SYNTACTIC INFORMATION CONTAINED IN PROSODIC FEATURES OF JAPANESE UTTERANCES

Kazuhiko Ozeki, Kazuyuki Kousaka, and Yujie Zhang

The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182 Japan
{ozeki, kousaka, zhang}@achilleus.cs.uec.ac.jp

ABSTRACT

This paper is concerned with measuring the amount of syntactic information contained in prosodic features of Japanese utterances. Five prosodic features are employed, and the statistical relationship between those features and the inter-phrase dependency distance is estimated by using training data. Then parsing experiments are conducted in two different ways: one utilizing the posterior distribution of the inter-phrase dependency distance given the prosodic feature values, and the other without using such information. It has been shown that significant improvement in parsing accuracy is attained by utilizing the prosodic information, and that the duration of pause between adjacent phrases is more effective than prosodic features related to the fundamental frequency and the power.

1. INTRODUCTION

There is a certain relationship between prosody and syntax in Japanese [1] as well as in other languages [2]. In fact, when we speak a Japanese sentence, we can express, *to some extent*, its syntactic structure by such prosodic features as pause and intonation.

There have been many studies to utilize the relationship between prosody and syntax in the field of speech synthesis. Hirose *et al.* [3], for example, have described a method of generating prosodic parameters from the result of syntactic analysis of written Japanese sentences. A couple of other authors have addressed the inverse problem: how to reconstruct the syntactic structure by using the prosodic information. Komatsu *et al.* [4] defined a heuristic measure of inter-phrase association strength. Then, by dividing sentences at phrase boundaries in ascending order of the association strength, they obtained something like parse trees. Sekiguchi *et al.* [5] reported that by exploiting the prosodic features, it was possible to judge if two adjacent phrases were in modification relation 75% correctly.

In this paper, we attempt to measure the effectiveness of prosodic information in parsing of Japanese utterances. We employ five prosodic features, related to the fundamental frequency, the power, and the pause, to represent prosodic information, and define a measure of inter-phrase association strength on the basis of posterior distribution of the inter-phrase dependency distance given the prosodic feature values. The parameters of the posterior distribution function are estimated from a speech database in a training stage [6]. Then full parsing of read Japanese sentences is conducted using the inter-phrase association strength for linguistic knowledge. A novel parser called the *minimum total penalty method* [7] is employed, which finds the most probable syntactic structure efficiently based on given syntactic constraints and the inter-phrase association strength. The results of parsing are compared with baseline results, where no prosodic information is involved, to measure the performance gain attained by incorporating the prosodic information.

2. SYNTACTIC ANALYSIS

2.1. Dependency Structure of Japanese

The structure of a Japanese sentence can be looked from a dependency grammatical point of view; that is, it can be described by specifying which phrase modifies (in a wide sense) which phrase in a sentence. A phrase here is a syntactic unit called *bunsetsu* in Japanese, consisting of a content word with or without being followed by a string of function morphemes such as particles and auxiliary verbs. If phrase X modifies phrase Y in a sentence, X is called a modifier, and Y its head. In Japanese, a modifier normally precedes its head. There are various syntactic constraints governing the modification relation, which are divided into the following two categories.

2.1.1. Global Syntactic Constraints

There are two major global constraints that govern the syntactic structure of a whole sentence:

- Each non-final phrase in a sentence has one and only one head.

- Two modifier-head pairs never cross with each other. For example, in a phrase sequence $X_1 X_2 X_3 X_4 \cdots$, it is impossible that X_1 modifies X_3, and X_2 modifies X_4, simultaneously.

A global modification structure satisfying the above global syntactic constraints is referred to as a *dependency structure.*

2.1.2. Local Syntactic Constraints

There are another type of syntactic constraints, which govern modification relation between two phrases independently of the global syntactic constraints. For example, an adjective can modify phrases starting with a noun only, or phrases starting with a verb or adjective only, depending on its inflection. Thus, whether or not phrase X is allowed to modify phrase Y, with no semantic factors taken into consideration, is decided basically by the combination of the last morpheme (and its inflection if it is an inflecting morpheme) in X and the part of speech of the content word in Y.

2.2. Minimum Total Penalty Parsing

Parsing here is a process of deciding the most probable dependency structure taking the local syntactic constraints and prosodic information into account. Since prosodic information takes an analogue form in its physical manifestation, we need a parser that can treat continuous quantities as linguistic knowledge. For that reason, we have employed the *minimum total penalty method*[7].

In this method, a *penalty function* $F(X,Y)$, the value of which represents the difficulty for phrase X to modify phrase Y, is prescribed. The function value is non-negative, and should be small if X and Y are tightly associated. The penalty function values are added up over all the modifier-head pairs in a dependency structure, yielding the *total penalty.* Then the dependency structure that gives the minimum total penalty is selected. For example, a phrase sequence $X_1 X_2 X_3 X_4$ has five possible dependency structures:

$$\begin{aligned}
S_1 &: (X_1,X_2),(X_2,X_3),(X_3,X_4),\\
S_2 &: (X_1,X_2),(X_2,X_4),(X_3,X_4),\\
S_3 &: (X_1,X_3),(X_2,X_3),(X_3,X_4),\\
S_4 &: (X_1,X_4),(X_2,X_3),(X_3,X_4),\\
S_5 &: (X_1,X_4),(X_2,X_4),(X_3,X_4),
\end{aligned}$$

where (X_i,X_j) denotes a modifier-head pair. The total penalty of S_1, for example, is given as

$$R(S_1) = F(X_1,X_2) + F(X_2,X_3) + F(X_3,X_4).$$

If $R(S_k)$ is the minimum among $\{R(S_j)\}$, then S_k is selected as the most probable dependency structure.

Although the number of dependency structures grows exponentially with respect to the phrase sequence length, this combinatorial optimization problem can be solved in polynomial-time by using the principle of dynamic programming [7].

The penalty function is defined on the basis of posterior probability of the inter-phrase dependency distance given prosodic feature values, as will be discussed later.

2.3. Deterministic Parsing

We have employed another parsing method called the *deterministic analysis method*[8]. Since this method does not involve prosodic information, the parsing result can be a baseline in evaluating the effectiveness of prosodic information.

In this method, the analysis proceeds backward starting with the last non-final phrase in the sentence. For each non-final phrase X, its head is search for, the global and local syntactic constraints being taken into consideration. If there are more than one phrases that can be the head of X, the one that is closest to X is selected. In this way a dependency structure is decided deterministically.

It should be noted that the same global and local syntactic constraints are imposed in both of the minimum total penalty parsing and the deterministic parsing.

3. PROSODIC INFORMATION

3.1. Prosodic Features

Let A be a phrase in a sentence, and B its immediate successor. We have employed five prosodic features associated with A, which are defined in relation to B in the following way [6]:

- *Pause* is the interval between the ending time of A and the starting time of B.

- *Pitch-Gap* is the difference between the ending value of f_A and the starting value of f_B, where f_A and f_B are regression line segments fitted to the log-pitch contours of A and B, respectively.

- *Pitch-Slope* is the regression coefficient for the log-pitch contour of A.

- *Power-Gap* is the difference between the ending value of g_A and the starting value of g_B, where g_A and g_B are regression line segments fitted to the log-power contours of A and B, respectively.

- *Power-Slope* is the regression coefficient for the log-power contour of A.

Thus we have an n-dimensional prosodic feature vector $\boldsymbol{p}_n = (p_1, \ldots, p_n)$ associated with each non-final phrase in a sentence, where p_i is one of the above five features. When all the features are used, the dimension n equals 5.

3.2. Definition of Penalty Function

For a phrase pair (X, Y), the penalty function value $F(X, Y)$ is defined as

$$F(X,Y) = \begin{cases} -\log P(d \mid \boldsymbol{p}_n), & \text{if } (X,Y) \in LSC, \\ \infty, & \text{otherwise,} \end{cases}$$

where d is the inter-phrase distance between X and Y in the sentence to be analyzed, and $P(d \mid \boldsymbol{p}_n)$ is the posterior probability of d given the prosodic feature vector $\boldsymbol{p}_n$ for X [6, 9]. The notation $(X, Y) \in LSC$ signifies that X is allowed to modify Y by the local syntactic constraints.

By assuming that $\boldsymbol{p}_n$ follows a Gaussian distribution for a given d, the parameters (the mean vector and the covariance matrix) of $P(\boldsymbol{p}_n \mid d)$ can be easily estimated from training data [6]. The probability $P(d)$ is also estimated by the relative frequency of phrase pairs having dependency distance d. Then, calculation of the posterior probability $P(d \mid \boldsymbol{p}_n)$ is straightforward by using the Bayes' rule

$$P(d \mid \boldsymbol{p}_n) = \frac{P(\boldsymbol{p}_n \mid d)P(d)}{\sum_d P(\boldsymbol{p}_n \mid d)P(d)}.$$

4. EXPERIMENTAL RESULTS

4.1. Speech Material and Evaluation Criteria

An ATR speech database (Set B) was used for speech material. It contains 503 Japanese sentences taken from newspapers, magazines, and etc., read by two male and two female professional narrators. These sentences are divided into 10 groups A∼J. In Case (a), the groups A∼E, containing 1497 modifier-head pairs, are used for training data, and F∼J, containing 1426 modifier-head pairs, for test data. In Case (b), the training groups and the test groups are interchanged. All the sentences are annotated with the correct dependency structure, with which the parsing result is to be compared.

We define some criteria concerning evaluation of parsing results.

- *Parsing accuracy* is the number of test sentences whose dependency structures determined by parsing coincide with the annotation in the database, divided by the total number of test sentences.
- *Dependency accuracy* is the number of non-final phrases whose heads determined by parsing coincide with the annotation in the database, divided by the total number of non-final phrases in the test sentences.
- *Adjacency accuracy* is the accuracy of judgement if two adjacent phrases are in modification relation.

The evaluation results are averaged over the four speakers, and over the two cases (a) and (b).

4.2. Effectiveness of Prosodic Information

Table 1 shows the parsing accuracy (Par), the dependency accuracy (Dep), and the adjacency accuracy (Adj) for the minimum total penalty parsing (Prosody Used), and the deterministic parsing (Prosody Not Used). Under any criterion, a higher performance has been obtained when the prosodic information is used.

Table 1. The parsing accuracy (Par), the dependency accuracy (Dep), and the adjacency accuracy (Adj).

Prosody	Par(%)	Dep(%)	Adj(%)
Used	52.6	86.7	93.1
Not used	45.8	83.6	91.7

4.3. Comparison of Prosodic Feature Combinations

In order to investigate which prosodic feature is more effective than others, parsing experiments have been conducted for various combinations of the prosodic features. Table 2. shows that the *pause* is more effective than the *pitch* and the *power*. When the *pause* is included in the feature vector, the performance is high, otherwise low. It is also noted that no performance improvement has been made by adding other features to the *pause*.

Table 2. The performance for various combinations of prosodic features. *Pitch* means *pitch-gap* and *pitch-slope* combined, and *power* means *power-gap* and *power-slope* combined.

Feature(s)	Par(%)	Dep(%)	Adj(%)
Pause	53.2	87.0	93.2
Power	49.5	86.0	91.4
Pitch	50.5	86.2	91.5
Pause + Power	53.2	86.8	93.2
Pause + Pitch	53.0	86.8	93.1
Power + Pitch	48.1	85.4	91.1
All	52.6	86.7	93.1

4.4. Parsing Accuracy vs. Sentence Length

Fig.1. shows how the parsing accuracy changes with the sentence length. Even if the prosodic information is used, long sentences are difficult to parse. It should be noted, however, that the performance is consistently improved with the use of prosodic information over the whole range of sentence length.

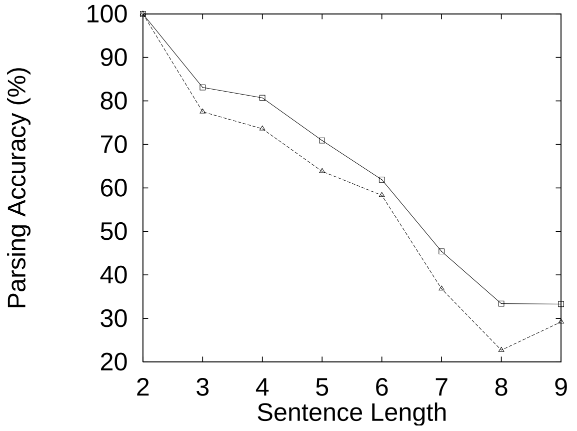

Fig.1. Parsing accuracy vs. sentence length.
Solid line: prosodic information used.
Dotted line: prosodic information not used.

4.5. Dependency Accuracy vs. Dependency Distance

Fig.2. illustrates the dependency accuracy as a function of inter-phrase dependency distance. The graph shows that long dependency distance is difficult to predict. However, prosodic information is still effective over virtually the whole range of dependency distance.

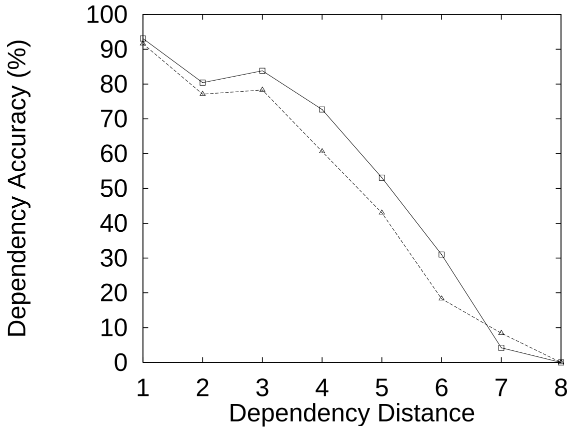

Fig.2. Dependency accuracy vs. dependency distance.
Solid line: prosodic information used.
Dotted line: prosodic information not used.

5. CONCLUSION

Five prosodic features were chosen, and the posterior probability distribution of the inter-phrase dependency distance given those feature values was estimated by using training data. By using the distribution together with the basic global and local syntactic constraints for linguistic knowledge, parsing experiments were conducted for read Japanese sentences. The results were compared with baseline results where no prosodic information was involved.

It has been shown under various evaluation criteria, and for various sentence length, that the prosodic features contain a significant amount of syntactic information. Among the employed prosodic features, the duration of pause was more effective than other features related to the fundamental frequency and the power. It does not follow immediately, however, that the fundamental frequency and the power contain less amount of syntactic information than the pause does, because only a part of relevant information they contain might have been utilized in this work.

Thus, our research plans include a search for physical features that better represent prosodic information, as well as their better use for syntactic information source. Also, a wider range of utterances including spontaneous speech will have to be tested before we reach the final conclusion.

REFERENCES

[1] H. Fujisaki, K. Hirose, and N. Takahashi, "Manifestation of linguistic information in the voice fundamental frequency contours of spoken Japanese", *IEICE Trans.* ,Vol.E76-A, No.11, pp.1919-1926, 1993.

[2] N. M. Veilleux and M. Ostendorf, "Probabilistic parse scoring with prosodic information", *Proc. ICASSP'93* II-51~54, 1993.

[3] K. Hirose and H. Fujisaki, "A system for the synthesis of high-quality speech from texts on general weather conditions", *IEICE Trans.* ,Vol.E76-A, No.11, pp.1971-1980, 1993.

[4] A. Komatsu, E. Ohira, and A. Ichikawa, "Conversational speech understanding based on sentence structure inference using prosodics, and word spotting", *IEICE Trans.* Vol.J71-D, No.7, pp.1218-1228, 1988.

[5] Y. Sekiguchi, Y. Suzuki, T. Kikukawa, Y. Takahashi, and M. Shigenaga, "Existential judgement of modifying relation between successively spoken phrases by using prosodic information", *IEICE Trans.* Vol.J78-D-II, No.11, pp.1581-1588, 1995.

[6] N. Eguchi and K. Ozeki, "Dependency analysis of Japanese sentences using prosodic information", *The Journal of the Acoustical Society of Japan*, Vol.52, No.12, pp.973-978, 1996.

[7] K. Ozeki, "Dependency structure analysis as combinatorial optimization", *Information Sciences*, Vol.78, pp.77-99, 1994.

[8] S. Kurohashi and M. Nagao, "A syntactic analysis method of long Japanese sentences base on coordinate structures' detection", *Journal of Natural Language Processing*, Vol.1, No.1, pp.35-57, 1994.

[9] Y. Zhang and K. Ozeki, "Dependency analysis of Japanese sentences using the statistical property of dependency distance between phrases", *Journal of Natural Language Processing*, Vol.4, No.2, pp.3-19, 1997.