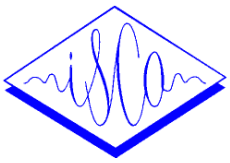

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

EFFECTIVENESS OF PROSODIC FEATURES IN SYNTACTIC ANALYSIS OF READ JAPANESE SENTENCES

*Yukiyoshi HIROSE** *Kazuhiko OZEKI* *Kazuyuki TAKAGI*

The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan
http://www-oz.cs.uec.ac.jp/

ABSTRACT

Prosody contains information that is lost when utterances are transcribed into letters or characters. This paper is concerned with exploiting such information for syntactic analysis of read Japanese sentences. In our previous work, we employed 12 prosodic features, and made a statistical model to represent the relationship between those features and dependency distances. Then, by incorporating the model in our parser, which allows the use of numerical information as linguistic knowledge, we showed that prosodic information is in fact effective for syntactic analysis. In the present work, we took up 24 prosodic features, and conducted an extensive search for effective ones. Also the statistical model was modified to account for the actual distribution of the feature values. In open experiments using an ATR 503-sentence database, parsing accuracy was improved by 21.2% compared with the case where no prosodic information was used. The duration of pauses at phrase boundaries was consistently effective in both closed and open experiments, while the effectiveness of other features, when used together with the duration of pause, was not clear in open experiments.

1. INTRODUCTION

Prosody contains information that is lost when utterances are transcribed into letters or characters. Such information may be useful for language processing, which has traditionally relied solely on information extracted from written materials. Many authors have suggested possible use of prosodic information for syntactic analysis. Uyeno *et al.* observed systematic changes in the duration of pauses at phrase boundaries and pitch contours of read Japanese sentences depending on their syntactic structures [1]. Komatsu *et al.* obtained something like a parse tree by successively dividing a sentence at phrase boundaries using prosodically defined association strength between adjacent phrases [2]. Veilleux *et al.* reported the use of prosodic information for parse scoring [3]. Sekiguchi *et al.* showed that prosodic information is effective for determining if adjacent phrases are in modification relation [4]. Eguchi *et al.* employed five prosodic features, and constructed a model to represent a statistical relationship between those features and dependency distances between phrases. Then, by incorporating the model into a Japanese parser, which can utilize numerical information as linguistic knowledge, they found that those prosodic features, especially the duration of pause, are effective to improve the parsing accuracy [5]. Up to the authors' knowledge, this is the first work for the Japanese language that demonstrated quantitatively the effectiveness of prosodic information in parsing. This line of work has been further developed, increasing the number of test speakers, and adding new prosodic features [6,7].

In the series of works [5,6,7], statistical distributions of prosodic features are modeled by Gaussian density functions. However, actual distributions of feature values, especially of the duration of pause, are significantly different from Gaussian distributions. Thus, the first point of this paper is an improvement on distribution functions for the prosodic features. The second point is an extensive search for effective prosodic features, employing 24 candidate features including new ones.

2. PROSODIC FEATURES

A Japanese sentence is a sequence of phrases, where a phrase is a syntactic unit called the *bunsetsu* in Japanese, consisting of a content word followed by (possibly 0) function words such as particles and auxiliary verbs. Given an utterance, prosodic features associated with a phrase X are defined on the basis of the duration of pause, log-power contour, log-pitch contour, and speaking rate. Many of the features are defined relative to the immediately succeeding phrase Y of X.

Most of the features employed in [7] were also taken up in this work. Those are *pause*, *front-duration*, *back-duration*, *pitch-slope*, *pitch-gap1*, *pitch-gap2*, *mean-pitch-gap*, *power-slope*, *power-gap1*, *power-gap2*, and *mean-power-gap*. In addition to these features, the following new features were examined:

(1) *Front-pause* is the duration of pause immediately preceding X.

*Currently with Sony Corporation.

(2) *Front-pitch-slope* is the regression coefficient of the pitch contour of X before its maximum.

(3) *Back-pitch-slope* is the regression coefficient of the pitch contour of X after its maximum.

(4) *Front-power-slope* is the regression coefficient of the power contour of X before its maximum.

(5) *Back-power-slope* is the regression coefficient of the power contour of X after its maximum.

(6) *Speaking-rate1* is the speaking rate of X measured in morae/sec.

(7) *Speaking-rate2* is the speaking rate of Y measured in morae/sec.

(8) *Front-phrase-command* is the magnitude of the phrase command just before X.

(9) *Back-phrase-command* is the magnitude of the phrase command just after X.

(10) *Phrase-command-time* is the time interval between the end of X and the phrase command just after X.

(11) *Phrase-command-gap1* is the difference between the magnitudes of the phrase commands just before and after X.

(12) *Phrase-command-gap2* is the phrase-command-gap1 divided by the duration of X.

(13) *No-accent-command* is the number of accent commands in X divided by the duration of X.

Thus, 24 prosodic features in total were examined in this work. The features (8)~(13) are related to the accent command and the phrase command [8], which were extracted by a prosody analysis tool *"PROSODY"* developed at Hirose Laboratory, Tokyo University.

3. DEPENDENCY STRUCTURE ANALYSIS

3.1. Global Syntactic Constraints

From a dependency grammatical point of view, the structure of a Japanese sentence can be described by specifying which phrase modifies which phrase in the sentence. Thus, the syntactic structure of a sentence $w_1 w_2 \cdots w_m$, represented as a sequence of phrases, is described by specifying a function that maps a modifier phrase to its modificand:

$$S: \ \{1, 2, \cdots, m-1\} \longrightarrow \{2, 3, \cdots, m\}.$$

Reflecting syntactic properties of the Japanese language, the function S must satisfy the following constraints:

- $\forall i \in \{1, 2, \cdots, m-1\}: \ i < S(i)$.
- $\forall i, j \in \{1, 2, \cdots, m-1\}: \ i < j \ \Rightarrow \ S(i) \geq S(j)$.

A function that satisfies the above two constraints is referred to as a *dependency structure* on $w_1 w_2 \cdots w_m$. There are ${}_{2(m-1)}C_{(m-1)})/m$ dependency structures on a phrase sequence of length m. Under a dependency structure S, $S(i) - i$ is called the *dependency distance* between w_i and $w_{S(i)}$, or simply dependency distance of w_i. For a pair of phrases w_i and w_j $(j > i)$, $j - i$ is called the *inter-phrase distance*, or simply *distance* between w_i and w_j regardless of whether $S(i) = j$ or not.

3.2. Local Syntactic Constraints

Besides the global syntactic constraints, there are local syntactic constraints concerning whether two given phrases can be in modification relation. These constraints are determined by the morphemes composing those phrases. For example, a phrase consisting of a single adjective can modify a phrase starting with a noun only, or a phrase starting with a verb or adjective only, depending on its inflected form. The set of the local syntactic constraints is referred to as the *dependency rule*.

3.3 Minimum Penalty Parser

Most of classical Japanese parsers are based on the idea of searching for dependency structures that are permitted by the dependency rule. However, this leaves too much syntactic ambiguity. Thus, instead of relying on just yes/no information given by the dependency rule, the use of numerical information such as *probability* or *preference* of modification is becoming more popular. In our parser, linguistic knowledge is represented by a function $F(x, y)$ that measures the amount of penalty when x modifies y. The parser then searches for a dependency structure S that minimizes the total penalty $\sum_{i=1}^{m-1} F(w_i, w_{S(i)})$, given a sentence $w_1 w_2 \cdots w_m$ [9].

3.4 Penalty Function

In this work, the penalty function $F(x, y)$ is defined on the basis of statistical knowledge about the relationship between the prosodic features and the dependency distances.

Let d be the dependency distance of a phrase in a sentence, and $\overline{p} = (p_1, \ldots, p_n)$ the prosodic feature vector associated with the phrase. The conditional probability of d given $\overline{p}$ is denoted by $P(d \mid \overline{p})$, which can be rewritten by Bayes theorem as

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

Thus, $P(d \mid \overline{p})$ can be calculated from $P(\overline{p} \mid d)$ and $P(d)$. $P(\overline{p} \mid d)$ is estimated by

$$P(\overline{p} \mid d) = \prod_{i=1}^{n} P_i(p_i|d),$$

where $P_i(p_i|d)$ is the conditional p.d.f. of p_i estimated from a set of phrases having the dependency distance d in a corpus. Also, $P(d)$ is estimated as $P(d) = N_d / \sum_d N_d$, where N_d is the number of phrases having the dependency distance d. Then the penalty function $F(x, y)$ is defined as

$$F(x, y) = \begin{cases} -\log P(d(x, y) \mid \overline{p}), & \text{if } (x, \ y) \in DR \\ \infty, & \text{otherwise,} \end{cases}$$

where $d(x, y)$ is the distance between the phrases x and y, $\overline{p}$ is the prosodic feature vector associated with x, and

$(x,\ y) \in DR$ signifies that x is allowed to modify y by the dependency rule.

4. EFFECTIVENESS OF PROSODIC FEATURES

4.1 Speech Material

An ATR speech database [10] was used in this work. This database contains 503 Japanese sentences extracted from newspapers, journals, novels, letters, textbooks, and etc., which are divided into 10 groups A $\sim$ J. The sentences have labels that indicate their dependency structures. It also contains the speech waveforms for the sentences read by professional announcers/narrators. Two male speakers' (MHT, MTK), and two female speakers' (FKN, FYM) voices were used here.

In the following experiments, the sentence groups A $\sim$ J were divided into training data and test data as in Table 1. Exp(i) is for closed experiments, while Exp(ii) and Exp(iii) are for open experiments. All the experiments in this paper are speaker-closed. Results of parsing were evaluated by *parsing accuracy*: the percentage of test sentences whose dependency structures determined by parsing coincide exactly with those described in the database.

Table 1. Training data and test data.

	training data	test data
Exp(i)	A–J (503 sentences)	A–J (503 sentences)
Exp(ii)	D–J (353 sentences)	A–C (150 sentences)
Exp(iii)	A–G (350 sentences)	H–J (153 sentences)

4.2 Distribution Function for Duration of Pause

In the series of works [5,6,7], the distribution of the duration of pauses is assumed to be Gaussian for every dependency distance. However, the actual distribution differs significantly from a Gaussian distribution especially for dependency distances 1$\sim$ 3; the histogram of pauses has a sharp peak at duration = 0, then a deep *dip* appears at a small value of duration. In order to approximate this peculiar distribution, every combination of Gaussian distribution, Poisson distribution, exponential distribution, and normalized histogram was tested for dependency distances 1$\sim$3. For dependency distances greater than 3, normalized histograms were used. Table 2 shows the best three combinations under the condition Exp(i). Table 3 shows the parsing accuracy for those combinations. In the table, *dist.* means a case when $P(d(x,y) \mid \overline{p})$ was replaced with $P(d(x,y))$ in the definition of the penalty function $F(x,y)$, and *det.* means a case when a deterministic analysis [11] was employed, in which no prosodic information was used. Although C_3 gave the best average accuracy, the best combination differs from speaker to speaker. To resolve this point, one more prosodic feature was combined with each combination, measuring the parsing accuracy. As a result, C_2 turned out to be the best for almost all the speakers. Therefore, throughout the experiments below, the combination C_2 was used.

Table 2. Combinations of distribution functions for the duration of pause. For dependency distances greater than 3, normalized histograms were used. D stands for dependency distance.

C_1	Gaussian distributions for $D = 1, 2, 3$.
C_2	Normalized histogram for $D = 2$. Gaussian distributions for $D = 1, 3$.
C_3	Gaussian distributions for $D = 1, 2$. Poisson distribution for $D = 3$.

Table 3. Parsing accuracy(%) for combinations of distribution functions.

Comb.	MHT	MTK	FKN	FYM	Av.
C_1	59.6	57.3	57.3	55.5	57.4
C_2	59.4	57.3	57.1	55.7	57.4
C_3	60.2	56.9	57.3	55.7	57.5
dist.					52.3
det.					47.3

4.3 Combinations of Prosodic Features

It has been reported that the duration of pause is most effective [5]. In order to search for other effective features, the next most effective feature was determined one by one, and added successively to the components of the prosodic feature vector.

4.3.1 Closed Experiments

The experimental condition was Exp(i) in this case. As shown in Table 4, 11 features were found effective. The parsing accuracy for combinations of those features are shown in Table 5. Thus, parsing accuracy was improved by 28.3% compared with that of deterministic analysis where no prosodic information was used.

Table 4. Combinations of prosodic features.

Comb.	Features successively added
C_a	"pause" only
C_b	C_a + "speaking-rate2"
C_c	C_b + "power-gap1"
C_d	C_c + "front-pitch-slope"
C_e	C_d + "front&back-power-slope"
C_f	C_e + "no-accent-command"
C_g	C_f + "back-duration"
C_h	C_g + "phrase-command-gap2"
C_i	C_h + "power-gap2"
C_j	C_i + "pitch-gap2"
C_k	C_j + "front-pause"

Table 5. Parsing accuracy(%) for combinations of prosodic features.

Comb.	MHT	MTK	FKN	FYM	Av.
C_a	59.4	57.3	57.1	55.7	57.4
C_b	60.4	58.4	58.1	57.1	58.5
C_c	60.4	59.0	58.1	57.3	58.7
C_d	61.0	59.2	57.9	57.3	58.9
C_e	61.6	58.4	58.3	58.3	59.2
C_f	62.6	59.8	58.3	58.3	59.8
C_g	63.0	59.2	59.2	58.1	59.9
C_h	63.0	60.0	59.6	59.2	60.5
C_i	63.4	60.8	58.6	59.4	60.5
C_j	63.2	60.8	58.6	59.6	60.6
C_k	63.8	60.4	59.2	59.4	60.7
dist.					52.3
det.					47.3

4.3.2 Open Experiments

In this case, the experimental conditions were Exp(ii) and Exp(iii), and the results were averaged over the two cases. As shown in Table 6, 3 features were found effective. It is seen in Table 7 that parsing accuracy was improved by 21.2% compared with that of deterministic analysis. However, only a very small number of features showed effectiveness. Moreover, most of the improvement was due to the duration of pause, and contribution of other features was not clear.

Table 6. Combinations of prosodic features.

Comb.	Features successively added
C_A	"pause"only
C_B	C_A + "power-gap1"
C_C	C_B + "back-duration"

Table 7. Parsing accuracy for combinations of prosodic features (averaged over Exp(ii) and Exp(iii)).

Comb.	MHT	MTK	FKN	FYM	Av.
C_A	60.4	59.8	59.7	56.8	59.2
C_B	62.3	60.0	59.8	57.4	59.9
C_C	62.3	61.3	59.1	57.4	60.0
dist.					54.5
det.					49.5

5. CONCLUSION

Twenty-four prosodic features were employed, and an extensive search was conducted for effective ones. Also, an effective combination of distribution functions was sought to better model the actual distribution of the duration of pauses. Parsing accuracy was improved by 21.2% with the use of prosodic information in open experiments. This figure is 4.7 point higher than our previous result [7]. The duration of pause was consistency effective in both closed and open experiments, while contribution of other features, related to the pitch, power, and speaking rate, was not clear when used together with the duration of pause. More work is needed to assess the amount of syntactic information contained in those features, and to find a way of fully exploiting them in syntactic analysis.

Acknowledgment

The authors gratefully acknowledge Prof. K.Hirose for providing a prosody analysis tool *PROSODY*.

REFERENCES

[1] T. Uyeno, H. Hayashibe, K. Imai, H. Imagawa, and S. Kiritani: "Syntactic structure and prosody in Japanese: a study on pitch contours and the pauses at phrase boundaries," Annual Bulletin of Research Institute of Logopedics and Phoniatrics, University of Tokyo, Vol.15, pp.91-108, 1981.

[2] 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.

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

[4] 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.

[5] 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.

[6] K. Ozeki, K. Kousaka, and Y. Zhang: "Syntactic information contained in prosodic features of Japanese utterances," *Proc. Eurospeech'97*, pp.1471- 1474, 1997.

[7] K. Ozeki, K. Kousaka, and Y. Zhang: "The use of prosodic information in syntactic analysis of Japanese utterances," *Proc. SPECOM'98*, pp.157- 160, 1998.

[8] H. Fujisaki and K. Hirose: "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, Vol.5, No.4, pp.233-242, 1984.

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

[10] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara: "Manual of Japanese Speech Database", ATR, 1990.

[11] 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.