP</td>
<td>$s_1^b, h_1^b, t_1^b, 1, d_1, s_2^p, h_2^p, t_2^p$</td>
</tr>
<tr>
<td>(ii) 1st AP + $p_1$</td>
<td>$s_1^b, h_1^b, t_1^b, 1, d_1, s_2^p, h_2^p, t_2^p + p_1$</td>
</tr>
<tr>
<td>(iii) 1st AP + $p_1$ + head of 2nd AP</td>
<td>$s_1^b, h_1^b, t_1^b, 1, d_1, s_2^p, h_2^p, t_2^p + p_1 + 2nd AP$</td>
</tr>
<tr>
<td>(iv) 1st AP + $p_1$ + 2nd AP</td>
<td>$s_1^b, h_1^b, t_1^b, 1, d_1, s_2^p, h_2^p, t_2^p + p_1 + 2nd AP + p_2$</td>
</tr>
<tr>
<td>(v) 1st AP + $p_1$ + 2nd AP + $p_2$</td>
<td>$s_1^b, h_1^b, t_1^b, 1, d_1, s_2^p, h_2^p, t_2^p + p_1 + 2nd AP + p_2 + head of 3rd AP$</td>
</tr>
</tbody>
</table>

When using local parameters of only the last mora, the correct rate of discrimination for subtrees reaches a high accuracy of 81%, which is not significantly worse than the results obtained using global parameters. Moreover, for the case of using all local parameters (see the center row of Table 2), the correct rate is 84% which is about the same as that obtained using global parameters. We consider that this result is reasonably good because this method uses only local prosody information, and moreover, it requires no information about pauses which greatly affect prosody.

On the other hand, in the final result when using both parameters (see the lowest row), there is an improvement of approximately 3% in accuracy compared with the previous method. These results indicate that the syntactic structure can be estimated from prosody information alone, such as the F0 contour and time duration. It is considered that the obtained prosodic tree can appropriately represent the general syntactic structure because no substantial difference exists between the rates of the closed and open tests. This observation also holds true upon comparing the results of speaker-dependent and speaker-independent conditions.

4. RESULTS

4.1. Evaluation of parameters

Initially, we investigate the effectiveness of local parameters. Table 2 shows the correct rate of discrimination at the subtree level in each parameter set. In this table, the upper row shows the results obtained using new local parameters for only the first mora of the AP, only the last mora of the AP, and both of these two regions (as mentioned in Sect. 2.2.2). To allow a comparison, the lower row shows the results in the case of using the previous global parameters (in Sect. 2.2.1) and the case of using these local and global parameters together.
4.2. Results for discrimination

Table 3 shows the correct rate of discrimination at the subtree level under each condition presented in Sec.3.3.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Closed test</th>
<th>Open test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) up to 1st AP</td>
<td>77.2%</td>
<td>75.8%</td>
</tr>
<tr>
<td>(ii) up to 1st AP &amp; p₁</td>
<td>83.3%</td>
<td>83.4%</td>
</tr>
<tr>
<td>(iii) up to head of 2nd AP</td>
<td>83.8%</td>
<td>83.0%</td>
</tr>
<tr>
<td>(iv) up to 2nd AP</td>
<td>87.0%</td>
<td>85.9%</td>
</tr>
<tr>
<td>(v) up to 2nd AP &amp; p₂</td>
<td>87.0%</td>
<td>85.4%</td>
</tr>
<tr>
<td>(vi) up to head of 3rd AP</td>
<td>87.3%</td>
<td>86.8%</td>
</tr>
</tbody>
</table>

In Table 3, it is evident that the correct rate of discrimination saturates under the condition of (iv), which means up to the second AP. To take account of condition (iii), which means up to the head of the second AP, the accuracy is reasonable compared with the results obtained using global parameters, shown in Table 2. Thus it is reasonable to suggest that the syntactic structure of each sequence of three APs can be judged by using only the features up to the second AP. On the basis of these results, it is considered that it might be possible for us to understand the syntactic structure on the way of utterances.

4.3. Typical errors

Some cases in which the judgment of the subtree type seemed particularly difficult were observed. In a typical example, the discriminant function gave an incorrect estimation because a pause was inserted between sequential noun phrases with no particle. In this example, the inserted pause is important for understanding the speech in real time, because it is difficult for us to distinguish between the two phrases without a pause. The syntactic relationship between these phrases is very close in practice, but our method yields a contrary structure owing to the pause, which usually acts to separate phrases and thus influences the peripheral prosodic features. Therefore, this type of error seems to be unavoidable when using our method, although it is observed that the accuracy was slightly improved in the case of using local parameters.

Another typical error is caused by prominence. If a word or a phrase is emphasized, the F0 contour and time duration near the emphasized word or phrase are changed, resulting in an estimation error.

5. DISCUSSION

In Table 2, the obtained accuracy of 84% for a subtree level when using only local parameters surpassed our expectations. This result is comparable to the result obtained using global parameters which also include the features in the following part. In the case of using both global and local parameters, the accuracy of 87% at the subtree level is quite good. We consider this result to be reasonably good since this method uses only local information and also has the difficult condition of using only prosody.

Moreover, from the results obtained under various conditions where the regions of parameters were increased from the head of the subtree in order (see Table 3), it is considered that it might be possible to judge the syntactic structure of each sequence of three APs by using only the features up to the second AP. It may be suggested that this result supports our hypothesis that we can infer a syntactic relationship with the following AP before we actually hear it. It might also be suggested that we perform a predictive function with respect to understanding the syntactic structure.

These results suggest that prosody represents syntactic structure to a large extent, at least in polite read speech such as that used this experiment, and that local prosodic information is very useful for estimating a syntactic structure, and moreover, is helpful in enabling listeners to understand utterances more clearly.

On the other hand, there are some cases in which judgment is very difficult, such as the example described in Sect.4.3. In these cases, it is considered that prosody can provide additional information to make it easier for a listener to understand the speaker’s intention in real-time communication.

6. CONCLUSION AND FUTURE WORK

In this study, we introduced a method of estimating syntactic structure in Japanese speech using prosody information alone, such as the F0 contour and time duration. Through preliminary experiments, it is found that in the case of using the parameters for the local prosodic feature, the method has the estimation accuracy of 84% for judging the branching of each sequence of three APs, even though it does not use pause information which significantly affects prosody. In the case of using the parameters for both local and global prosodic features, the accuracy is increased to 87%. In addition, it is observed that the accuracy of discrimination saturates around the features up to the second AP. In the case of using the parameters up to the head of the second AP, the correct rate reaches a high accuracy of 83%, which is comparable to the results obtained using global parameters. These results indicate that the syntactic structure can be estimated using only prosody.

In the future, to improve this method, we will use other prosodic features such as power and peak of the F0 contour. In order to deal with cases in which it is difficult to apply this method, we will take into account the variations due to interrogative sentences and the typical case described above. Moreover, we plan to take into consideration factors such as semantic structure and prominence to enable the application of the method to spontaneous speech in which prosody is considered to contain much more information than assumed in this study.

7. REFERENCES


