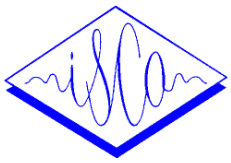

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PREDICTION OF PROSODIC PHRASE BOUNDARIES USING STOCHASTIC CONTEXT-FREE GRAMMAR

Shigeru FUJIO, Yoshinori SAGISAKA and Norio HIGUCHI

ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-02 Japan

ABSTRACT

We propose a method of predicting prosodic phrase boundaries using a SCFG (stochastic context-free grammar) for non-bracketed input word attributes. In this paper, prosodic phrase boundaries are predicted using information generated by a SCFG trained using phrase dependency bracketings and prosody bracketings. Using the Inside-Outside algorithm for training, corpora with phrase dependency brackets are first used to train the SCFG from scratch. Next, this SCFG is re-trained using the same corpora with prosody brackets. Then, the probability of each bracketing structure is computed using the SCFG, and is used as a parameter in the prediction of the prosodic phrase boundaries. To examine the effect of these parameters, prosodic phrase boundaries were predicted. Held-out data: 92.9% of the prosodic phrase boundaries, and 83.1% of the non-prosodic phrase boundaries, were correctly predicted.

1. INTRODUCTION

Prediction of prosodic phrase boundaries is important for natural-sounding speech synthesis. A F_0 pattern of Japanese speech in a sentence can be described by partial ups and downs grouped into one or two phrases (accent phrase) and superimposed on a gentle downslope. At most boundaries between accent phrases, the downslope is kept and creates accent phrases that are in the same prosodic group. But at some boundaries, the underlying F_0 declination is reset and a prosodic phrase boundary is made. The problem of predicting these prosodic phrase boundary locations has not been completely solved yet. Until now, various heuristic rules[1] and rules based on statistical analysis[2] have been used to predict these boundaries. However, these rules have required information about linguistic brackets, which are difficult to determine automatically. This paper describes a method of predicting prosodic phrase boundaries for non-bracketed input word attributes and using a SCFG (stochastic context-free grammar), and reports experimental results using this prediction method.

2. A METHOD FOR PREDICTING PROSODIC PHRASE BOUNDARIES

Phrase dependency structures and words before and after boundaries have commonly been used for predicting prosodic phrase boundaries. The accurate formulation of these has been difficult because phrase dependency structures (which are important factors) are decided by syntactic structures, linguistic relationships among the words and so on. In this paper, phrase dependency structures and prosodic structures are learned by a SCFG, enabling us to predict prosodic phrase boundaries without phrase dependency structures as input. With this method, prosodic phrase boundaries are predicted using information generated by a SCFG trained using phrase dependency and prosody bracketings. Prosodic phrase dependency structure is learned by the SCFG. The parameters representing stochastic information about the prosodic phrasing are computed using the probability parameter of the SCFG. Then, prosodic phrase boundaries are predicted using these parameters and words before and after the boundaries. The flow of the generation algorithm for the parameters is shown in Figure 1.

2.1. Learning a prosodic phrase dependency structure using SCFG

For the training of the SCFG, an efficient Inside-Outside algorithm has already been proposed[3] and an extension of this has been applied to partially bracketed text corpora[4]. We are using this algorithm to learn a prosodic phrase dependency structure using this algorithm[5]. Using this method for training the SCFG, corpora with phrase dependency brackets are first used to train the SCFG from scratch. Next, this SCFG is re-trained using the same corpora with prosody brackets. Determining good terminal symbols and number of non-terminal symbols is necessary for obtaining SCFG with high accuracy.

Considering the limitations of data size and computational cost, part of speech(POS) and divided postpositional particles were selected as terminal symbols. In this paper, a further examination of terminal symbols that are POS and some words is carried out to obtain SCFG with higher accuracy by dividing some POS. Because postpositional particles strongly influence the degree of connection between phrases[6], we think that dividing postpositional particles will have a good effect on the accuracy of SCFG.

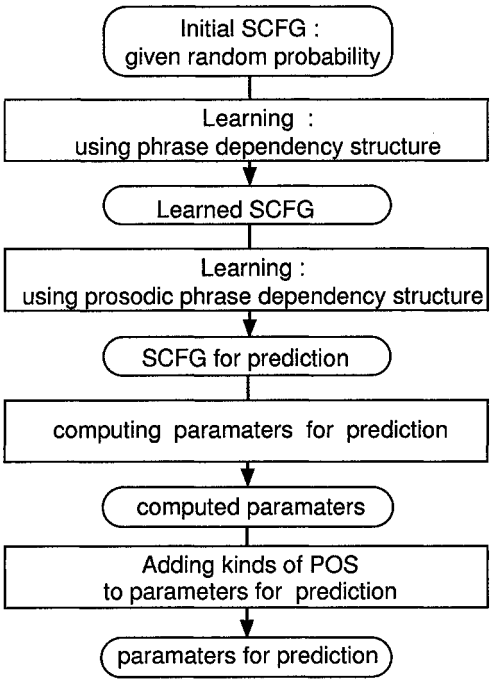

Figure 1. Flow of generation of the parameters for prediction

In the division of postpositional particles, words used more than fifty times in the continuous speech data set(503 sentences)[7] were used as terminal symbols. Therefore, the following sets of terminal symbols were proposed.

(1) POS (adjective, nominal noun, verbal noun, pronoun, quantifier, adverb, attributive, conjunction, exclamation, auxiliary verb, adverbial particle, conjunctive particle, case particle, final particle, suffix, prefix, proper noun, adjectival noun, verb, adnominal particle, coordinate particle, modal particle, others)

(2) (1) excepted case particle + divided case particles("ga","no","ni","wo","de","to",others)

(3) (1) excepted conjunctive particle + divided conjunctive particles("te",others)

(4) (1) excepted case particle + divided modal particles("ha",others)

The terminal symbols of (1) were only POS, and 23 pieces. The terminal symbols of (2) were POS and divided case particles, and 29 pieces. The terminal symbols of (3) were POS and divided conjunctive particles, and 24 pieces. And the terminal symbols of (4) were POS and divided modal particles, and 24 pieces.

2.2. Parameters for prediction of prosodic phrase boundaries

For the prediction of the prosodic phrase boundaries, the parameter computed using the SCFG is considered to use stochastic information of prosodic phrase grasped by SCFG. As shown in Figure 2, at each word, P_m:the probability that the word is part of a left-branching structure

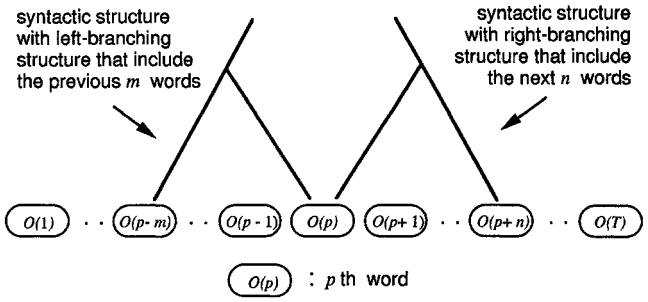

Figure 2. syntactic structure

which includes the previous m words and Q_n:the probability that the word is part of a right-branching structure which includes the next n words can be calculated using a SCFG as follows. $a[i,j,k]$ is the probability that the non-terminal symbol i will generate the pair of non-terminal symbols j and k. $b[i,m]$ represents the probability that the non-terminal symbol i will generate a single terminal symbol m. These parameters are sufficient to describe any SCFG. In Inside-Outside algorithm, the inner probability $e(s,t,i)$ is defined as the probability of the non-terminal symbol i generating the observation $O(s),\cdot\cdot,O(t)$ and can be expressed as follows:

CASE 1: $s = t$

$$e(s,s,i) = b[i,O(s)]$$

CASE 2: $s \neq t$

$$e(s,t,i) = \sum_{j,k}\sum_{r=s}^{t-1} a[i,j,k]e(s,r,j)e(r+1,t,k)$$

And the outer probability $f(s,t,i)$ may be thought of as the probability that, in the rewrite process, i is generated and that the strings not dominated by it are $O(1),\cdot\cdot,O(s-1)$ to the left and $O(t+1),\cdot\cdot,O(T)$ to the right. Hence:

$$\begin{aligned} f(s,t,i) &= \sum_{j,k}[\sum_{r=1}^{s-1} f(r,t,j)a[j,k,i]e(r,s-1,k) \\ &\quad + \sum_{r=t+1}^{T} f(s,r,j)a[j,i,k]e(t+1,r,k)\,] \end{aligned}$$

$$and \quad f(1,T,i) = \begin{cases} 1 & \text{if } i = S \\ 0 & \text{otherwise} \end{cases}$$

The non-terminal symbol i can be thought of having two possible settings $j \rightarrow ik$ or $j \rightarrow ki$. $f_l(s,t,i)$ is the probability when only settings $j \rightarrow ik$ are considered and $f_r(s,t,i)$ is the probability when only settings $j \rightarrow ki$ are considered. These are expressed as follows:

$$f_l(s,t,i) = \sum_{j,k}\sum_{r=t+1}^{T} f(s,r,j)a[j,i,k]e(t+1,r,k)$$

$$f_r(s,t,i) = \sum_{j,k}\sum_{r=1}^{s-1} f(r,t,j)a[j,k,i]e(r,s-1,k)$$

Therefore, the probability that the observation $O(1), \cdot\cdot, O(T)$ has a left-branching structure which includes the observation $O(s), \cdot\cdot, O(t)$ is:

$$\sum_i e(s,t,i) f_l(s,t,i)$$

The probability that the observation $O(1), \cdot\cdot, O(T)$ has a right-branching structure which includes the observation $O(s), \cdot\cdot, O(t)$ is:

$$\sum_i e(s,t,i) f_r(s,t,i)$$

And the probability generated for the entire observation $O(1), \cdot\cdot, O(s), \cdot\cdot, O(t), \cdot\cdot, O(T)$ is $e(1,T,S)$. Therefore, P_m and Q_n at the pth word are as follows:

$$P_m = \frac{\sum_i e(p-m,p,i) f_l(p-m,p,i)}{e(1,T,S)}$$

$$Q_n = \frac{\sum_i e(p,p+n,i) f_r(p,p+n,i)}{e(1,T,S)}$$

These probabilities are used as parameters for the prediction of the prosodic phrase boundaries because these represent phrase dependency structures that affect determination of prosodic phrase boundaries.

3. EXPERIMENTS

3.1. Learning SCFG

The continuous speech data set parsed morphemes are given the following two types of bracketing information, and served as the corpora for the training the SCFGs.

- Phrase dependency bracketing
  Phrase dependency structure was hand-tagged by trained transcribers according to conventional Japanese school grammar. As hand-tagged phrase dependency structure is determined using both syntactic and semantic relations, we only expect that syntactic information implicitly manifested in this bracketing will be captured through SCFG training.
- Prosody bracketing
  Only accent phrase boundaries and prosodic phrase boundaries were considered in bracketing. By listening to speech data and observing analyzed F_0 contour, accent phrase-sized units were manually bracketed. Prosodic phrase bracketing was automatically carried out by finding F_0 boosting boundaries were two succeeding accent phrases did not show declining F_0 characteristics.

3.1.1. Influence of terminal symbol

To examine the influence by different terminal symbols, the following experiments were carried out. For learning prosodic phrase structures, several SCFGs each using different terminal symbols (refer to paragraph 2.1.) were first trained using corpora with phrase dependency brackets. Next, these SCFGs were re-trained using the same corpora with prosody brackets. All SCFGs used 15 non-terminal symbols. To evaluate these SCFGs, the percentage of compatible bracketings (two compatible bracketings don't overlap each other) predicted for the target corpus(for example, corpora with prosody brackets) were computed. Table 1 shows the compatibility scores of SCFG each using different terminal symbols. The score of experiments for held-out data in table 1 were obtained as follows. The corpus were divided into ten parts, and nine parts were used for training, and the remaining one was used as test corpus. Ten experiments were carried out using each part as test corpus, and the average results were computed. The results show that the addition of terminal symbols by the division of case particles has a better effect on the accuracy of the SCFG.

Table 1. Comparison among compatibility scores of SCFGs each using different terminal symbols

corpus	terminal symbols	compatibility score: target corpus: linguistic brackets	compatibility score: target corpus: prosodic brackets
training corpus	POS	80.73 %	88.37 %
	POS + divided case particles	82.52 %	90.46 %
	POS + divided conjunctive particles	80.65 %	88.65 %
	POS + divided modal particles	81.43 %	89.50 %
test corpus	POS	79.83 %	87.65 %
	POS + divided case particles	81.28 %	88.30 %
	POS + divided conjunctive particles	79.81 %	87.55 %
	POS + divided modal particles	79.93 %	87.53 %

3.1.2. Influence of non-terminal symbol

The training speed of SCFG using Inside-Outside algorithm is determined as $O(N^3)$ by the number of non-terminal symbols N. Therefore, we should select a good number of non-terminal symbols. Experiments to examine the influence of the number of terminal symbols were carried out as follows. Several SCFGs (each using a different number of non-terminal symbols) were trained as described in the previous paragraph, and the compatibility scores of the SCFGs were evaluated. Each SCFG used the same terminal symbols: POS excepted case particle + divided case particles. Table 2 shows the compatibility scores for the SCFGs using different number of non-terminal symbols. The results show that the accuracy of SCFGs were not improved by increasing the number of non-terminal symbols above 15.

Table 2. Comparison among compatibility scores of SCFGs with different number of non-terminal symbols

corpus	number of non-terminal symbols	compatibility score target corpus: linguistic brackets	compatibility score target corpus: prosodic brackets
training corpus	10	77.99 %	86.19 %
	15	82.52 %	90.46 %
	20	83.09 %	90.43 %
	25	82.45 %	91.22 %
test corpus	10	77.71 %	85.26 %
	15	81.28 %	88.30 %
	20	80.16 %	88.37 %
	25	81.02 %	89.05 %

3.2. Prediction of prosodic phrase boundaries

To examine the effect of the parameters proposed in paragraph2.2., prosodic phrase boundaries were predicted. In these experiments, a feed-forward type neural network was used to predict a category from many analogue parameters. The parameters to predict prosodic phrase boundaries were computed using the trained SCFG which has 20 (non-terminal symbols) and POS + divided case particles (terminal symbols).

3.2.1. Training of neural network

The structure of the neural network used for prediction is four layers: input layer with 50 units, two hidden layers with 25 units, and an output layer with 2 units. The input parameters are P_m and Q_n at words, the independent words (because Q_n is nearly zero at attached words) around the boundary, and the kind of terminal symbol of words around the boundary as follows.

- P_m and Q_n at the following words
  (m,n=1,2,3,4 and sum of more than 5)
  - The independent word preceding the word before the boundary
  - The word before the boundary
  - The word after the boundary
  - The independent word following the word after the boundary
- The terminal symbols of the 5 words preceding the boundary
- The terminal symbols of the 5 words following the boundary

The neural network was trained using sets of the above parameters and a pair of parameters(prosodic phrase boundary:(0,1), others:(1,0)) which were decided by finding F_0 boosting boundaries where two succeeding accent phrases did not show declining F_0 characteristics.

3.2.2. Prediction of prosodic phrase boundaries

The prediction of prosodic phrase boundaries was carried out using the neural network described in the previous paragraph. As prosodic phrase boundaries were (0,1) and others are (1,0) in training, prediction of prosodic phrase boundaries was carried out by comparing the outputs of the two units in the output layer.

Table 3. Prediction results

corpus	percentage of correct prediction: boundaries that all speakers reset F_0	percentage of correct prediction: boundaries that no speaker resets F_0
training corpus	99.4 % (676/680)	92.4 % (5280/5715)
test corpus	92.9 % (632/680)	83.1 % (4751/5715)

As some prosodic phrase boundaries are optional, there are boundaries where all speakers reset F_0, boundaries where no speaker reset F_0, and others where some speakers reset F_0. Therefore, prosodic phrase boundaries are predicted at boundaries at which four speakers in the database reset F_0. The results, as shown in table 3, indicate that a using SCFG is effective for prediction of prosodic phrase boundaries.

4. CONCLUSION

We examined the prediction method for prosodic phrase boundaries using a SCFG from non-bracketed input word attributes, and proposed the prediction parameters computed by the SCFG trained using prosodic phrase structure. Our results show that the proposed parameters are effective for the prediction of prosodic phrase boundaries. In the future, the parameters to predict prosodic phrase boundaries will be examined for improvement in prediction accuracy, and the method of learning the prosodic phrase dependency structure by SCFG will be considered for obtaining SCFG with high accuracy.

Acknowledgments

We would like to thank Dr. Y.Schabes and Dr. F.Pereira for providing the program for Inside-Outside training.

REFERENCES

[1] K.Hirose,H.Fujisaki,H.Kawai,M.Yamaguchi :"Manifestation of linguistic and para-linguistic information in the voice fundamental frequency contours of spoken japanese",Proc. ICSLP, pp.485-488, 1990

[2] N.Kaiki,Y.Sagisaka :"Optimization of intonation control using statistical F_0 resetting characteristics", Proc. ICASSP, Vol.2pp.49-52, 1992

[3] K.Lari,S.J.Young :"The estimation of stochastic context-free grammars using the Inside-Outside algorithm", J.Computer Speech and Language, Vol.4pp.35-56,1990

[4] F.Pereira,Y.Schabes :"Inside-outside reestimation from partially bracketed corpora", Proc. ACL, pp.128-135,1992

[5] Y.Sagisaka,F.Pereira :"Inductive learning of prosodic phrasing characteristics using stochastic context-free grammar", The Acoustic Society of Japan Spring Meeting Proc., 2-8-10(1994,3)

[6] S.Nakajima,K.Kabeya :"Relations between phrase structure and pitch contour", The Acoustic Society of Japan Spring Meeting Proc., 2-2-8(1984,3) (in Japanese)

[7] M.Abe,Y.Sagisaka,H.Kuwabara :"The integration of linguistic, prosodic information and fundamental frequency in a continuous speech database", The Acoustic Society of Japan autumn Meeting Proc., 2-3-22(1989,10) (in Japanese)