PROSODIC STRUCTURE IN JAPANESE SPONTANEOUS SPEECH

Yasuo Horiuchi[1] *Akira Ichikawa*[1]

[1]Dept. of Information and Image Sciences, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, Japan

ABSTRACT

In this paper, we introduce a method of generating a prosodic tree structure from the F0 contour of an utterance in order to analyze the information expressed by prosody in Japanese spontaneous dialogue. The connection rate, which means the strength of the relationship between two prosodic units, is defined. By repeatedly combining the two adjacent prosodic units where the rate is high, the tree structure is gradually generated. To determine the parameters objectively, we applied the principal component analysis to 32 dialogues from the Chiba Map Task Dialogue Corpus. Then we applied our method to one dialogue. The results suggested that the prosodic tree based on the first principal component was concerned with the information telling what the speaker wanted to do next and that the prosodic tree based on the second principal component represented the syntactic and grammatical structure.

1. INTRODUCTION

There is much information expressed by prosody in spontaneous speech. If a computer understands this kind of information, we can access the computer in a more familiar manner (i.e. using spontaneous speech). In this paper, we aim to analyze the information expressed by intonation in Japanese spontaneous dialogue.

In order to analyze intonation information, we improve the method introduced by Komatsu et al.[1]. Their method was primarily used for infering the sentence structure (a tree structure) which is calculated from the fundamental frequency (F0) contour and represents some syntactic structure modified by the speaker's feelings. Applying the improved method, we can obtain a prosodic tree. By analyzing the resulting tree, we try to extract some information expressed by intonation in Japanese spontaneous speech. In this paper, we will describe how the prosodic tree is generated and what kinds of information might be represented in the resulting tree structure.

2. MATERIALS AND METHOD

2.1. Materials

As speech data, Komatsu et al. used conversation with a PBX telephone operator (the operator understands what the user says, and connects the user to the specified extension telephone line). However, because of the situation, the conversation is relatively polite. We want to apply their method to more natural spontaneous speech. Therefore we select the Chiba Map Task Dialogue Corpus [2] as the target of our investigation. This corpus is a complete replication, in Japanese, of the Edinburgh HCRC Map Task Corpus [3]. In the Map Task dialogues, two participants talk to each other spontaneously and naturally, induced by the design of the task, in which they look at similar but significantly different maps, each unseen by the other, of the same region. One party with a given route (the instruction giver) instructs the other (the instruction follower) to draw it through the landmarks on the map from the start to the goal.

2.2. Unit of Analysis (Utterance Unit)

Komatsu et al. generated a prosodic tree from a sentence, but in our corpus, a sentence cannot be recognized as a unit of analysis. The speech used by Komatsu et al. is polite, so a sentence can be recognized easily, while our speech is very natural and familliar, and there are various phenomena, such as repairs, hesitations and interruptions by the other speaker, which make utterances ungrammatical. Hence as a unit of analysis, we use a pause-bound unit which is a stretch of a single speaker's speech bound by pauses of more than 400 milliseconds long. We call such a unit an utterance unit for convenience.

2.3. Generation of Prosodic Tree

Komatsu et al.'s method generates a tree structure from the F0 contour. In the following paragraphs, we describe how the tree structure is generated.

Prosodic Unit (PU) First of all, we must decide on a unit which can be act as a leaf of a prosodic tree. A prosodic unit (PU) is defined as a stretch of a single speaker's speech bound by a local minimum point of an F0 contour pattern. Komatsu et al. took no account of pauses in a sentence, but when we apply this method to our corpus we must take into account pauses in an utterance unit (i.e., pauses less than 400 milliseconds long). In addition to the local minimum point, we regard such pauses as boundaries of PUs. PUs may be sentences, clauses, phrases, words or parts of words.

Line Approximation of F0 Contour The F0 contour pattern of a PU is represented as a single approximate straight line which has the least squares error between the line and actual F0 values. This line is usually a declining line.

Connection of PUs An input utterance unit 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 R_i, 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 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 whole utterance unit 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. Komatsu et al. introduce three parameters (Gap of F0, Length of leading PU and Declining slope). In order to examine natural spontaneous speech, we add the parameter of Pause duration to the above three parameters. The connection rate R_i at the point i is defined as follows[1] (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 [Hz]
l: length of leading PU [s]
d: difference between the slope of the approximation line and the normal slope[2] (absolute value)[Hz/sec]
p: pause duration between two PUs [s]
W_g, W_l, W_d, W_p: weighting coefficients

2.5. Weighting Coefficients

In order to infer the sentence structure, Komatsu et al. determined the weighting coefficients based on only one speech example [4] and showed that this heuristic manner is applicable to a variety of speech input including another speaker's utterances and speech in a foreign language (English, French, Chinese etc.).

Our aim is not to make a parsing tree but to elucidate what kinds of information are expressed by prosody. Hence it is desirable to adopt some statistical method and to determine the coefficients based on many actual dialogues. For this purpose, 32 dialogues (16 different male speakers) are selected from the Chiba Map Task Dialogue Corpus.

First, we investigated the correlation between each parameter. There is only a slight correlation between d and l and no correlation between any other two parameters. This means that a set of four parameters is reasonable for analyzing spontaneous speech.

Second, we applied principal component analysis to utterances in these 32 dialogues, and then two sets of coefficients based on first and second principal components were calculated (Table. 1).

[1]In practice, we represent the F0 contour on a logarithmic scale to allow variations of speakers.

[2]In Japanese, the slope of a normal utterance is -25 [Hz/s] empirically.

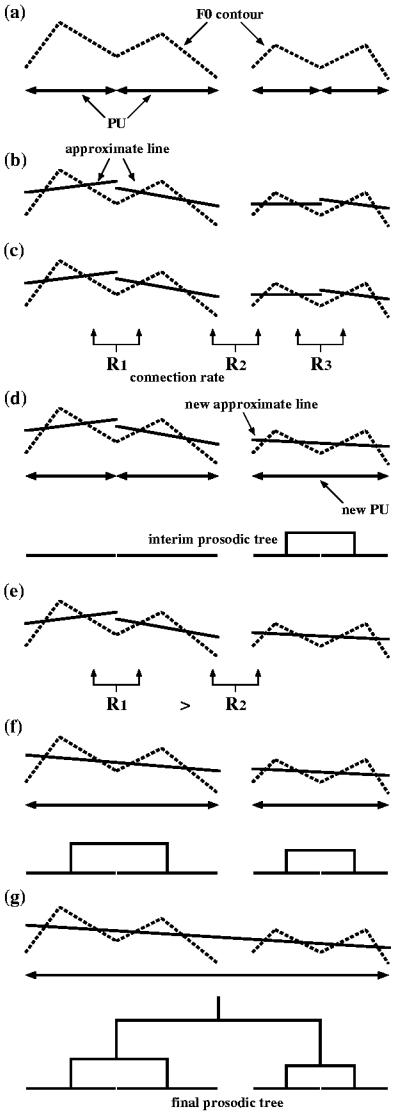

Figure 1: Generation of a prosodic tree.

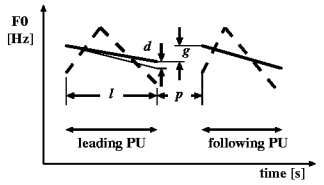

Figure 2: Four parameters.

Table 1: Weighting coefficients based on first and second principal components.

	W_d	W_l	W_g	W_p
1st	0.65	-0.65	0.17	0.20
2nd	0.18	-0.20	-0.67	-0.60

3. EXPERIMENTAL RESULTS

Using the set of parameters, we apply our method to one dialogue selected from the above 32 dialogues. The number of utterance units is 150 and the total number of PUs is 512. Figure 3-6 show examples of the resulting prosodic tree. Within the figures, the utterances are written using Japanese kana character. The first example is "de (Then,) tu ki ma si ta ra (when you reach there,) ha i o ku no u e o (over the abandoned cottage,) to o t te (pass by.)". Figures 3 and 4 show the prosodic trees based on first and second principal components, respectively. The second example is "a (Ah,) de wa de wa (then,) so re zya na ku te (it is not that,) zya (anyway,) ko t ti no i u to o ri ni (as I say,) si te (do that,) ku da sa i (please.)". Figures 5 and 6 show the prosodic trees based on first and second principal components, respectively.

3.1. Prosodic Tree Based on Second Principal Component

First we examine the trees based on the second principal component. The tree from the 1st example is constructed in the following way

[[[de tu ki ma si ta][ra]] (Then, when you reach there,)
[[ha i o ku no u e o] (over the abandoned cottage,)
[[to o t][te]]]] (pass by.)

The tree from the 2nd example is as follows

[[[[a][de wa de wa]] (Ah, then,)
[[so re zya na ku][te]]] (it is not that.)
[[[zya ko t ti no][i u to o ri ni]] (Anyway, as I say,)
[[si te][ku da sa i]]]] (do that, please.)

Looking at these structures carefully, the trees based on the second principal component seem to represent syntactic and grammatical information. These trees are similar to the prosodic structure introduced by Komatsu et al.

3.2. Prosodic Tree Based on First Principal Component

The trees based on the first principal component are quite different from the above trees. The tree from the 1st example is as follows

[[[de tu ki ma si ta] (And you reach there.)
[[ra][ha i o ku no u e o]]]
(Then, over the abandoned cottage.)
[[to o t][te]]] (Pass by.)

The tree from the 2nd example is as follows

[[[[a][de wa de wa]] (Ah, then.)
[[so re zya na ku] (It is not that.)
[[te][zya ko t ti no]]]]
(Therefore, pay attention to my words.)
[[i u to o ri ni] (In accordance with my instruction,)
[[si te][ku da sa i]]]] (do that, please.)

In the above representation, English texts within brackets are not the translation. They represent the Japanese nuance expressed by the PUs. Of course the units are not correct in terms of general Japanese grammar, for example, the division of [ta] and [ra] in the first example is not allowed.

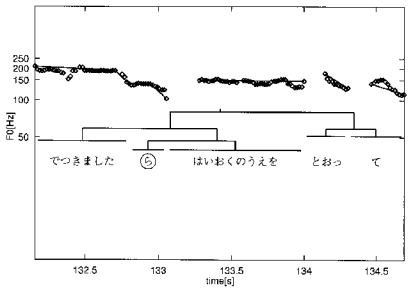

Figure 3: The prosodic tree from the 1st example based on first principal component

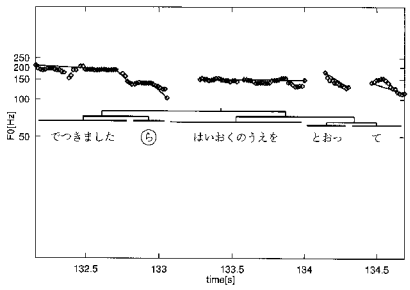

Figure 4: The prosodic tree from the 1st example based on second principal component

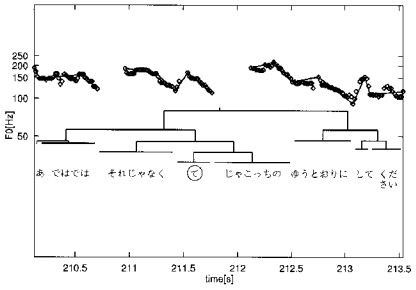

Figure 5: The prosodic tree from the 2nd example based on first principal component

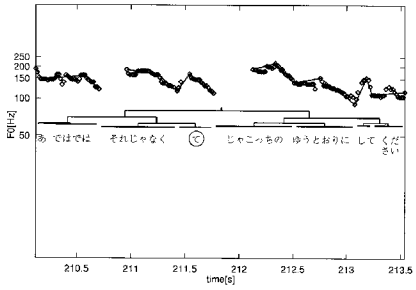

Figure 6: The prosodic tree from the 2nd example based on second principal component

When we examine these two trees without regard to Japanese grammar, some interesting findings are revealed. In the first example, the PU [de tu ki ma si ta] is complete enough to give the information "you reach there". The next PU [ra] (then) tells the listener that there is more information which the speaker wants to impart and the PU is combined with the following PU [ha i o ku no u e o] (over the abandoned cottage). After this PU, the listener seems to understand what the speaker wants to say, i.e. "over the abandoned cottage, go or pass by". It is suggested that because the listener predicts the next PU, the following PU ([[to o t][te]]) is not combined with the leading PU ([ha i o ku no u e o]).

From this viewpoint, we can understand the 2nd example. When the speaker says [a de wa de wa so re zya na ku] (Ah, then. It is not that.), the listener understands what he wants to say (i.e., "You are wrong about it"), and [te] (therefore) indicates to the listener that there is more information. When the speaker says [zya ko t ti no], he seems to instruct the listener to pay attention to his words, so the PU is combined with the leading PU [te] and not with the following PU [i u to o ri ni si te ku da sa i] (Please do that as I say). In this example, it is suggested that the speaker intends for the listener to pay attention before he says important information.

4. DISCUSSION

As mentioned above, the trees based on the 2nd principal component represent the syntactic and grammatical information. The set of coefficients is $W_d = 0.18, W_l = -0.20, W_g = -0.67$ and $W_p = -0.60$. Namely, g and p mainly contribute to generating the tree, and both are governed by the relation between the leading PU and the following PU. When a new grammatical phrase is started (i.e. at the begining of a sentence or a clause), g might be greater. Speakers have a tendency to take a breath when a grammatical phrase is ended (i.e. after a sentence or a clause), hence p might be longer.

In contrast, the trees based on the 1st principal component are concerned with the information which the speaker tells the listener. The set of coefficients is $W_d = 0.65, W_l = -0.65, W_g = 0.17$ and $W_p = 0.20$. Namely, d and l mainly contribute to generating the tree, and both are governed by only the leading PU. Keep in mind the observation that these trees indicate that the speaker has more information to tell and that he wants to continue to speak or that the speaker intends to the listener to pay attention before he says important information. Thus it is suggested that this kind of information are expressed in only the leading PU and the listener understands the speaker's intention by listening to only the leading PU. These suggestions are in agreement with our intuition that when we listen to someone's utterance, before his utterance is ended, we can understand whether he will continue to speak or not, and that before important information is spoken, we can infer that we must pay attention.

In our investigation, the syntactic or grammatical information is expressed based on the 2nd principal component, not the 1st principal component. This may be because this kind of information are mainly expressed on phonetic information, whereas prosodic information only supports this information.

5. CONCLUSION

In this paper, we introduced a method of generating a prosodic tree, improving the method presented by Komatsu et al. In order to avoid subjectivity, we applied the principal component analysis to 32 dialogues and then determined the set of coefficients. It was suggested that the prosodic trees based on the first principal component were concerned with the information indicating the speaker wants to do next, and that the tree based on the second principal component had a syntactic and grammatical structure similar to that determined by Komatsu et al.

The results suggest that prosody is concerned with the semantic and rhetorical structures [5] in spontaneous speech, although there is not enough evidence to prove this at present. It is reasonable that prosody helps the listener to understand, in real time, these kinds of information embedded deeply in the speech. Our method is suitable for realtime processing, because the connection rate is simply calculated using four independent parameters. We must investigate whether this method of calculating the connection rate is appropriate. We will analyze many other spontaneous dialogues in terms of prosody using this method. This kind of information may be applied not only in speech recognition systems, but also in speech synthesis to generate utterances that are more natural and easy to understand.

6. REFERENCES

1. A. Komatsu, E. Oohira, and A. Ichikawa. Prosodical Sentence Structure Inference for Natural Conversational Speech Understanding. In *Proc. of Eurospeech*, pages 400–403, 1989.
2. Y. Horiuchi, A. Yoshino, M. Naka, S. Tutiya, and A. Ichikawa. The Chiba Map Task Dialogue Corpus Project. *Journal of Faculty of Engineering, Chiba University*, 48(2):33–60, 1997. (in Japanese).
3. A. H. Anderson, M. Bader, E. G. Bard, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.
4. A. Komatsu, E. Oohira, and A. Ichikawa. Conversational Speech Understanding Based on Sentence Structure Inference Using Prosodics, and Word Spotting. *Trans. IEICE Japan*, J71-D(7):1218–1228, 1988. (in Japanese).
5. W. C. Mann and S. A. Thompson. Rhetorical structure theory. In G. Kempen, editor, *Natural Language Generation*, chapter 7, pages 85–95. Martinus Nijhoff, 1987.