Prediction of Prosodic Phrase Boundaries considering Variable Speaking Rate

Yeon-jun Kim, Yung-hwan Oh

Department of Computer Science
Korea Advanced Institute of Science and Technology,
373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea
e-mail: yjkim@adam.kaist.ac.kr

ABSTRACT

This paper proposes a model for predicting the prosodic phrase boundaries of speech with variable speaking rates.

Speakers can produce a sentence in several ways without altering its meaning or naturalness, i.e., a sequence of words can have a number of prosodic phrase boundaries. There are many factors which influence the variability of prosodic phrasing, such as syntactic structure, focus, speaker differences, speaking rate and the need to breathe.

In this work, we adopt dependency grammar, similar to link grammar, to efficiently combine speaking rates. The proposed model reduced prosodic phrase boundary prediction error by 20% compared the model using only syntactic informations. We show a potential way to make use of a read speech corpus in the training of prosodic phrasing for spontaneous speech. The proposed model is expected to make synthesized speech more natural, and improve the robustness of spontaneous speech recognition.

1. INTRODUCTION

In continuous speech, speakers tend to group words into phrases whose boundaries are marked by duration and intonational cues, and many phonological rules constrain operation only within such phrases, usually termed prosodic phrases. While it is generally agreed that prosodic structure has some relationship to syntactic structure, the two are not isomorphic.

Many researchers have proposed methods and theories to explain the transformation from syntactic to prosodic phrasing. Many theories of prosodic phrasing define a hierarchy of prosodic constituents based on phrase structured grammar. For example, the *readjustment rules* proposed by Chomsky and Halle would modify the syntactic structure to produce the prosodic phrases which match the intonation [1]. And the *verb balancing rule* proposed by Gee and Grosjean is another theory based on phrase structured grammar [3].

In many cases, the boundaries of syntactic constituency are not aligned with the prosodic phrasing. Because prosodic structures are determined linearly from left to right. They are not embedded recursively while a phrase structure tree reveals the recursive expression in terms of groupings of its actual elements.

Many researchers have remarked on the need to flatten syntactic structures to predict prosodic structures. Hunt's work employs a flat syntactic representation. He made use of *link grammar*, which draws links between syntactically related words not to cross. He showed that each *Surface-Syntactic Relation* labeled by a link had an intrinsic prosodic coupling strength and link grammar was effective in predicting natural phrasing [4].

In this work, we adopt *dependency grammar*, which is similar to link grammar and more effective for analyzing languages with freer surface word order as Melcuk contend [2]. The word order of Korean is quite free unlike the fixed word order of English.

Another issue dealt with is the variability of prosodic phrasing caused by the *speaking rate*. Speakers can produce a sentence in several ways without altering its meaning or naturalness, i.e., a sequence of words can have a number of prosodic phrase boundaries though it can be represented by only one syntactic structure. Therefore a model which can resolve this variability may be needed.

There are many factors which influence the variability of prosodic phrasing, such as focus, speaker differences, speaking rate and the need to breathe. In speech spoken at a normal rate, prosodic boundaries correlate well with syntactic structures, while changing speaking rates cause variations in prosodic phrasing. Speaking rate is one of the most important factors in spoken language systems, because the speaking rate of spontaneous speech is higher and more variable than that of read speech and more variabilities in speaking rate may be observed in spontaneous speech than in read speech.

In this paper, we propose a prediction model based on dependency grammar closely related to prosodic phrasing and capable of resolving the variability in prosodic phrasing caused by the speaking rate.

2. PROSODIC PHRASING MODEL

2.1. Dependency Grammar

Dependency grammar is a formal means of representing the syntactic structure of a sentence. Dependency grammar is based on the relations between syntactic units, while phrase structured grammar is defined in terms of the recursive groupings of the actual elements. Assuming a dependency relation of w_i on w_j, it can be denoted by '$w_i \leftarrow w_j$' where w_i is a *dependent* and w_j is a *governor*. The dependency tree in Fig. 1 shows which items are related to which other items and in what way.

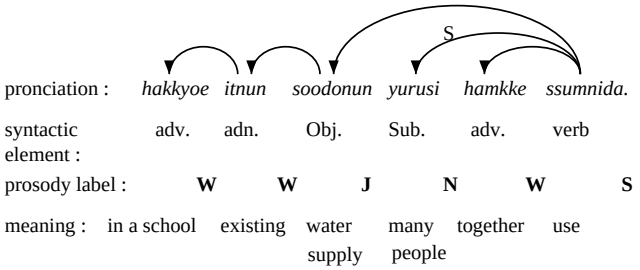

Figure 1: The dependency tree for the Korean sentence, "*hakkyoe itnun soodonun yurusi hamkke ssumnida.*"

In general, when people speak they consider the syntactic structure of sentence to determine when to pause. They don't pause in consideration of the syntactic structure itself, but decide on pauses after judging the relation between the word being uttered and the words that are to follow. From this point of view, dependency grammar is much closer to human prosodic phrasing and a dependency tree allows for a natural representation of this prosodic phrasing. We can make the following observation about prosodic phrasing.

• **observation 1** : The greater the distance of the dependency relation, i.e., the more words there are between *dependent* and *governor*, the more probable it is that the prosodic phrase boundary will follow after the dependent.

Another advantage of dependency grammar is that in parsing a language whose word order is freer it is more effective than other syntactic representations such as phrase structure grammar and link grammar. From the genealogical point of view, Korean is an *Ural-Altaic* language whose word order is free.

While a dependency tree contains only terminal nodes, most nodes in a phrase structured tree are non-terminal. A phrase structure tree reveals the recursive structure of an expression in terms of groupings of its actual elements: maximal blocks, which consist of smaller blocks, which consist of still smaller blocks. Therefore, the number of non-terminal node to represent a word group grows to be uncontrollable.

2.2. Stochastic Model

To resolve the variability of prosodic phrasing we employ a stochastic model, which has the advantage that it can be trained automatically [5].

We assume that all potential prosodic boundaries, b_i, in a sentence are between adjacent words, (w_i, w_{i+1}). A boundary(b_i) is a random variable, which can be one of a finite number of values, a 'word boundary(W)', a 'minor phrase boundary(N)', a 'major phrase boundary(J)', or a 'sentence boundary(S)'.

Given a sequence of words, $w_{1..n}$, i.e., a sentence, we can get a sequence of prosodic phrase boundaries from the stochastic prediction model defined by eq.(1).

$$\phi(w_{1..n}) \stackrel{def}{=} \arg\max_{b_{1..n-1}} P(b_{1..n-1}|w_{1..n-1}) \quad (1)$$

$$= \arg\max_{b_{1..n-1}} \frac{P(w_{1..n-1}|b_{1..n-1})P(b_{1..n-1})}{P(w_{1..n-1})} \quad (2)$$

$$= \arg\max_{b_{1..n-1}} P(w_{1..n-1}|b_{1..n-1})P(b_{1..n-1}) \quad (3)$$

$$\cong \arg\max_{b_{1..n-1}} \prod_{i=1}^{n-1} P(w_i|b_i)P(b_i|b_{i-1}) \quad (4)$$

Since a sequence of prosodic boundaries is assumed to be a *Markov* process, eq.(3) can be approximated to eq.(4).

Generally, the set of features, F_i, is used in the prediction of prosodic phrase boundaries instead of a word(w_i) itself as in the following eq.(5).

$$\phi(w_{1..n}) \cong \arg\max_{b_{1..n-1}} \prod_{i=1}^{n-1} P(F_i|b_i)P(b_i|b_{i-1}) \quad (5)$$

The stochastic prediction model allows the use of such syntactic information as syntactic structure, part-of-speech labels and constituent length.

- According to *observation 1*, we employed the distance (number of syllables, r_i) from the current word(w_i) to its *governor* determined by a *dependency relation* as the **syntactic structure**.
- **Part-of-speech information** is another factor widely used to predict boundary locations, particularly in text-to-speech. As in other Ural-Altaic languages, a Korean sentence is composed of larger grammatical units formed of several morphemes, and we call the larger grammatical unit a *word-phrase* similar to *bunsetsu* in Japanese. We use the part-of-speech of morpheme in the rear of word-phrase, which determines the component of a word-phrase in a sentence.

- Constituent length can be defined as the distance (number of syllables, l_i) from the current word(w_i) to the previous prosodic phrase boundary. It is observed that the constituent length in higher speaking rates is longer than in a normal rate, i.e., the constituent length can be determined from the speaking rate. Therefore, from now on, we call it the **speaking rate** instead of the constituent length.

Assuming that each factor is independent of the others to make the stochastic model simple, we can describe the features as the product of each condition probability and its weighting factors(w_f) as in eq.(6). We obtained the dependency relations and the part-of-speech information of the sentence via Lee's text analyzer [7].

$$P(F_i|b_i) \cong \underbrace{w_r\ P(r_i|b_i)}_{dependency\ relation} \cdot \overbrace{w_t\ P(t_i|b_i)}^{part-of-speech} \cdot \underbrace{w_l\ P(l_i|b_i)}_{speaking\ rate} \tag{6}$$

Finally, a stochastic prediction model in eq.(7) is established, which is based on the dependency relation and takes the speaking rate into consideration.

$$\phi(w_{1..n}) \cong \arg \max_{b_{1..n-1}} \prod_{i=1}^{n-1} P(r_i|b_i)P(t_i|b_i)P(l_i|b_i)P(b_i|b_{i-1}) \tag{7}$$

3. VARIABLE SPEAKING RATE

People can produce a sentence in various ways, in their own speaking style without altering meaning or naturalness. The speech of non-professional speakers especially is rather inconsistent in comparison with professional speech which generally considered to contain more consistent prosodic cues.

The results of prosodic phrasing in Fig.2 and Fig.3 show that prosodic phrasing is affected primarily by syntax but a syntactic structure can produce several ways of prosodic phrasing. Another phenomenon is described in observation 2.

• **observation 2** : The more short dependency relations there are in the diagram of a sentence, the more variabilities of prosodic phrasing there are, i.e., short dependency relations cause speakers to be confused about where to pause.

Though prosodic phrasing varies according to speaker, most existing prosodic boundaries prediction models maintain a constant speaking rate or modify only the duration of syllables for the spoken dialogue system. In some cases speaker identity has been used to improve the performance of prosodic phrase prediction [6].

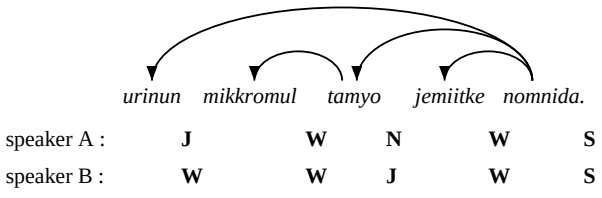

Figure 2: Dependency tree of "*urinun mikkromul tamyo jemiitke nomnida.*"

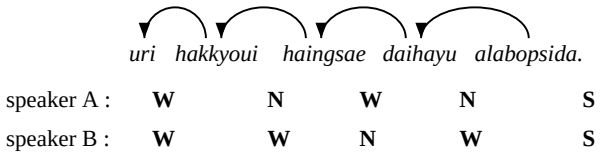

Figure 3: Dependency tree of "*uri hakkyoui haingsae daihayu alabopsida.*"

In this paper, it is assumed that the variability of speaker-dependent prosodic phrasing is caused by speaking style or breathing capacity, i.e. speaking rate. We observed that in high speaking rate speech phone durations were shorter than in a normal rate and there were more words in a prosodic phrase at higher speaking rates.

• **observation 3** : The higher the speaking rate is, the more words there are in a prosodic phrase, i.e., the number of boundaries decreases as the speaking rate rises.

Based on the above observations, we built up a stochastic prediction model considering variable speaking rates. In eq.(8) $\tilde{l}_i$ denotes variable speaking rate.

$$\phi(w_{1..n}) \cong \arg \max_{b_{1..n-1}} \prod_{i=1}^{n-1} P(\tilde{l}_i|b_i)P(t_i|b_i)P(r_i|b_i)P(b_i|b_{i-1}) \tag{8}$$

Provided the above model is trained using a large corpus including various speaking rates, it will be able to represent the wide variety of speaking styles and produce a robust means of discovering the relationship between syntax and prosody. But it is very difficult to collect speech with various speaking rate.

For this reason, we trained the prediction model with the speech spoken by a professional anouncer and then adapted the model to the speaking rate of the speech which was tested. Our hope was that the professional speaking style would be easier to label prosodically and that it would be easier to train the prediction model. And according to observation 3 we adapted the model by shifting the output probability distribution as in Fig. 4.

In this work, we obtained the best sequence of prosodic

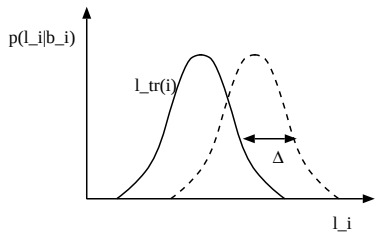

Figure 4: Shifting of the trained output probability distribution ($P(l_{tr(i)}|b_i)$) considering the *speaking rate*

phrase boundaries applying the *Viterbi* algorithm to the model(eq.9).

$$\phi(w_{1..n}) \cong \arg \max_{b_{1..n-1},\Delta} \prod_{i=1}^{n-1} P(l_{tr(i)}+\Delta|b_i)P(t_i|b_i)P(r_i|b_i)P(b_i|b_{i-1}) \quad (9)$$

4. EXPERIMENTAL RESULTS

The speech corpus for training the stochastic prediction model was recorded by a professional female announcer and for testing the model recorded by two non-professional speakers. Most of the related studies treated the speech of the professional speaker only to maintain the consistency of prosodic cues. But in this work we concentrate on the modeling of prosodic variability. Prosodic phrase boundaries, {W, M, J, S}, in the speech corpus were hand-labeled by two listeners.

To test the proposed model, we used a corpus of 10 sentences recorded by two non-professional speakers, a man and a woman whose speaking styles were different from the announcer's. The performance of the stochastic prediction model was evaluated by comparing the hand-labeled prosodic boundaries in the speech. We can observe the following from the experimental results.

	prediction rate(%)
Only syntactic information (train)	76.7
Only syntactic information (test)	48.5
Considering speaking rate (train)	76.7
Considering speaking rate (test)	62.8

Table 1: The result of prosody boundaries prediction

When the probabilities distribution shift in speaking rate(Δ) was +5, the prediction accuracy for the test data was the best. As Table 1 shows, the prediction considering speaking rate is more accurate than when only syntactic information was used for the test data. As a result, we can see evidence that considering the speaking rate is one way to model the variability of prosodic phrasing and speaker individuality.

5. CONCLUSION

In this work, speaking rate, i.e. speaking style, is considered both a feature as well as syntactic information for determining prosody boundaries.

The prediction error decreased by about 20% comparing the existing prediction model to the test data. This work shows a potential way to make use of a read speech corpus in the training of prosodic phrasing for spontaneous speech. The proposed model is expected to make synthesized speech more natural, and improve the robustness of spontaneous speech recognition.

The experimental results using this model are encouraging, but it is clear that much more data must be used for training and testing. Our model has been tested on a narrow set of sentences, so it may be necessary to continue training on more speech corpora to improve the robustness and accuracy of the prediction model.

6. REFERENCES

1. N. Chomsky, M. Halle, *The Sound Pattern of English.* Harper and Row, New York, 1968.
2. I. A. Melcuk, *Dependency Syntax: Theory and Practice.* State University of New York Press, 1988.
3. J. Bachenko, E. Fitzpatrick, "A computational grammar of discourse-neutral prosodic phrasing in English". *Computational Linguistics*, Vol.16(3) pp.155-170, 1990.
4. A. J. Hunt, *Models of Prosody and Syntax and their Application to Automatic Speech Recognition.* Phd thesis, University of Sydney, 1995.
5. M. Ostendorf, N. Veilleux, "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location", *Computational Linguistics*, Vol.20(1) pp.27-52, 1994.
6. Michelle Q. Wang, Julia Hirschberg, "Automatic classification of intonational phrase boundaries", *Computer Speech and Language*, Vol.6 pp.175-196, 1992.
7. Sangho Lee, Yung-hwan Oh, "A Text Analyzer for Korean Text-to-Speech Systems", *Intl. Conf. on Spoken Language Processing 96*, 1996, (to appear).

ACKNOWLEDGEMENTS

The authors greatly appreciate the help of Sangho Lee for the text analyzer and the prosodic labeled corpus used in this study.