ESTIMATING SYNTACTIC STRUCTURE FROM F0 CONTOUR AND PAUSE DURATION IN
JAPANESE SPEECH

Yasuo Horiuchi    Tomoko Ohsuga    Akira Ichikawa

Chiba University
Graduate School of Science and Technology
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, Japan
(hory@tj.chiba-u.ac.jp)

ABSTRACT
In this study, we introduce a method for estimating the syntactic structure of Japanese speech from F0 contour and pause duration. We defined a prosodic unit (PU) which is bound by a local minimum point of an F0 contour pattern or pause. Combining PUs repeatedly (a pair of PUs is combined into one PU), a tree structure is gradually generated. Which pair of PUs in a sequence of three PUs should be combined is decided by the discriminant function based on the discriminant analysis of many speech data. We applied the method to the ATR 503 Phonetically Balanced Sentences read by four Japanese speakers. As a result, the correct rate of judgement for each sequence of three PUs is 79% and the estimation accuracy of the entire syntactic structure for each sentence is 26%. We consider this result to be fairly good for the difficult task of estimating a syntactic structure only from prosody.

1. INTRODUCTION
Prosody provides much information, i.e., syntactic information, paralinguistic information, nonlinguistic information and so forth. Syntactic information is mainly represented by phonemic information, but although prosody might seem to be redundant in giving syntactic information, we believe that it is very important for understanding speech in real time communication.

In previous studies, [1, 2] researchers have tried to automatically determine boundaries of words or phrases for speech recognition, [3] used prosody for syntactic and dependency analysis and [4] introduced a method which determines dependency relations (modifications) between two adjacent phrases based on prosody. Other researchers [5] have tried to infer parsing trees only from prosody. In this study, we also aim to estimate syntactic structure from only prosody.

In order to generate a tree structure (we call this a prosodic tree) from prosody which is close to a syntactic tree, we improved the method introduced by Komatsu et al. [5, 6]. While the parameters in their method were determined by analyzing only one utterance [6], we introduce a method based on the discriminant analysis of many Japanese speech samples of reading out in order to determine the parameters more objectively.

2. THE METHOD OF KOMATSU ET AL.
Komatsu et al.’s method generates a tree structure only from the F0 contour[5, 6]. In the following sections, we describe how the tree structure is generated.

2.1. Prosodic Unit (PU)
Initially, we must decide on a unit which can act as a leaf of a prosodic tree. A prosodic unit (PU) is bound by a local minimum point of an F0 contour pattern. Komatsu et al. did not take into account pauses in a sentence, but pauses can also divide prosodic units, so in addition to the local minimum point, we regard such pauses as boundaries of PUs[7].

2.2. Line Approximation of F0 Contour
The F0 contour pattern of a PU is represented as a single approximately straight line where the least squares error is between the line and actual F0 values. This line slopes downward in many cases.

2.3. Connection of PUs
An input utterance is divided into PUs (Figure 1(a)) and the contour of each PU is approximated by a straight line (Figure 1(b)). Then, between adjacent pairs of all PUs, the connection rate Ri, described later, is calculated (Figure 1(c)). The connection rate represents the strength of the connection. Among all junctions in an utterance unit, at the junction which has the highest connection rate, the two adjacent PUs are combined into one new PU and the new PU
is approximated by a new line. In this example, $R_3$ has the highest connection rate and the two PUs on the right-hand side are combined. Figure 1(d) shows a new PU, a new approximate line and an interim prosodic tree. Then $R_i$ is calculated again and the above process is repeated until the entire utterance is joined into one PU (see Figures 1(e) to 1(g)).

### 2.4. Connection Rate

The connection rate between two adjacent PUs is introduced as a measure of the strength of the relationship between PUs in terms of prosody. Several parameters contribute to setting the value of the connection rate. Komatsu et al. introduced three parameters, gap of F0, length of leading PU and declining slope (in practice, we represent the F0 contour on a logarithmic scale to allow for variations among speakers). In order to take pauses into account, we added the parameter of pause duration to the above three parameters[7]. The connection rate $R_i$ at point $i$ is defined as follows (Figure 2).

$$R_i = W_g \times g + W_l \times l + W_d \times d + W_p \times p$$

* $g$: gap in the F0 value between the tail of the leading PU and the head of the following PU
* $l$: duration of leading PU
* $d$: difference between the slope of the approximation line and the normal slope (in Japanese, the slope of a normal utterance is empirically $-25$ [Hz/s])
* $p$: pause duration between two PUs
* $W_g$, $W_l$, $W_d$, $W_p$: weighting coefficients
3. OUR NEW METHOD

In order to infer the sentence structure, Komatsu et al. determined the weighting coefficients based on only one speech sample [6] and showed that this heuristic is applicable to a variety of speech inputs including another speaker’s utterances and speech in a foreign language (English, French, Chinese, etc.). In our previous study [7], our aim was to elucidate what kinds of information are expressed by prosody, and hence we applied principal component analysis to spontaneous speech to decide these four coefficients.

In this study, we would like to generate a prosodic tree which is close to a syntactic tree in a more objective manner, so it is desirable to adopt a statistical method and to determine the coefficients based on analysis of many actual speech data. For this purpose, we introduce a new method of generating a prosodic tree.

3.1. Type of Subtree

A prosodic tree generally consists of several subtrees which have three branches. We categorize these subtrees into two types: one type is the left-branching tree and the other is the right-branching tree (see Figure 3).

(a) Left-Branching Tree  (b) Right-Branching Tree

Fig. 3. Two Types of Subtree.

3.2. Connection of PUs

As mentioned in section 2.3, PUs are combined gradually. In previous research [5, 6, 7], the two PUs which have the highest connection rate are combined. We consider this procedure as deciding whether a given subtree is left-branching or right-branching.

The new algorithm is as follows:

(1) Determination of PUs
(2) Line approximation of F0 contour
(3) The first two PUs are pushed into a stack
(4) Obtaining a new PU (it becomes the current PU)

(4-1) If there are no more new PUs, PUs in the stack are combined from right to left by deciding the type of subtree in a similar way to that in (5) below, and when the number of PUs remaining in the stack becomes two, they are combined into one PU and then this algorithm is complete.

(5) Combining PUs by deciding the type of subtree consisting of three PUs in which two are in the stack and one is the current PU.

(5-1) If the subtree is left-branching, then the two preceding PUs are combined into one PU and approximated by a new line and pushed into the stack. If the number of PUs in the stack is no more than one, then the current PU is pushed into the stack and goto (4), otherwise goto (5).

(5-2) If the subtree is right-branching, the current PU is pushed into the stack and then goto (4).

We applied discriminant analysis for deciding the type of subtrees in step (5). The discriminant function is obtained by analyzing many speech data.

4. EXPERIMENT AND RESULTS

4.1. Speech Data

We use the ATR Phonetically Balanced Sentences (503 sentences) read by four Japanese speakers for assessment of the above-mentioned method. The speech data have phoneme labels with precise timestamps, F0 value, syntactic structure (syntactic tree) and so forth. We considered the syntactic trees provided by the corpus to be ideal trees, and the discriminant function is obtained by using these trees. In this experiment, the leaves of the syntactic tree were used as PUs.

The number of sentences is 503 × 4 = 2012, the total number of PUs is 3426 × 4 = 13704 and the number of subtrees is 3604 × 4 = 14416. In these subtrees, there are 7216 left-branching trees and 7200 right-branching trees.

ATR Phonetically Balanced Sentences (503 sentences) consist of 10 sets of phonetically balanced sentences (about 50 sentences). We used one set as a test set and the others as learning data for the discriminant analysis, and changed the test set in turn (open test).

4.2. Results

The correct rate of judgement for each sequence of three PUs was 79.2%. For each sentence, the estimation accuracy of the entire syntactic tree was 26.4%, because if there is one error of discrimination, the entire sentence fails, and therefore the longer the sentence, the lower the estimation accuracy of the entire syntactic tree.
4.3. Typical Errors

One of the typical errors is abbreviation of the Japanese particle ‘to (and)’. For example, in the following Japanese phrase

\[ [N_A \text{ to}] [N_B \text{ no}] [N_C \text{ wo}] \]

where \(N_A, N_B\), and \(N_C\) are nouns, to, no, and wo are particles and [ ] indicates the PU.

(This phrase means \(N_C\) of \(N_A\) and \(N_B\) in English)

the particle ‘to’ can be abbreviated as

\[ [N_A] [N_B \text{ no}] [N_C \text{ wo}] \]

In this abbreviated case, the Japanese speaker usually inserts a pause between \([N_A]\) and \([N_B \text{ no}]\). Although the relation between \([N_A]\) and \([N_B \text{ no}]\) is closer than that between \([N_B \text{ no}]\) and \([N_C \text{ wo}]\) syntactically, the discriminant function judges incorrectly because of the inserted pause. In this case, the pause is important for understanding the speech in real time, because if there is no pause, it is difficult to separate the two words \(N_A\) and \(N_B\).

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

4.4. Neural Net Approach

We attempted to use the neural net approach instead of the discriminant analysis. As a result, the correct rate of judgement for each sequence of three PUs was 80.0% and the estimation accuracy of the entire syntactic tree for each sentence was 26.1%. The levels of correctness for the discriminant analysis and for the neural net approach were not significantly different.

5. DISCUSSION

We consider that the correctness of about 80% is reasonably good because there are some cases where the judgement is very difficult, such as the above typical example (section 4.3). Although the estimation accuracy of entire syntactic structure is 26%, we also consider that this result is fairly good for the difficult task of estimating a syntactic structure from only prosody.

From these results, it is suggested that prosody represents syntactic structure but is not always identical to it. In some cases including the above-mentioned example, prosody can provide to make it easier for a listener to understand the speaker’s meaning and therefore prosody clearly helps listeners to understand speech in real time communication.

6. CONCLUSION AND FUTURE WORK

In this study, we introduced a method of estimating syntactic structure in Japanese speech only from prosody. As preliminary experimental results, the method has the correctness of 79% for the branching judgement of each sequence of three PUs and the estimation accuracy of the entire syntactic tree for each sentence was 26%.

In future, we will use other prosodic characteristics such as power, accentual pattern etc. and we will apply this method to spontaneous speech.

7. REFERENCES


