Dependency Analysis of Spontaneous Monologue Speech Using Pause and F_0 Information: A Preliminary Study

Kazuyuki Takagi & Kazuhiko Ozeki

The University of Electro-Communications,Tokyo, Japan

{takagi,ozeki}@ice.uec.ac.jp

Abstract

This paper deals with the problem of exploiting prosodic information in syntactic analysis of spontaneous monologue utterances of non-professional speakers. Duration of pauses at phrase boundaries and relative F_0 contour features, which improve parsing accuracy of read sentences, were also found to be effective for parsing spontaneous speech. Dependency analysis was performed by the minimum penalty parser on academic presentation speech recorded in Corpus of Spontaneous Japanese, a large-scale database of spontaneous Japanese with rich linguistic annotations. Preliminary experiments on relatively clean parts of the monologue data utterances showed that the pause and F_0 features are effective to improve the accuracy of dependency analysis of spontaneous utterances, and that combined use of both features will give further improvement. It was also found that the effectiveness of pause information was larger when pause models were estimated separately for zero-duration and non-zero-duration pauses, which better model the actual distribution of pause duration than a simple Gaussian distribution. Although this is a preliminary study, the results are promising.

1. Introduction

Researches related to the problem of recovering syntactic structure based on prosodic information are found in the literature [1, 2, 3, 4]. The problem of exploiting prosodic information in ASU is found in [5], in which prosodic information is used to recognize phrase boundaries and to select alternative string hypotheses. However, very little work has been done in NLP field to incorporate prosodic information directly into a parser as linguistic knowledge, and exploit it in the search process. Eguchi and Ozeki presented a method of incorporating prosodic information into a Japanese dependency structure parser [6]. The parser can handle both symbolic information such as syntactic rule and numerical information such as the probability of dependency distance of a phrase in a unified way as linguistic information. As results of our previous work in which an optimal combination of various prosodic features was sought for [7], the duration of pauses at phrase boundaries and relative F_0 contour features were found to be most effective.

Our recent work [8] supports that pause and F_0 information can be applied effectively to dependency analysis in a speaker-independent condition by normalizing the prosodic features and by employing simpler (i.e., with more generalization ability) models for the distributions of the features. It has also been shown that linear combination of pause and F_0 information [9] improves parsing accuracy more than the cases where pause or F_0 information is used by itself in a speaker-independent condition as well as in a speaker-dependent condition.

A natural application of our framework will be analysis of spontaneous utterances. In this paper, the dependency analysis method is applied to spontaneous monologue utterances recorded in Corpus of Spontaneous Japanese (CSJ) [10], which is a large-scale database of spontaneous Japanese with rich linguistic annotations. As a preliminary study, relatively clean parts of the monologue data of CSJ is analyzed to examine the effectiveness of the pause and F_0 features that has been shown to be effective in dependency analysis of read sentences by professional speakers.

2. Parser

2.1. Dependency distance

A Japanese sentence is a sequence of syntactic units called *bunsetsu* (hereafter simply referred to as "phrase"), where a *bunsetsu* consists of content words followed by function words. Let $w_1 w_2 \dots w_m$ be a sentence represented as a sequence of phrases. If w_i modifies w_j, then $j - i$ is referred to as the *dependency distance* of w_i.

From a dependency grammatical point of view, the structure of a Japanese sentence can be determined by specifying the dependency distance of each phrase in the sentence. Thus any information related to the dependency distance is expected to be useful for dependency analysis.

2.2. Dependency structure

The dependency structure of a sentence $w_1 w_2 \dots w_m$ is determined by specifying a function S that maps a modifier phrase to the modified phrase:

$$S : \{1, 2, \dots, m-1\} \rightarrow \{2, 3, \dots, m\}.$$

The function S must satisfy the following constraints in order to reflect syntactic properties of Japanese language:

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

A function that satisfies these constraints is referred to as a *dependency structure* on $w_1 w_2 \dots w_m$. Note that our parser can be easily applied to other languages provided that their syntax satisfies these constraints.

2.3. Minimum penalty parser

In our parser, linguistic knowledge is represented by a function $F(w_i, w_j)$ that measures the amount of penalty when a phrase w_i is to modify a phrase w_j. The parser searches for a depen-

Table 1: Frequency of dependency distances in CSJ and ATR

Dependency Distance	CSJ *Core* APS		ATR PB Sentences	
	Frequency	%	Frequency	%
1	38570	59.5	1909	65.3
2	9154	14.1	500	17.1
3	4724	7.3	253	8.7
4	2822	4.4	126	4.3
5	1940	3.0	73	2.5
> 6	7585	11.7	61	2.1
Total	64795	100.0	2922	100.0

dency structure S that minimizes the total penalty

$$\sum_{i=1}^{m-1} F(w_i, w_{S(i)})$$

given a sentence $w_1 w_2 \dots w_m$ [6].

2.4. Penalty function

The penalty function $F(w_i, w_j)$ is defined on the basis of conditional probability of the dependency distance given the prosodic feature [6]:

$$F(w_i, w_j) = \begin{cases} -\log P(d \mid \boldsymbol{p}), & \text{if } (w_i, w_j) \in DR \\ \infty, & \text{otherwise} \end{cases} \quad (1)$$

where $d = j - i$ is the dependency distance of w_i, $\boldsymbol{p}$ is the prosodic feature vector associated with w_i, and "$(w_i, w_j) \in DR$" signifies that w_i is allowed to modify w_j by the local syntactic constraints, or *dependency rule*, *DR*, which is based on the morphological structure of the phrases.

3. Database

Corpus of Spontaneous Japanese (CSJ) [10] is a large-scale database of spontaneous Japanese, which contains about 660 hours or 7.5 million words of speech recordings of academic presentation speech (APS), simulated public speech, and some other miscellaneous speech sources. Among them is *Core* dataset, to which transcriptions, phonetic labels, dependency structure information as well as detailed linguistic annotations (e.g., POS, filled pauses, word fragment, mispronunciation) are provided. APS is a collection of live recordings of 9 different academic societies whose fields cover engineering, social science, and humanities.

There is no clear boundary marks of sentences in the spontaneous speech data. So, *clause unit* that is defined in CSJ is treated as sentence in this paper. Table 1 shows the statistics of dependency distance in 70 APS monologues of *Core* dataset, and that of sentence reading speech of ATR DB [12]. Although both distributions are almost the same for the distances less than 6, CSJ APS data contains longer dependency distances. This is because APS is orally presented written languages, which often contain filled pauses and inserted phrases that make clause units lengthy.

4. Prosodic models for dependency analysis

Given an utterance, many of the prosodic features of the phrase w_i ($1 \leq i \leq m-1$) are defined relative to the immediately succeeding phrase w_{i+1}.

4.1. Pause duration

The pause duration x of a phrase w_i in question is defined as the interval between the ending time of w_i and the starting time of w_{i+1}. The mean pause duration grows linearly with the dependency distance up to $d = 4$, though the slope depends on the speaker as illustrated in Fig. 2-(a), for 10 speakers in ATR SET B [12]. This fact shows that the duration of pause contains information about dependency distance [6]. The average mora duration was calculated sentence-by-sentence, and then duration of pauses in each sentence was normalized by the average mora duration of the sentence [8].

The same relationship between dependency distance and pause duration holds for monologue speech in CSJ corpus in this distance range as illustrated in Fig. 2-(b), though slopes vary more than ATR case.

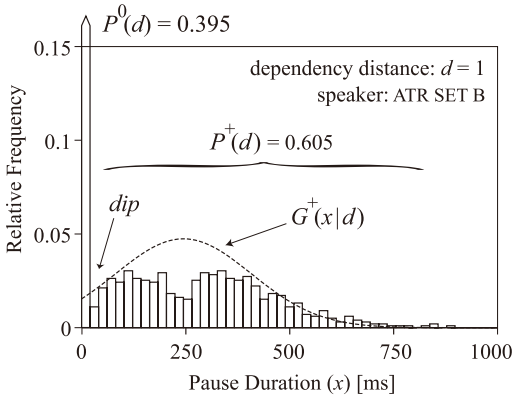

Figure 1: Histogram of pause duration x for $d = 1$ calculated over all speakers of ATR DB SET B, and probability models that fit the distribution. $P^0(d)$, $P^+(d)$: discrete probability of pause for $x = 0$, and $x > 0$; $G^+(x|d)$: single Gaussian p.d.f. for $x > 0$.

4.1.1. Simplification of $P(d \mid x)$

When the pause duration for each dependency distance is modeled by Gaussian p.d.f., there are large overlaps among the distributions for distances greater than 1. Also, the occurrence frequency of the dependency distance greater than 4 is so low ($<$ 3%) that an estimation of Gaussian parameters is not reliable [8]. Therefore, as in our previous work, the number of distance classes was limited in estimation of $P(d \mid x)$, that is, pause data of a phrase whose dependency distance is greater than the predefined upper limit are treated as single class. Table 2 shows the dependency accuracy, which is averaged over all APS monologues tested in the experiment. *Dependency accuracy* is the percentage of test phrases whose dependency distance were estimated correctly.

By limiting the number of distance classes to 3, best dependency accuracy of 65.8% was obtained, as in "no separation" case in Table 2. Although the improvement was very small (compared to 65.3% in the case distance limit is ∞), the number of parameters to be estimated was reduced to one-third of the case where there was no limit on the number of distance classes. This result is consistent with our previous experiment in which sentences read by professional announcers were analyzed [8].

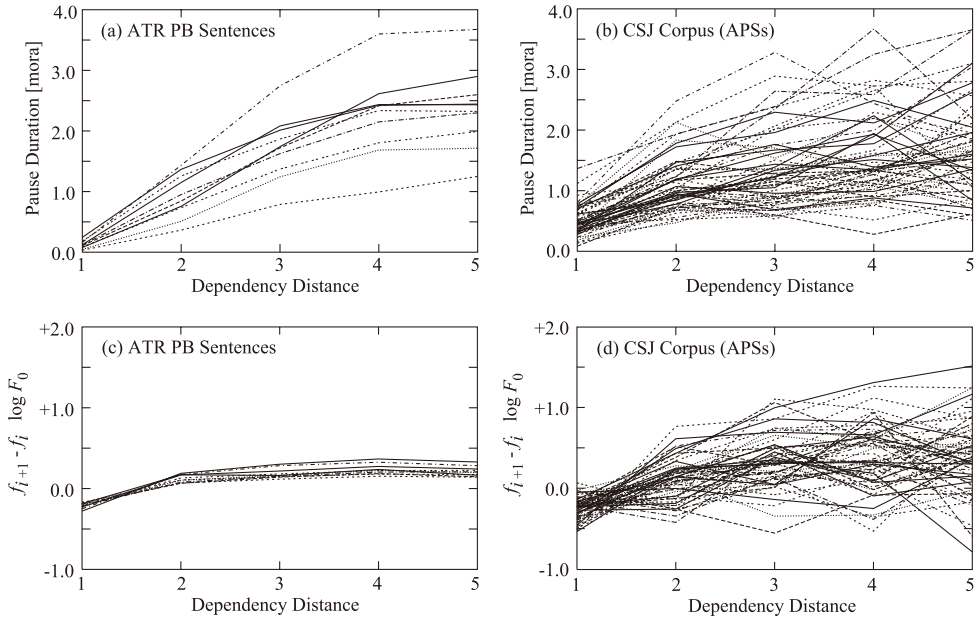

Figure 2: Pause duration and F_0 feature value as a function of dependency distance in ATR PB 503 sentences and in CSJ APS monologue. Different lines correspond to different speakers of ATR in (a) and (c), or different monologues in (b) and (d).

Table 2: Dependency accuracy (%) distance limit for pause p.d.f. estimation.

Distance limit	$d > 1$	$d > 2$	$d > 3$	$d > 4$	∞
no separation	65.6	**65.8**	65.3	65.2	65.3
separation	66.3	66.3	**66.5**	66.2	65.4

4.1.2. Separation of zero-duration-pause

Fig. 1 shows the histogram of pause duration for dependency distance $d = 1$ calculated over 10 speakers in ATR SET B. It has a sharp peak at duration=0, then a deep *dip* appears at a small value of duration. The pause duration modeling method that was effective in our work on ATR data [8, 11] was tested for CSJ data. In this model, $P(x \mid d)$ is defined as

$$P(x \mid d) = \begin{cases} P^0(d), & \text{if } x = 0 \\ P^+(d)\, G^+(x \mid d), & \text{if } x > 0 \end{cases} \qquad (2)$$

where $G^+(x \mid d)$ is a Gaussian p.d.f. estimated from the pause data of dependency distance d whose duration is greater than 0. Also, $P^0(d) = N_d^0/N_d$, $P^+(d) = N_d^+/N_d$, where N_d^0 and N_d^+ are the number of phrases of dependency distance d whose pause duration is equal to 0 and greater than 0, respectively. Limitation of distance classes was also effective for this model, yielding the best dependency accuracy of 66.5% when the number of distance class is 4, as shown in "separation" case in Table 2.

4.2. F_0 contour feature

The log-F_0 contour of the phrase w_i was first smoothed by fitting a parabola. Then, the F_0 at the time-center of the curve, f_i, was picked up as the F_0 feature of the phrase w_i. Figure 3 is a typical example of the distribution of $f = f_{i+1} - f_i$. There is a significant difference between the distribution for $d = 1$ and those for $d > 1$. In other words, $f = f_{i+1} - f_i$ has information about whether the phrase modifies the immediately succeeding phrase or not.

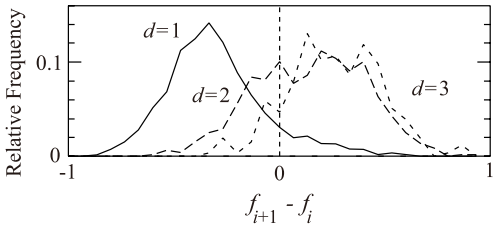

Figure 3: Distribution of $f_{i+1} - f_i$ for $d = 1$, $d = 2$, and $d = 3$.

Lower half of Fig. 2 illustrates the value of $f = f_{i+1} - f_i$ as a function of dependency distance. The property of this feature is clear in the ATR case (Fig. 2-(c)): the value changes from negative to positive as d changes from 1 to 2, and remains unchanged as d becomes larger. For 38 out of the 49 monologues that are plotted in Fig. 2-(d), value of f is negative for $d = 1$ and positive for $d > 1$ as in ATR data, while for the rest the similar relation is not observed.

4.3. Linear combination of pause and F_0 information

In order to combine pause and F_0 information to make a single penalty function, a weighted sum of the log conditional probability of dependency distance given the pause duration and that given F_0 information [9] is calculated as

$$F(w_i, w_j) = -\{\alpha \log P(d \mid x) + (1 - \alpha) \log P(d \mid f)\}, \qquad (3)$$

where $d = j - i$. The optimum value of α was determined experimentally.

5. Experiments

5.1. Experimental condition

A small part of the data was selected from the *Core* APS of CSJ in such a way that it does not contain restarts, nor filled pauses that do not have a phrase to modify. Also, clause units that consist of more than 31 phrases were excluded. By this filtering, 373 sentences (clause units) or 3211 phrases chosen from 44 monologues were extracted for the analysis dataset consisting of 3211 dependency relations. This filtering was done because we wanted to check, as a preliminary study in this paper, whether our method that is effective for read speech is promising or not.

Sentences in this dataset was then split into 3 subsets of equal size so that the distributions of sentence length in terms of the number of phrases are balanced. One subset consists of about 145 sentences or about 1300 phrases, chosen from 36 or 38 monologues. One of the 3 subsets was left for evaluation and the other 2 datasets were used for estimating pause and F_0 feature value distributions. Dependency accuracy was obtained by averaging over three different evaluation subsets. The results are not speaker independent because there are overlaps in monologues among the evaluation datasets, although no clause unit is included in multiple evaluation datasets.

Pause duration and F_0 features were extracted from the label files of CSJ. As for dependency rule *DR*, the same rule that had been used in read sentence cases in our previous work was used, by mapping POS system of CSJ to that of ATR DB.

5.2. Results and discussion

Table 3 shows the dependency accuracy obtained in various conditions. The baseline dependency accuracy was 60.4% by the deterministic analysis [13] in which no prosodic information was used.

With the use of pause information, the dependency accuracy was improved, as in our previous works on read sentences by professional announcers [8, 9, 11]. The improvement was 5.4 points over the baseline, or 6.1 points when the pause distribution was modeled separately for zero-duration and non-zero-duration pause (Pause_Separated). F_0 information was also effective, with the same improvement rates (6.1 points) as in the case of Pause_Separated.

The last two rows show the results when pause and F_0 information were linearly combined. The accuracy values are the average of the best dependency accuracies obtained at the optimum values of α. Dependency accuracy was improved slightly in both cases, Pause+F_0 and Pause_Separated+F_0, giving 6.6 point improvement over the baseline. In Pause_Separated+F_0 case, the optimum values of α shifted close to 1. This indicates that the pause information was better represented by Pause_Separated model.

Table 3: Dependency accuracy in various conditions of prosodic information usage.

Condition	Dependency Accuracy (%)
Deterministic (no prosody)	60.4
Pause	65.8
Pause_Separated	66.5
F_0	66.5
Pause +F_0	67.0
Pause_Separated +F_0	67.0

6. Conclusion

In this paper, we conducted a dependency analysis of spontaneous monologue speech by the minimum penalty parser utilizing pause and F_0 information. Duration of pauses at phrase boundaries and relative F_0 contour features, which improve parsing accuracy of read sentences, were found to be also effective for parsing spontaneous speech. It was shown that better performance was obtained by employing simpler models for pause duration and F_0 feature distributions, and that linear combination of them yielded further improvement over the baseline. Although the experiments are preliminary, the results are promising.

The baseline dependency accuracy was not enough: it is 86.3% in ATR DB case. So, as a future work a dependency rule tuned to spontaneous speech analysis should be developed. Making a framework to process events specific to spontaneous speech, such as filled-pauses and restarts, is also a part of our future work.

7. References

[1] 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, 1218–1228, 1988.

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

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

[4] J. Venditti, S. Jun and M. Beckman, "Prosodic cues to syntactic and other linguistic structures in Japanese, Korean and English,", J. L. Morgan and K. Demuth (eds.), Signal to syntax: bootstrapping from speech to grammar in early acquisition (Hillsdale, NJ: Lawrence Erlbaum), 287–311, 1996.

[5] R. Kompe, "Prosody in speech understanding systems," Lecture Notes in Artificial Intelligence 1307, Springer, 1997.

[6] N. Eguchi and K. Ozeki, "Dependency analysis of Japanese sentences using prosodic information," J. Acoust. Soc. of Japan, Vol. 52, No. 12, 973–978, 1996.

[7] Y. Hirose, K. Ozeki, and K. Takagi, 2001, "Effectiveness of prosodic features in dependency analysis of read Japanese sentences," Natual Language Processing, Vol. 8, No. 4, 71–89.

[8] K. Takagi, and K. Ozeki, "Dependency analysis of read Japanese sentences using pause and F_0 information: a speaker independent case," Proc. ICSLP 2004, Vol .4, 3021–3024, 2004.

[9] K. Takagi, H. Kubota, and K. Ozeki, "Combination of pause and F0 information in dependency analysis of Japanese sentences," Proc. ICSLP2002, 1173–1176, 2002.

[10] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), paper MMO2, 2003.

[11] K. Takagi, and K. Ozeki, "Pause information for dependency analysis of Japanese sentences," Proc. Eurospeech2001, 1041–1044, 2001.

[12] Y. Sagisaka, K. Tanaka, M. Abe, S. Katagiri, T. Umeda, and H. Kuwabara, "A large-scale Japanese speech database," Proc. ICSLP1990, Vol. 2, 1089–1092, 1990.

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