ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PROSODIC PHRASING IN KOREAN; DETERMINE GOVERNOR, AND THEN SPLIT OR NOT

Yeon-Jun Kim, Heo-Jin Byeon, and Yung-Hwan Oh
Department of Computer Science, KAIST
373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea
yjkim@bulsai.kaist.ac.kr
http://bulsai.kaist.ac.kr

10.21437/Eurospeech.1999-139

ABSTRACT

This paper introduces a prosodic phrasing method in Korean to improve the naturalness of speech synthesis, especially in text-to-speech conversion. In prosodic phrasing, it is necessary to understand the structure of a sentence through a language processing procedure, such as POS tagging and parsing, since syntactic structure correlates better with the prosodic structure of speech than with other factors.
In this paper, the prosodic phrasing procedure is treated from two perspectives: dependency parsing and prosodic phrasing using dependency relations. This is appropriate for Ural-Altaic, since a prosodic boundary in speech usually concurs with a governor of dependency relation. From experimental results, using the proposed method achieved 12% improvement in prosody boundary prediction accuracy with a speech corpus consisting 300 sentences uttered by 3 speakers.

Keywords: prosodic phrase, dependency relation

1. INTRODUCTION

In continuous speech, most speakers tend to group words into phrases whose boundaries are marked by a pause and a intonational change. Spoken language as well as written language both contain phrases consisting of several words. However the phrases of spoken language are different from that of written language. While a phrase in written language is determined by a consistent grammar, a phrase in spoken language is determined by the speaker's utterance, which is termed *prosodic phrase* to distinguish it from the written language phrase.

The derivation of the prosodic phrase structure of a sentence is important in understanding spoken language. Speech synthesized without its appropriate prosodic structure is unnatural, monotonous and boring, whose meaning is also difficult to follow in longer passages. Because the speaker uses prosody for a number of reasons including the conveyance of meaning. Assigning appropriate prosody is also important in improving intelligibility, particularly in longer sentences.

Compared with other factors, syntactic phrase structure correlates relatively well with the prosodic structure of speech uttered at a normal rate, without emotional and other contextual influences. In many cases, however, the boundaries of the syntactic constituents are not aligned with the prosodic phrase boundary. Because most theories of prosodic structure define a hierarchy of prosodic constituents based on phrase structured grammar, prosodic structure is known to be determined linearly from left to right.

To resolve this problem, many linguists have proposed methods and theories to explain the relation between syntactic and prosodic structure [1][2]. Though the *readjustment rules* by Chomsky and Halle and the *verb balancing rule* by Gee and Grosjean were proposed, they were incomplete explanations. Many researchers have noticed on the need to flatten syntactic structures to predict prosodic structures. A recent approach by Hunt employs a flat syntactic representation, *link grammar*, which draws links between the syntactically related words themselves [4]. He showed that each *Surface-Syntactic Relation* labeled by a link has an intrinsic prosodic coupling strength.

In this paper, we adopt another flat syntactic scheme for finding prosodic structure:*dependency grammar*. This is more effective in analyzing languages with freer surface word orders, such as Ural-Altaic. The word order of Korean is quite free unlike the fixed word order of English.
And we propose a probabilistic model for predicting prosodic boundaries and show the effectiveness of the dependency relation, which can be be incorporated into existing text-to-speech systems.

2. DETERMINE GOVERNOR

2.1. Definition of Dependency Relation

While phrase structured grammar is defined in terms of the recursive groupings of syntactic units, *Dependency grammar* is based on the relations between actual elements. Assuming a dependency relation between two words, w_i, w_j, which can be denoted by '$w_i \leftarrow w_j$' where w_i is the *dependent* and w_j is the *governor*. Also it can be said that w_i depends on w_j or w_j governs w_i.

And a sequence of words can be a sentence of the language defined by dependency grammar if there exists a set of relations joining those words that satisfies three conditions:

1. *Unique Governor*: Each word in a sentence has its own unique governor except for the sentence governor which is usually the main verb.

2. *Planarity*: Relations do not cross (when drawn above the words).
3. *Governor Post-position* : A *governor* is always a post-position of its *dependent* (in Korean).

Condition 3, *Governor Post-position*, is modified because of the characteristics of Korean. Since Korean has the governor post-positioning property, dependency parsing can be implemented more easily using the restricted CNF (Chomsky normal-form) rules with fewer consequent ambiguities. The three types of CNF rules for Korean dependency parsing are the following:

1. **S** $\rightarrow$ **A**
2. **A** $\rightarrow$ **B A**
3. **A** $\rightarrow$ **a**

where **S** is a start symbol, **A** and **B** are the part-of-speech sequence of word-phrase which consists of several morphemes, and **a** is a word phrase which is a basic unit in Korean. As in other Ural-Altaic language, a Korean sentence is composed of larger grammatical units formed of several morphemes, called a *word-phrase* similar to *bunsetsu* in Japanese.

2.2. Probabilistic Dependency Parsing

Probabilistic Dependency paring can be done by finding the most probable dependency tree while satisfying the above three conditions. A *dependency tree* is defined as a set of governor-dependent relationships which reveals the structure of an expression in terms of hierarchical links among its actual elements. To find the most probable dependency tree, we make use of two types of probabilities based on the above CNF rules.

1. *dependency probability*: the probability that a dependency relation between a tag seq. **B** and another tag seq. **A** may occur,

$$Pr(\mathbf{A} \rightarrow \mathbf{B}\mathbf{A}) \stackrel{def}{=} Pr(\mathbf{B}, \mathbf{A}|\mathbf{A}) = Pr(\mathbf{B}|\mathbf{A}) \quad (1)$$

2. *lexical probability*: the probability that a word-phrase **a** belongs to a POS tag seq. **A**,

$$Pr(\mathbf{A} \rightarrow \mathbf{a}) \stackrel{def}{=} Pr(\mathbf{a}|\mathbf{A}) \quad (2)$$

The probability based on the first type of CNF rule (**S** $\rightarrow$ **A**) is not considered in the dependency parsing phase because the last word-phrase always becomes the head of a Korean sentence. Dependency probabilities can be extracted from the corpus bracketed with dependency relations, and lexical probabilities are generated at the POS tagging phase.

3. SPLITTING ALGORITHM FOR PROSODIC PHRASE

In this section, a prosodic phrasing algorithm based on a probabilistic approach is described. Splitting prosodic group into two is more emphasized in determining prosodic phrase boundaries while the recursive grouping of the syntactic units is a key operation. Our splitting algorithm assigns probabilities to potential prosodic boundaries for a given input sentence.

3.1. Basic Algorithm

In a given sentence with word sequence, $w_1^N = \{w_1, w_2, \ldots, w_N\}$, the probabilistic model is defined as in Eq.(3) to get an optimal sequence of prosodic phrase boundaries, $b_1^N = \{b_1, b_2, \ldots, b_N\}$ where b_i exists between adjacent words, $\langle w_i, w_{i+1}\rangle$.
By applying *Bayes decision rule* to $Pr(b_1^N|w_1^N)$ in Eq.(3), we can transform the problem into a likelihood maximization problem with a priori probability as shown in Eq.(4).

$$\phi(w_1^N) \stackrel{def}{=} \arg\max_{b_1^N} Pr(b_1^N|w_1^N) \quad (3)$$

$$= \arg\max_{b_1^N} \frac{Pr(w_1^N|b_1^N)Pr(b_1^N)}{Pr(w_1^N)} \quad (4)$$

Given a word (w_i), the POS information of w_i is chosen to maximize the likelihood of the training data as Eq.(6). This is a factor generally used to determine prosodic phrase boundaries especially in text-to-speech conversion [3].
As mentioned in the previous section, a Korean word consists of several morphemes. The POS of a morpheme in the post-position of a word-phrase and that of a content word of the next word-phrase are used because the relation between them is a dominant factor in determining whether there is a prosodic boundary or not.

$$Pr(w_1^N|b_1^N) = \prod_{i=1}^{N} Pr(w_i|w_{i+1}^N, b_1^N) \approx \prod_{i=1}^{N} Pr(w_i|w_{i+1}, b_1^N) \quad (5)$$

$$\approx \prod_{i=1}^{N} Pr(POS(f_{w_i})|POS(c_{w_{i+1}}), b_1^N) \quad (6)$$

In addition to POS information, a physiological factor is adopted to solve the sparseness of a sequence of prosodic boundaries, b_1^N.
Generally, a longer utterance may tend to be produced with more boundaries for reasons of physiological production. The distance of the boundary site from the beginning and end of the utterance is another variable likely to be correlated with boundary locations. In our work, the distance is defined as the number of syllables from the current word (w_i) to the previous prosodic phrase boundary, Δ_{b_i}. Assuming that w_i is depend only on b_i and the distance (Δ_{b_i}), we can simplify Eq.(6) into Eq.(7).

$$Pr(w_1^N|b_1^N) \approx \prod_{i=1}^{N} Pr(POS(f_{w_i})|POS(c_{w_{i+1}}), \langle\Delta_{b_i}, b_i\rangle) \quad (7)$$

$$= \prod_{i=1}^{N} \frac{Pr(POS(f_{w_i}), POS(c_{w_{i+1}}), \Delta_{b_i} | b_i)}{Pr(POS(c_{w_{i+1}}), \Delta_{b_i} | b_i)} \quad (8)$$

To get $Pr(w_1^N | b_1^N)$ in Eq.(7), such methods as the use of a POS bigram or constituent length information were proposed, however these concentrate only on local relations between adjacent words and overlook many important factors, such as intra-sentential syntactic influences.

3.2. Extended Algorithm using Dependency Relation

In this paper, two kinds of intra-sentential features based on dependency relation are introduced in considering non-local syntactic characteristics. One is syntactic information which is the POS of the governor and dependent in a dependency relation, and the other is the physiological information related to potential boundary locations.

First, we assume that each dependency relation has an intrinsic prosodic coupling strength, which has a probability distribution estimated from the speech corpora [4]. Some relations may be more or less likely than others to be intonationally separated by prosodic boundaries.
As shown in the experimental results [Tbl.1, 2], the existence or level of the prosodic boundary has correlations with the syntactic meaning of the dependency relation, the POS information of the current word (w_i) and its governor (w_{g_i}).

Table 1: Probability distribution of occurrence of prosodic boundaries according to dependency relation
(Δ_{phy}: the distance of physiology based on dependency relation)

	Major		Minor		no-break	
(f_{w_i}, c_{w_g})	Pr(·)	Δ_{phy}	Pr(·)	Δ_{phy}	Pr(·)	Δ_{phy}
(sc, nc)	1.0	24.5	0	-	0	
(sc, vb)	0.96	40.1	0	-	0.03	1.5
(px, nc)	0.90	26.8	0	-	0.10	16.5
(pt, nc)	0.75	30.1	0.17	15.8	0.07	27.2
(pt, vb)	0.70	33.8	0.15	17.7	0.13	18.8
(ec, vb)	0.69	30.1	0.08	12.6	0.22	12.8
(ec, nc)	0.69	25.0	0.17	16.3	0.14	22.4

Table 2: Probability distribution of non-occurrence of prosodic boundaries according to dependency relation

	Major		Minor		no-break	
(f_{w_i}, c_{w_g})	Pr(·)	Δ_{phy}	Pr(·)	Δ_{phy}	Pr(·)	Δ_{phy}
(ex, vb)	0	-	0.11	3.5	0.89	11.7
(dn, nc)	0.02	6.0	0.20	7.2	0.78	9.3
(nc, vb)	0.03	22.0	0.27	11.7	0.70	12.8
(pc, ad)	0	-	0.28	11.7	0.69	12.8

In addition to the basic model (Eq.5), the syntactic meaning of the dependency relation can be used to approximate $Pr(w_1^N | b_1^N)$. Assuming that w_i is dependent on the adjacent word (w_{i+1}) as well as on the w_i's governor (w_{g_i}), the model can be extended as in Eq.9.

$$Pr(w_1^N | b_1^N) \approx \prod_{i=1}^{N} Pr(w_i | w_{i+1}, w_{g_i}, b_1^N) \quad (9)$$

$$\approx \prod_{i=1}^{N} Pr(POS(f_{w_i}) | POS(c_{w_{i+1}}), POS(c_{w_{g_i}}), b_1^N) \quad (10)$$

where $i+1 \leq g_i \leq N$ since the governor of a Korean sentence always exist in the post-position.

In the previous section, the elapsed distance from the last prosodic boundary to the current word is used to determine whether the prosodic boundary exist or not. Similarly, we can foresee the next prosodic boundary from the fact that a prosodic boundary in speech usually concurs with the governor of the dependency relation.

The influence of the dependency relation distance on prosodic boundary location was already investigated in a previous work [6]. Though it is clear that distance information is effective to some degree, using dependency relation distance alone has limitations.

In this paper, we propose a new physiological measure combining the elapsed distance from the previous prosodic boundary and the distance to the potential prosodic boundary using dependency relation. The physiological distance (Δ_{phy}) based on dependency relation refer to the number of syllable from the previous prosodic boundary to the right most immediate *governor*. As the distance increases, the probability of the occurrence of prosodic boundaries also increase [Fig. 1].

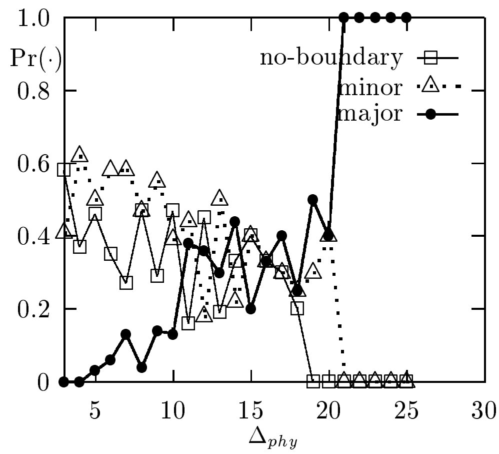

Figure 1: Probability distribution of each prosodic boundary according to the physiological distance based on dependency relation

$$
\begin{aligned}
&Pr(w_1^N|b_1^N)\\
&\approx \textstyle\prod_{i=1}^N Pr(POS(f_{w_i})|POS(c_{w_{i+1}}), \langle \Delta_{g_i}, POS(c_{w_{g_i}})\rangle, \langle \Delta_{b_i}, b_i\rangle)\\
&\approx \textstyle\prod_{i=1}^N Pr(POS(f_{w_i})|POS(c_{w_{i+1}}), POS(c_{w_{g_i}}), \Delta_{phy}, b_i)
\end{aligned}
\tag{11}
$$

where $\Delta_{phy} = \Delta_{b_i} + \Delta_{g_i}$.

Finally, a probability prosodic phrasing model in eq.(12) is established, which is included the proposed distanced measure. In this work, we obtained the best sequence of prosodic phrase boundaries applying the *Viterbi* algorithm.

$$
\phi(w_1^N) = \prod_{i=1}^{N} \frac{Pr(POS(f_{w_i}), POS(c_{w_{i+1}}), POS(c_{w_{g_i}}), \Delta_{phy}|b_i)}{Pr(POS(c_{w_{i+1}}), POS(c_{w_{g_i}}), \Delta_{phy}|b_i)} \tag{12}
$$

4. EXPERIMENTS

4.1. Corpus

The probabilistic models proposed in this paper are trained with hand-labeled corpora. In practice, this means that a large corpus of speech data must be available to accurately estimate the model parameters, such as the probability of a prosodic event given some text or acoustic- based features. Prosodic phrase boundaries in the speech corpus were hand-labeled by two listeners. The announcer was asked to read the sentences in context, using their standard radio style of speaking. Most related studies have used only the speech of professional speakers so as to maintain consistency in prosodic cues.

The performance of the probabilistic prediction model was evaluated by comparing the hand-labeled prosodic boundaries in the speech sample. The correspondence between the predicted and the observed breaks is tabulated in a confusion matrix.

4.2. Experimental Results

In this section, the prosodic phrase break prediction for the proposed probabilistic model is evaluated. We compared the experimental results of a basic algorithm with the proposed model including knowledge of the dependency relation.
In the Tbl. 4, the proposed model shows better performance in predicting accuracy than the model without dependency information.

It is clear that the value of human perceptual judgments of synthetic speech is important enough to warrant serious investigation into experimental design. Recent studies have discussed the importance of minimizing confounding factors. From the work presented here, it has become evident that a finding task is sufficiently but not excessively difficult to reveal differences in tested models.

Table 3: Experimental results using the basic algorithm

	Predicted Prosodic Marks		
	Major	Minor	no-break
Major	483	64	101
Minor	88	147	187
no-break	187	135	652

Table 4: Experimental results using the extended algorithm based on dependency relation

	Predicted Prosodic Marks		
	Major	Minor	no-break
Major	523	41	84
Minor	92	183	147
no-break	128	111	735

5. CONCLUDING REMARKS

This paper has proposed an automatically trainable, computational prosody phrasing model, which can be expected to incorporated into existing text-to-speech systems.

From the experimental results, using the proposed method achieved 12% improvement of the prosody boundary prediction accuracy with a speech corpus consisting 300 sentence uttered by 3 speakers. We were able to show that the physiological distance measure based on dependency relation is effective to predict prosodic phrase boundaries.

In the future, the proposed model can be expected to make synthesized speech more natural and improve the robustness of spontaneous speech recognition.

6. REFERENCES

[1] J. P. Gee and F. Grosjean, “Performance structure: A psycholinguistic and linguistic appraisal”, Cognitive Psychology, Vol. 15 pp.411-458, 1983.

[2] J. Bachenko and E. Fitzpatrick, “A computational grammar of discourse-neutral prosodic phrasing in English”, Computational Linguistics, Vol.16(3) pp.155-170, 1990.

[3] Michelle Q. Wang, Julia Hirschberg, “Automatic classification of intonational phrase boundaries”, Computer Speech and Language, Vol.6 pp.175-196, 1992.

[4] A. J. Hunt, Models of Prosody and Syntax and their Application to Automatic Speech Recognition, PhD thesis, University of Sydney, 1995.

[5] Sangho Lee, Yung-Hwan Oh, “A Text Analyzer for Korean Text-to-Speech Systems”, Intl. Conf. on Spoken Language Processing 96, 1996.

[6] Yeon-Jun Kim, Yung-Hwan Oh, “Prediction of Prosodic Phrase Boundaries considering Variable Speaking Rate”, Intl. Conf. on Spoken Language Processing 96, 1996.