

A Neural Network Approach to Dependency Analysis of Japanese Sentences Using Prosodic Information

Kazuyuki Takagi Mamiko Okimoto Yoshio Ogawa Kazuhiko Ozeki

The University of Electro-Communications, Tokyo 182-8585, Japan
{takagi,ozeki}@ice.uec.ac.jp

Abstract

Prosody and syntax are significantly related with each other as has often been observed. In the field of speech synthesis, many efforts have been made to control prosody so that it reflects the syntactic structure of the sentence. However, the inverse problem, recovery of syntactic structure using prosodic information, has not been so much investigated. This paper focuses on syntactic information contained in prosodic features extracted from read Japanese sentences, and describes a method of exploiting it in dependency structure analysis. In this paper, a multilayer perceptron is employed to estimate conditional probability of dependency distance of a phrase given its prosodic feature, i.e., pause duration and F_0 contour. Parsing accuracy was improved by combining two different kinds of prosodic information by the perceptron.

1. Introduction

There is a significant relationship between prosody and syntax. In the field of speech synthesis, numerous papers have been published on prosody control based on the syntactic structure of a sentence [1, 2]. This paper is concerned with the inverse problem: recovery of syntactic structure based on prosodic information. Some work related to this problem can be found in the literature [3, 4, 5, 6]. However very little work has been done to incorporate prosodic information directly into a parser as linguistic knowledge, and exploit it in the search process.

Eguchi and Ozeki [7] presented a method of incorporating prosodic information into a Japanese dependency structure parser in 1996. The parser can handle both symbolic information such as syntactic rule and numerical information such as probability of dependency distance in a unified way as linguistic information. This work has been further extended by increasing the number of prosodic features and the number of speakers [8, 9]. An optimal combination of these features was also sought for [9].

Combination of pause and F_0 information is an important issue because it is indicated that pause duration and F_0 contour features seem to work complementarily in analyzing dependencies of short distance. We have tested various probability density

functions, in order to approximate the distributions of prosodic features [9, 10, 11]. The inventory of probability density functions that can be used to model joint probability distributions is limited. We also want a better way to integrate pause and F_0 information. In this paper, a multilayer perceptron was used to estimate conditional probabilities of dependency distance given the prosodic features. The multilayer perceptron may be useful to model complicated probability distributions, and to integrate unknown probability distributions of prosodic features from multiple sources.

2. Dependency distance and prosodic features

A Japanese sentence is a sequence of phrases, where a phrase is a syntactic unit called *bunsetsu* (hereafter simply referred to as “phrase”) in Japanese, consisting of a content word followed by (possibly zero) 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 except for the last phrase in the sentence. Thus any information related to the dependency distance is expected to be useful for dependency structure analysis.

3. Minimum penalty parsing

3.1. Parser

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

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

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

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

A function that satisfies these constraints is referred to as a *dependency structure* on $w_1 w_2 \dots w_m$.

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 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 $w_1 w_2 \dots w_m$ [7].

3.2. Penalty function

In this work, as in our previous works [7, 8, 9, 10, 11, 12], the penalty function $F(w_i, w_j)$ is defined on the basis of conditional probability of the dependency distance given the prosodic features:

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

where $d = j - i$, \mathbf{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*, which is based on the morphological structure of the phrases.

4. Prosodic features

4.1. Post-phrase pause duration

The post-phrase pause duration of a phrase in question is defined to be the time interval between the ending point of the phrase and the starting point of the immediately succeeding phrase. Fig. 1 illustrates the mean pause durations for four speakers, MHT, MTK, FKN, and FYM in a database [13], as functions of the dependency distance. The mean pause duration grows linearly with the dependency distance up to $d = 4$, though the slope depends on the speaker. This shows that the duration of pause contains information about dependency distance.

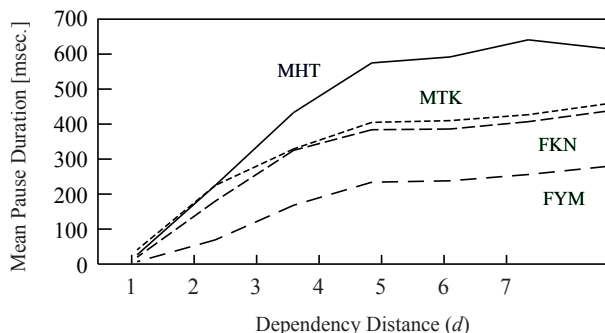


Figure 1: Mean pause duration for four speakers as functions of the dependency distance.

4.2. Fundamental frequency contour

The log- F_0 contour of a phrase in question was first smoothed by fitting a quadratic regression curve. Then, as illustrated in Fig. 2, three points were picked up from the curve: the center point, and the two points inside the end points by 10% of the phrase length. Let the log- F_0 frequencies at those points be f_1, f_2, f_3 , respectively. In the same way, three log- F_0 frequencies f_4, f_5, f_6 were measured for the immediately succeeding phrase. Then in order to get relative values, f_2 was selected to be a reference point, and all the values were represented relative to f_2 . Thus a 5-dimensional feature vector $\mathbf{f} = (f_1 - f_2, f_3 - f_2, f_4 - f_2, f_5 - f_2, f_6 - f_2)$ was obtained for each non-sentence-final phrase.

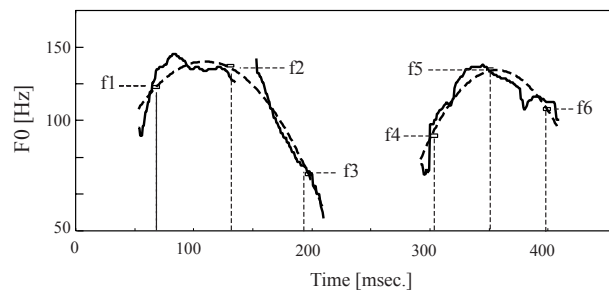


Figure 2: Extraction of a F_0 feature vector from adjacent F_0 contours. Solid lines show the original F_0 contours, and broken lines show quadratic regression curves.

Fig. 3 shows an example of the distribution of $f_5 - f_2$ for dependency distance $d = 1, 2, 3$. There is a significant difference between the distribution for $d = 1$ and those for $d > 1$, but no clear difference between $d = 2$ and $d = 3$.

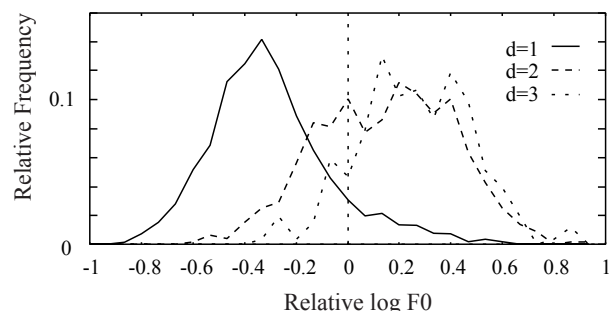


Figure 3: Distribution of $f_5 - f_2$ for $d = 1, 2, 3$. (Speaker: MHT)

4.3. Multilayer perceptron

We have tested combinations of Gaussian, Poisson, exponential distributions, normalized histogram[9], and Gaussian mixture distribution [10, 11], in order to approximate the distributions of post-phrase pause duration, because the actual distribution differs from a simple Gaussian distribution. However there were still not significant differences among the results.

Inventory of probability density functions that can be used to model joint probability distributions is limited. We also want a better way to integrate pause and F_0 information. A multilayer perceptron may be useful to integrate unknown probability distributions of prosodic features from multiple sources.

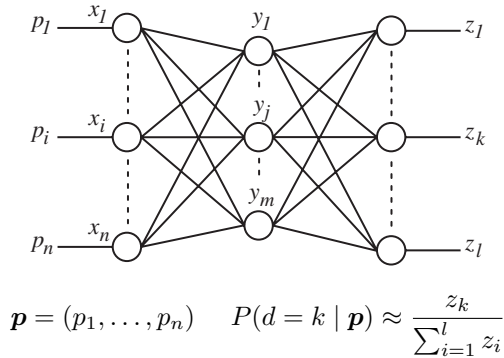


Figure 4: Three-layer perceptron for calculating $P(d = k | \mathbf{p})$.

Experiments were conducted using a three-layer perceptron in Fig. 4. The perceptron was so trained that the values of the output layer units represent the probabilities $P(d = k | \mathbf{p})$ of the phrase in question when the prosodic feature vector of the phrase is fed into the input layer. For this purpose, the training signal for the k -th output unit was set to 1 (0 for all other units) for the input feature vector of the training phrase whose dependency distance is k . After the training completes, the probability distribution of dependency distance d is estimated from the output z_i ($i = 1, \dots, l$) by

$$P(d = k | \mathbf{p}) \approx \frac{z_k}{\sum_{i=1}^l z_i}.$$

The number of input units was set identical with the dimension of the prosodic feature vector, e.g., 1 in *Pause-only* case, 5 in F_0 -only case, and 6 in *Pause+ F_0* case. The number of output units was fixed in all cases to 10, which is the maximum value of dependency distance in the database. The number of hidden units was determined experimentally considering the number of input units: 1 in *Pause-only* case, 7 in F_0 -only case, and 9 in *Pause+ F_0* case.

5. Experiments

An ATR speech database [13] was used in this work. The database contains 503 Japanese sentences extracted from newspapers, journals, novels, letters, textbooks, etc., which are divided into 10 groups A – J. The sentences have labels that indicate their dependency structures. It also contains the speech waveforms for the sentences read by professional announcers or narrators. Speech data of five male speakers (Mxx), and four female speakers (Fxx) were used, among which four speakers, MHT, MTK, FKN, and FYM, are common to the experiments of this paper and the previous work [12] (common set).

In the following experiments, the sentence groups A – J were divided into training data and test data as in Table 1. Results were averaged over Set(i) and Set(ii). All the experiments in this paper are speaker-dependent. Results of parsing were evaluated by parsing accuracy, i.e., the percentage of test sentences whose dependency structures determined by parsing are exactly the same as those described in the database.

Table 1: Training data and test data.

Dataset	Training data	Test data
Set(i)	D – J (353 snt.)	A – C (150 snt.)
Set(ii)	A – G (350 snt.)	H – J (153 snt.)

6. Results

Table 2 shows the best results obtained so far by the perceptrons trained with different convergence conditions. “Dist.” denotes a case where $P(d | \mathbf{p})$ is replaced with $P(d)$ in Eq. 1. “Det.” means a case where a deterministic analysis method [14] was employed. The parsing accuracy was improved from 49.5% in Det. to 54.5% by using the distance distribution information. The result obtained in the previous work [12] is also shown for comparison, in which $P(d | \mathbf{p})$ is modeled by a Gaussian probability density function. “Sub. Av.” is the average over the speakers in the common set, while “Total Av.” is the average over all the nine speakers.

Effectiveness of pause information was greater than that of F_0 , just as in our previous work that uses Gaussian p.d.f. for the estimation of $P(d | \mathbf{p})$ [12]. As for the speakers in the common set (MHT, MTK, FKN, FYM), better performance was obtained compared to the previous work except for FYM. In *Pause-only* case, performance was improved for all the four speakers in the common set with the improvement ranging from 1.6 points (FYM) to 3.3 points (MHT) compared to the previous work. As an average of the common set, the parsing accuracy was 2.4 points better than the previous work.

Table 2: Parsing accuracy (%) by using the perceptron, and by the previous work [11] in which Gaussian p.d.f. is used (in parentheses).

Cond.	<i>Pause-only</i>	<i>F₀-only</i>	<i>Pause+F₀</i>
MHT	62.4 (61.4)	60.7 (57.4)	60.5 (62.1)
MTK	61.4 (61.4)	57.1 (54.8)	61.0 (61.4)
FKN	61.4 (60.7)	59.7 (57.1)	61.4 (60.7)
FYM	58.1 (59.4)	57.4 (55.8)	59.4 (59.4)
Sub. Av.	60.8 (60.7)	58.7 (56.3)	60.6 (60.9)
MHO	58.7	60.1	61.0
MMY	59.4	60.1	61.0
MSH	59.4	57.8	60.7
FKS	59.4	59.4	61.0
FTK	59.4	57.7	60.4
Total Av.	60.0	58.9	60.7
Dist.	54.5		
Det.	49.5		

When pause and F_0 information were combined (*Pause+F₀* case), the average parsing accuracy was improved by 1.9 points compared to *F₀-only* case (60.6% vs. 58.7%) for the common set, although the effects differed from speaker to speaker. The performance was also better than the *Pause-only* case, except for the three speakers, MHT, MTK, and FKN. On the whole, the improvement by combining the two kinds of prosodic information by the perceptron was 0.7 points from *Pause-only* case, and 1.8 points from *F₀-only* case (See ‘Total Av.’ in Table 2).

The performance of the perceptrons depends on various parameters, e.g., the number of hidden layer units, initial weights of the connections, and convergence criteria. Better results may be obtained by adjusting these parameters optimally.

7. Conclusion

This paper focused on using the multilayer perceptron for estimating distributions of post-phrase pause duration, F_0 contour feature, and for integrating the two features, in dependency structure analysis of read Japanese sentences. By using pause information, the parsing accuracy was improved by 5.5 points compared with the case where only distance distribution information was used. Improvement of parsing accuracy by using the multi-dimensional feature of F_0 contour was 4.4 points, which was larger than Gaussian p.d.f. model case [12]. Parsing accuracy was further improved when pause and F_0 were integrated by the perceptron. It was shown that a multilayer perceptron is a promising device for estimating the probability distribution of dependency distance from the prosodic features. Our future work includes elaboration of the multilayer perceptron, as well as training and evaluation using a larger dataset. It should be noted that the upper bound of the parsing accuracy of our method is limited by the sentence

coverage rate of the dependency rule DR , which is currently 73%. Improving the coverage rate of DR will contribute to the performance of the parser.

8. References

- [1] N. Kaiki and Y. Sagisaka, “Study of pause insertion rules based on local phrase dependency structure,” IEICE Trans., Vol. J79-D-II, No. 9, pp. 1455-1463, 1996.
- [2] N. Kaiki and Y. Sagisaka, “ F_0 control based on local phrase dependency structure,” IEICE Trans., Vol. J83-D-II, No. 9, pp. 1853-1860, 2000.
- [3] 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.
- [4] N. M. Veilleux and M. Ostendorf, “Probabilistic parse scoring with prosodic information,” Proc. ICASSP’93, Vol. II, pp. 51-54, 1993.
- [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] 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), pp. 287-311, 1996.
- [7] N. Eguchi and K. Ozeki, “Dependency analysis of Japanese sentences using prosodic information,” J. Acoust. Soc. of Japan, Vol. 52, No. 12, pp. 973-978, 1996.
- [8] K. Ozeki, K. Kousaka, and Y. Zhang, “Syntactic information contained in prosodic features of Japanese utterances,” Proc. Eurospeech’97, Vol. 3, pp. 1471-1474, 1997.
- [9] Y. Hirose, K. Ozeki, and K. Takagi, “Effectiveness of prosodic features in syntactic analysis of read Japanese sentences,” Proc. ICSLP2000, Vol. 3, pp. 215-218, 2000.
- [10] K. Ozeki, K. Takagi, and H. Kubota, “The use of prosody in Japanese dependency structure analysis,” Proc. of ISCA Tutorial and Workshop on Speech Recognition and Understanding, pp. 123-126, 2001.
- [11] K. Takagi and K. Ozeki, “Pause information for dependency analysis of read Japanese sentences,” Proc. Eurospeech2001, Vol. 2, pp. 1041-1044, 2001.
- [12] K. Takagi, H. Kubota, and K. Ozeki, “Combination of pause and F_0 information in dependency analysis of Japanese sentences,” Proc. ICSLP2002, Vol. 2, pp. 1173 - 1176, 2002.
- [13] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, “Manual of Japanese Speech Database,” ATR, 1990.
- [14] 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, pp. 35-57, 1994.