Incorporating Second-Order Information Into Two-Step Major Phrase Break Prediction for Korean

*Seungwon Kim, *Jinsik Lee, **Byeongchang Kim and *Gary Geunbae Lee

*Department of Computer Science & Engineering
Pohang University of Science and Technology, Pohang, South Korea
*{rockzja, palcery, gblee}@postech.ac.kr
**School of Computer & Information Communication Engineering,
Catholic University of Daegu, Daegu, South Korea
** bckim@cu.ac.kr

ABSTRACT

In this paper, we present a new phrase break prediction method that integrates second-order information into general maximum entropy model. The phrase break prediction problem was mapped into a classification problem in our research. The features we used for the prediction of phrase breaks are of several layers such as local features (part-of-speech (POS) tags, a lexicon, lengths of eojeols[1] and location of juncture in the sentence), global features (chunk label derived from a eojeol parse tree) and second-order features (distance probability of previous and next phrase break). These three features were combined and used in the experiments, and we were able to generate good performance especially in the major phrase break prediction.

Index Terms: phrase break, prosodic phrasing, speech synthesis, ToBI

1. INTRODUCTION

One of the crucial problems in the high quality text-to-speech (TTS) system is assigning appropriate phrase breaks from raw text. Phrase breaks form a prosodic structure in the given sentence, which makes the synthesized speech more natural and intelligible. On the other hand, a phrase break that is missing or inserted into a wrong position in a sentence can change the original meaning of the sentence. Phrase break information in TTS system affects other modules like grapheme-to-phoneme conversion and prosodic feature generation such as the duration or the tone assignment. Therefore, using the wrong phrase break information decreases the performance of other modules that are using the information.

Many methods have been introduced to predict phrase break, such as Hidden Markov Models (HMM) [1], Classification and Regression Trees (CART) [2], Maximum Entropy (ME) [3] and Bayesian approach [4].

The recent research has two different approaches for the phrase break prediction. The first approach uses probabilistic methods on the large labeled corpus. The second approach uses the relationship between syntactic structure and prosodic phrasing structure. The phrase breaks appear mainly at the junctures between the major syntactic phrases. The major problem of the first approach is that the method uses only local information. The second approach can use some global information, but the unreliable performance of syntactic parser decreases the phrase break prediction accuracy.

Research on prosodic phrasing agrees that the distribution of the length of prosodic phrase is an important feature on the phrase break prediction [5]. Nevertheless, the previous phrase break prediction approaches were not able to use the length of prosodic phrase such as the number of syllables between the current juncture and the previous phrase break because the information was too difficult to be used for the real-time TTS system.

In this paper, we propose the use of the previous and the next major break's relative locations as the second-order information for a two-step major break prediction method. The probabilistic classification approach, which uses the second-order information such as a distance probability between phrase breaks, is proposed based on the conventional usual features of local information and global information such as syntactic information.

2. CORPUS STATISTICS

The SITEC (Speech Information Technology & Industry Promotion Center) TTS corpus, called SynthFemale01 corpus, is one of the standard corpus for Korean TTS, and contains 3,306 sentences which consist of 112,022 syllables or 37,765 eojeols. The corpus was labeled by an expert annotator following the K-ToBI (Korean tones and break indices) specification [6]. The break indices in the corpus consist of b_0, b_1, b_2 and b_3 (b_0 represents no boundary, b_1 for boundary of prosodic word, b_2 for boundary of accentual phrase (AP) and b_3 for boundary of intonational phrase (IP)). However, b_1 is too short for synthesizing speech sound, thus we combine b_1 with b_0. Therefore the corpus has three types of juncture: major break, minor break and non-break.

- major break: a strong phrasal juncture such as an IP boundary.
- minor break: minimal phrasal juncture such as an AP boundary.
- non-break: phrase-internal word boundary and a juncture smaller than a word boundary.

[1] An eojeol is a Koran spacing unit (similar to English word) which usually consists of one or more stem morphemes and functional morphemes.

The statistics of three types of junctures for SynthFemale01 are shown in Table 1.

Table 1: Statistics of break types in the corpus

Types of juncture	major break	minor break	non-break
The number of occurrence	15,062	16,438	6,265
Occurrence probability	39.88%	43.53%	16.59%

The distances between major breaks tend to be balanced. The short and long phrases are less frequent than the average distance phrases. Figure 1 shows the distribution of the distances between major phrase breaks expressed in terms of syllables.

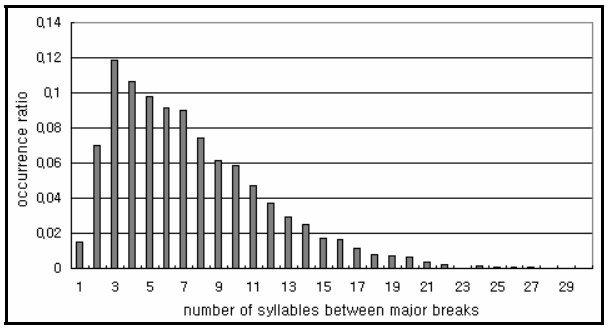

Figure 1: Distribution of the distance between major breaks.

In other words, the previous and the next phrase breaks can affect the current phrase break prediction. Therefore, we can expect that this kind of second-order information will improve the performance of the phrase break prediction.

3. TWO-STEP PREDICTION METHOD

In order to use the second-order information in the TTS system, we can consider several methods, such as using full parse tree, Conditional Random Fields (CRF) [7] and Maximum Entropy Markov Models (MEMM) [8]. But these methods have some problems; Full parse tree and CRF are too time consuming to apply to the TTS system. In addition, since the models are trained only with the reference data, they result in low performance in predicting the unseen data. On the other hand, the MEMM which uses the second-order information extracted from the reference data may introduce a "label bias problem" [7]. In particular, TTS system has high chance of encountering with unseen data.

Therefore we propose a new phrase break method which is well suited for using the second-order information. We specially modified the Stacked Sequential Learning (SSL) [9] with maximum entropy learner [3] by changing the three-way decision phrase break prediction problem into multiple two-way decision problems in a two-step architecture. Moreover, since SSL is trained and predicted on a predicted data, SSL is more robust to the unseen data.

3.1 Phrase Break Prediction Algorithm

In TTS system, the major breaks are much more important than the minor breaks. Also the major breaks are strongly influenced by the distance between the other major breaks. Therefore we inference the major breaks with the modified stacked sequential learning and separated the major phrase break problem into a two-step prediction problem. The minor break can be predicted with the usual single-step prediction.

Table 2 shows the algorithm of our phrase break prediction. The second-order features are weighted by the factor w. For example, if there is a non-major break in the previous 3 syllables, the weight factor w = 0.014+0.07+0.118, where 0.014, 0.07 and 0.118 are occurrence ratio for the number of syllables 1, 2 and 3 between major breaks respectively (see Figure 1).

Table 2: Our phrase break prediction algorithm.

Parameters:
$\boldsymbol{x}$ = feature sequence {POS, lexicon, distance, …}
$\boldsymbol{y}$ = break sequence {M (major break), m (minor break), N (non-break)}
k = cross validation parameter
l = second-order feature {(number of syllables between current juncture and previous M) × w, (number of syllables between current juncture and next M) × w}
w = the accumulated occurrence ratio for the number of syllables between M's
ME = base learner based on maximum entropy principle

Learning procedure:
Given a corpus $C = \{(\boldsymbol{x}, \boldsymbol{y})\}$
1. Change 3 class data into 2 class data:
   Corpus C_M = {Ms, ms, Ns} ➔ {Ms, non-Ms}
   Corpus C_m = {Ms, ms, Ns} ➔ {ms, non-ms}
2. Split 2-class corpus C_M into k equal-sized : $\{C_{M1}, \ldots, C_{Mk}\}$
3. For the second-order feature l, k-fold cross-validate C_M
4. Using the second-order feature l, we get new corpus $C_M' = \{(\boldsymbol{x}', \boldsymbol{y})\}$ where $\boldsymbol{x}' = (\boldsymbol{x}, l)$
5. Train three ME model

majorME$_1$: maximum entropy model trained with C_M
majorME$_2$: maximum entropy model trained with C_M'
minorME : maximum entropy model trained with Cm

Inference procedure:
Two-step major break inference
1. Prepare the second-order feature l by predicting the major break with model **majorME$_1$**.
2. Using second-order feature l, inference the major break with model **majorME$_2$**.

Single-step minor break inference
The **minorME** model is used to inference whether the non-major breaks are minor breaks or not.

Figure 2 shows a graphical representation of major break prediction integrated with the second-order information.

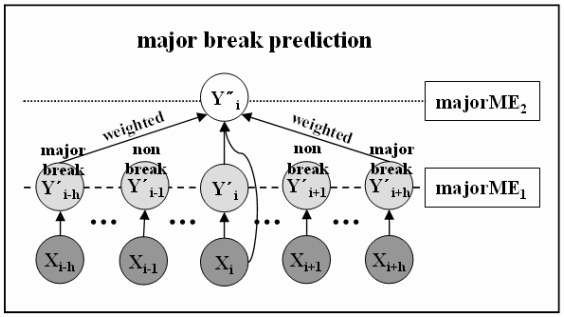

Figure 2: The two-step major break prediction.

3.2 Machine Learning Features

In our research, we used the previously developed Korean POS tagger [10] and the Korean dependency parser [11]. Using the linguistic analysis systems, we extracted various features, and tried the manifold combinations of the feature categories to obtain the optimal features. These features are as follow:

majorME$_1$ model and **minorME** model

Local features

- Part-of-speech tag: POS tag feature includes the previous five and the next four tags.
- Lexical eojeol: Lexical eojeol feature includes the previous lexical eojeol.
- Eojeol length: In Korean speech, the syllable is the basic pronunciation unit. Therefore, for an n-syllable eojeol, the eojeol length n is used to obtain some of the corresponding prosodic information. Eojeol length feature includes previous two and next two eojeol lengths.
- Distance: Distance feature is the distance in syllables from the current position to the beginning and end of the sentence. The distance feature is normalized by using the sentence length.

Global features

- Global syntactic chunk: Syntactic feature includes previous three terminal, previous three pre-terminal, next two terminal and next two pre-terminal chunk labels.

majorME$_2$ model

- Include **majorME$_1$** model's all features.

Second-order features

- Second-order information: Second-order feature includes the information about the previous major break and the next major break.

4. EXPERIMENTAL RESULTS

4.1 Experimental setup

For more extensive comparisons, we used the F-score to measure performance. The F-score is the harmonic mean of precision and recall. We divided the SynthFemale01 corpus into 10 parts and used the 10-fold cross validation. To confirm the validity of the method we proposed, we have performed the experiments using the following three methods.

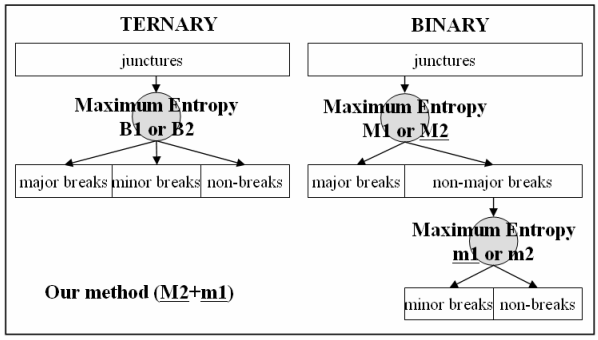

Figure 3: Ternary, Binary and our method for the experimental setup.

Ternary decision making method: Major break, minor break and non-break are predicted concurrently.

- baseline ME (B1): The phrase breaks are predicted using the basic ME. We used this method as the baseline method.
- baseline ME using second-order information (B2): The phrase breaks are predicted using local, global and second-order features.

Binary decision making method: Minor breaks are predicted after predicting major breaks.

- single-step major break prediction (M1) + single-step minor break prediction (m1): The phrase breaks are predicted using local and global features.
- two-step major break prediction (M2) + single-step minor break prediction (m1): This method is our method that is described in section 3.1. Major breaks are predicted using local, global and second-order features, and minor breaks are predicted using local and global features.
- two-step major break prediction (M2) + two-step minor break prediction (m2): The phrase breaks are predicted using local, global and second-order features.

Other method:

- CRF: The phase breaks are predicted using first order linear chain CRF model, which is a learner specifically designed for sequential labeling [7] without second-order information.

4.2 Experimental results and analyses

We achieved the performance of 75.4 in terms of F-score when applying the baseline ME (B1) method as shown in Table 3. The result of the baseline ME method of simply using the second-order information (B2) showed a slightly worse performance than the baseline ME method. However, our proposed method (M2+m1) outperformed all the other methods. For example, our method improved 4.2% in major break prediction and 2.1% in minor break prediction compared with the baseline ME (B1) method. Moreover, the result indicated that our method has achieved a better performance than even CRF method which is learner specifically designed for a sequential labeling. The execution time of our method is much shorter than that of CRF (about 4 times faster).

Table 3: Performance comparison of various methods

	major break	minor break	total break
Ternary Decision Making Method			
B1	75.4 (0.0)	73.0 (0.0)	86.6 (0.0)
B2	74.7 (-0.7)	72.9 (-0.1)	86.0 (-0.6)
Binary Decision Making Method			
M1 + m1	75.5 (+0.1)	72.7 (-0.3)	86.5 (-0.1)
M2 + m1	79.6 (+4.2)	75.1 (+2.1)	90.7 (+4.1)
M2 + m2	79.6 (+4.2)	73.9 (+0.9)	88.5 (+1.9)
Other Method			
CRF	76.1 (+0.7)	72.6 (-0.4)	86.8 (+0.2)

Our major and minor breaks well correspond to IP and AP boundaries respectively. Thus our phrase break prediction model can be compared to the related works on prosodic phrasing model. Table 4 shows the comparison results between our model and the previously related works.

Table 4: Comparison between our method and the related works on IP and AP boundary

Prosody boundary	Our method	Yoon [2]	Kwon [12]	Kim [13]
IP	79.6	71.2	66.9	75.2
AP	75.1	72.8	87.1	48.4
IP + AP	90.7	88.0	80.4	78.1

In Table 4, the corpora used are all different to each other, so direct comparison is meaningless. However the corpora in the related works are mostly domain limited or small sized, whereas our corpus is balanced for the tri-phone and the genre, has a bigger size (has more eojeols) and is specially built for TTS systems. So, these situations can reveal the superiority of our works in Table 4. Compared to the related works in F-scores, our method is superior to the others for IP boundary prediction task. For AP boundary prediction task, our method may not be the best, but still relatively performed well.

5. CONCLUSIONS

The previous TTS systems have some difficulties in using the second-order information such as the distance between the current juncture and the previous/next phrase break due to the lack of efficient computational algorithms. This paper presented a new efficient phrase break prediction method which integrates the second-order information with the general maximum entropy model. Our method employs a modified stacked sequential learning method which is well designed to use this kind of second-order information. As shown in the experimental results, our two-step architecture for major break prediction is much more effective than the methods which only use local and global information.

In future work, we will apply the useful syntactic parse trees to generate other useful second-order information with more improvements in the syntactic parser. We will also develop the prosody model by assigning IP and AP boundary tones to the predicted major and minor breaks.

ACKNOWLEDGEMENTS

We would like to thank Minwoo Jeong and Sangkeun Jung for their helpful discussion. This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment) (IITA-2005-(C1090-0501-0018)).

6. REFERENCES

[1] Taylor, P. and Black, A. W. "Assigning Phrase Breaks from Part-of-speech Sequences," *Computer Speech and Language*, 12(2):99-117, 1998.

[2] Yoon, K. "A Prosodic Phrasing Model for a Korean Text-to-speech Synthesis System," *Computer Speech and Language*, 20(1):69-79, 2006.

[3] Zheng, Y., Kim, B. and Lee, G. G., "Using Multiple Linguistic Features for Mandarin Phrase Break Prediction in Maximum-entropy Classification Framework," *In Proceedings of the 8th International Conference on Spoken Language Processing*, 2004.

[4] Zervas, P., Maragoudakis, M., Fakotakis, N. and Kokkinakis, G, "Bayesian Induction of Intonational Phrase Breaks," *Eurospeech '03,* 113-116, 2003.

[5] Ostendorf, M. and Veilleux, N. "A Hierarchical Stochastic Model for Automatic Predictions of Prosodic Boundary Location," *Computational Linguistics*, 20(1):27-54, 1994.

[6] Jun, S., K-ToBI (Korean ToBI) labeling conventions (version 3.1), http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html, 2000

[7] Lafferty, J., McCallum, A. and Pereira, F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *In Proceedings of the International Conference on Machine Learning (ICML-2001)*, 282-289, 2001.

[8] McCallum, A., Freitag, D. and Pereira, F., "Maximum Entropy Markov Models for Information Extraction and Segmentation," *In Proceedings of the International Conference on Machine Learning (ICML-2000)*, 591-598, 2000.

[9] Cohen, W. W. and Carvalho, R. V. "Stacked Sequential Learning," *In Proceeding of IJCAI*, 671-676, 2005.

[10] Lee, G. G. and Cha, J. "Syllable Pattern-based Unknown Morpheme Segmentation and Estimation for Hybrid Part-of-speech Tagging of Korean," *Computational Linguistics*, 28(1):53-70, 2002.

[11] Eun, J., Jeong, M., and Lee, G. G., "Korean Dependency Structure Analyzer based on Probabilistic Chart Parsing," *In Proceeding of the 17th Conference on Hangul and Korean Information Processing,* 105-111, 2005.

[12] Kwon, O., Hong, M., Kang, S. and Shin, J., "AP, IP Prediction for Corpus-based Korean Text-to-speech," *Journal of Speech Sciences,* 9(3):25-34, 2002.

[13] Kim, Y., Byeon, H. and Oh, Y., "Prosodic Phrasing in Korean; Determine Governor, and then Split or Not," *Eurospeech99,* 539-542, 1999.