LETTER

A Variable Break Prediction Method Using CART in a Japanese Text-to-Speech System

Deok-Su NA[†a)], ***Member and*** **Myung-Jin BAE**[††], ***Nonmember***

SUMMARY Break prediction is an important step in text-to-speech systems as break indices (BIs) have a great influence on how to correctly represent prosodic phrase boundaries. However, an accurate prediction is difficult since BIs are often chosen according to the meaning of a sentence or the reading style of the speaker. In Japanese, the prediction of an accentual phrase boundary (APB) and major phrase boundary (MPB) is particularly difficult. Thus, this paper presents a method to complement the prediction errors of an APB and MPB. First, we define a subtle BI in which it is difficult to decide between an APB and MPB clearly as a variable break (VB), and an explicit BI as a fixed break (FB). The VB is chosen using the classification and regression tree, and multiple prosodic targets in relation to the pith and duration are then generated. Finally, unit-selection is conducted using multiple prosodic targets. The experimental results show that the proposed method improves the naturalness of synthesized speech.

key words: *text-to-speech system, break prediction, variable break*

1. Introduction

Break indices (BIs) represent the structure of prosodic phrases in a text-to-speech (TTS) system, and rule-based and stochastic approaches to break prediction have been used. The rule-based approach is predicated upon the correlation between prosodic boundaries and syntactic structure, and requires sets of heuristic rules written by linguistic experts. The stochastic approach to break prediction, based on decision trees or Markov models, requires labeled training data, but is less dependent on the art of heuristic rule writing. There have been several studies of stochastic approaches for TTS systems that employ algorithms such as classification and regression trees (CART), Markov models, and so on [1]. However, their accuracy is not high enough due to the different reading styles of the speakers. In Campbell's research in 1996, the rate of congruence between BIs predicted automatically and BIs labeled manually was 69%. However, it could be increased to 90% when predicted BIs were modified to within ±1 [2]. Such a result shows the uncertainty among BIs that makes break prediction difficult. However, if such characteristics are used in a synthesizer, it is possible to generate more naturally synthesized speech. In order to use such characteristics, we define the variable break (VB) and use it to propose a unit-selection method.

Manuscript received August 19, 2008.

[†]The author is with Voiceware Co. Ltd., Sungsu-dong 2-ga 280–13, Sungdong-gu, Seoul, 133–120 Korea.

[††]The author is with the Department of Information and Telecommunication Engineering, Soongsil University, Korea.

a) E-mail: dsna@voiceware.co.kr

DOI: 10.1587/transinf.E92.D.349

2. Break Index and Variable Break

Table 1 shows the BIs used in a TTS system, which are generated through a simple rule-based method. BIs 0 and 1 can be generated by a phoneme sequence, which is the output of the text processing and word separation. BIs 2 and 3 can be decided upon according to the accent and structure of the sentence. And 4 and 5 can be distinguished by a comma (、) and a period (。). Among the BIs, it is easy to distinguish 2 and 3 with the other BIs (0, 1, 4, and 5). However, 2 and 3 are very similar in that complicated rules are required in order to decide between the two.

BI 2 is one in which a pause does not exist in spoken Japanese while the accent type is changed. On the other hand, BI 3 is one in which a short pause appears while the accent type is changed. In the corpus of BIs labeled manually, there are cases in which BI 2 and 3 are not distinguished according to the context. Furthermore, the labeled BIs can be different for the same context. However, the break generation module of a synthesizer only assigns one BI in a word boundary, and it is used in the unit-selection process. Thus, speech segments with similar BIs are not to be used for candidate-units together, and the optimal speech segment for more natural synthesized speech cannot be searched. To overcome such a problem, we have defined a subtle BI that is difficult to decide upon between APB and MPB clearly as a variable break (VB), and an explicit BI as a fixed break (FB), and we distinguish between the FB and VB if the generated BI is 2 or 3. If it is a VB, we generate multiple prosodic targets, which are used to conduct unit-selection. In the proposed unit-selection, added candidate-units, which include units having the generated BI (BI of 2 or 3) and other units having a similar BI (BI 2 or 3), are used.

2.1 Variable Break Prediction Using CART

The VB is predicted using a decision tree and decision rule.

Table 1 Break indices.

Break Indices	Prosodic Boundary
0	No prosodic break
1	Word boundary
2	Accentual phrase boundary
3	Major phrase boundary
4	Intonational phrase boundary
5	Sentence boundary

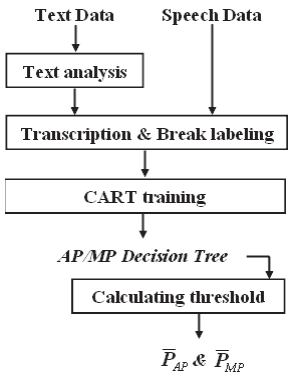


Fig. 1 Modeling of variable break.

Table 2 Features used for CART modeling.

No	Factors
1	Number of mora in ($W_{k-2}, W_{k-1}, W_k, W_{k+1}, W_{k+2}$)
2	Number of mora before/after W_k within IP
3	Part-of-speech types of ($W_{k-2}, W_{k-1}, W_k, W_{k+1}, W_{k+2}$)
4	Tone pattern of consecutive five moras before M_n
5	Tone pattern of consecutive five moras after M_n
6	Phoneme of moras in ($W_{k-2}, W_{k-1}, W_k, W_{k+1}, W_{k+2}$)
7	Kind of morphs ($W_{k-2}, W_{k-1}, W_k, W_{k+1}, W_{k+2}$)
8	BIs of (W_{k-2}, W_{k-1})

In order to distinguish between BI 2 and 3, the AP/MP decision tree is trained using CART. The output of the decision tree is represented by a probability value. The value of $\hat{P}_i$ represents the output-form of the AP/MP decision tree. It consists the probability that it will be an APB, $P_{AP}(i)$, and the probability that it will be an MPB, $P_{MP}(i)$.

$$\hat{P}_i = [P_{AP}(i), P_{MP}(i)] \tag{1}$$

$$P_{AP}(i) = 1 - P_{MP}(i) \tag{2}$$

Figure 1 shows the modeling of variable break. The VB modeling consists of training the AP/MP decision tree and calculating the thresholds ($\bar{P}_{AP}$ & $\bar{P}_{MP}$) that will be used in the decision making process. For the CART training and testing, 8,586 sentences and manually labeled data were used. 5,769 sentences were used for training, and 2,718 sentences were used for testing. One professional conducted the BI labeling by listening to a recorded speech. Table 2 shows the features used in the CART training. Term W_k is the kth word in a sentence, and M_n is the last mora of W_k. The 7th feature concerns the morph types. Like Table 3, morphs types classified by a similar grammatical function were used. For example, words of structural nouns and auxiliary verbs, which are strongly connected to the preceding word, were classified. The 8th feature is the BIs of the preceding words. The trained tree has 81 terminal nodes.

Table 4 shows the test results of the AP/MP decision tree. For the APB, it is decided that if $P_{AP}(i) > P_{MP}(i)$, it is TRUE, and in other cases, it is FALSE. For the MPB, it is also decided that if $P_{MP}(i) > P_{AP}(i)$, it is TRUE,

Table 3 Kinds of morphs for 7th feature.

Index	Morphs
0	**structural nouns** (ex : こと, ため, 上, 内, etc.)
1	**auxiliary verbs** (ex : した, た, ます, etc.)
2	**auxiliary adjectives** (ex :ない, なく, etc.)
3	**time nouns** (ex :一昨日, 昨日, 前, 午後, etc.)
• n	(ex : お祈り, お願い, お電話, お知らせ, etc.)

Table 4 Performance of AP/MP decision tree.

	Precision	Recall	F1-Score
AP	90.92%	94.46%	92.63%
MP	79.42%	69.28%	74.00%

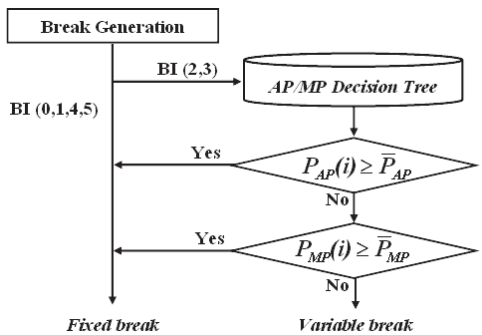


Fig. 2 Decision of variable break.

Table 5 Rate of variable break.

Target \ Prediction	AP		MP	
	FB	VB	FB	VB
AP	70.64%	29.26% (a)	32.55%	67.45% (b)
MP	19.22%	80.88% (c)	60.05%	39.95% (d)

while in other cases it is FALSE. In order to evaluate the AP/MP decision tree we measured the precision, recall, and F1-Score. The overall accuracy was 88.52%. The thresholds used in the decision process, $\bar{P}_{AP}$ and $\bar{P}_{MP}$, are as follows.

$$\delta_{AP}(i) = \begin{cases} 1 & i \in AP,\ P_{AP}(i) > P_{MP}(i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\bar{P}_{AP} = \frac{1}{N_{AP}} \sum_{i}^{N_{AP}} P_{AP}(i)\delta_{AP}(i) \tag{4}$$

$$\delta_{MP}(i) = \begin{cases} 1 & i \in MP,\ P_{MP}(i) > P_{AP}(i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\bar{P}_{MP} = \frac{1}{N_{MP}} \sum_{i}^{N_{MP}} P_{MP}(i)\delta_{MP}(i) \tag{6}$$

Term $\bar{P}_{AP}$ is the expectation of $P_{AP}(i)$ when the prediction of an APB is accurate, while $\bar{P}_{MP}$ is the expectation of $P_{MP}(i)$ when the prediction of an MPB is accurate. Terms N_{AP} and N_{MP} are the numbers of the APBs and MPBs. The value of $\bar{P}_{AP}$ was 0.9245 and $\bar{P}_{MP}$ was 0.8085.

2.2 Decision of Variable Break

The generated BIs 0, 1, 4, and 5 become an FB, and the VB is chosen only for BIs 2 and 3. The VB is decided upon as in Fig. 2 using the AP/MP decision tree and threshold. If the generated BI is 2 and 3, the prediction probability is calculated by the AP/MP decision tree. Then, if it is higher than the threshold, it becomes an FB, and if it is lower, it becomes a VB. When this is not the case (when it is an FB), $P_{AP}(i)$ and $P_{MP}(i)$ are adjusted to 1 or 0 as in Eq. (7).

$$\tilde{P}_i = \begin{cases} [1,0] & P_{AP}(i) \geq \bar{P}_{AP} \\ [0,1] & P_{MP}(i) \geq \bar{P}_{MP} \\ \hat{P}_i & \text{otherwise} \end{cases} \quad (7)$$

Table 5 shows the VB rates in the test results of the AP/MP decision tree. The VB rates of inaccurate prediction, (b) and (c), are higher than the VB rates of accurate prediction, (a) and (d). This means that the prediction is more likely to be decided as a VB when a BI generation error occurs within the synthesizer, and it is possible to complement the BI generation error through the proposed unit-selection process.

3. Multiple Prosodic Targets

In the synthesizer, the prosodic target (PT) consisting of BI, phoneme duration, and pith contour is generated for unit-selection. In a conventional synthesizer, a single PT was used because the generated BI is not changed until the end of the unit-selection. However, if the proposed VB is used, the generated BI can be changed in the unit-selection results. Thus, multiple prosodic targets are needed for the proposed unit-selection. Figure 3 describes the proposed prosody generation. The VB prediction is conducted after break generation, and multiple pitch and duration generation are conducted by using the VB information.

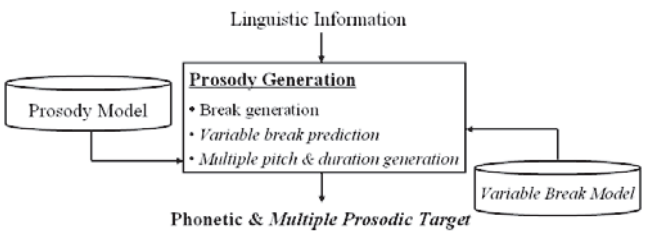
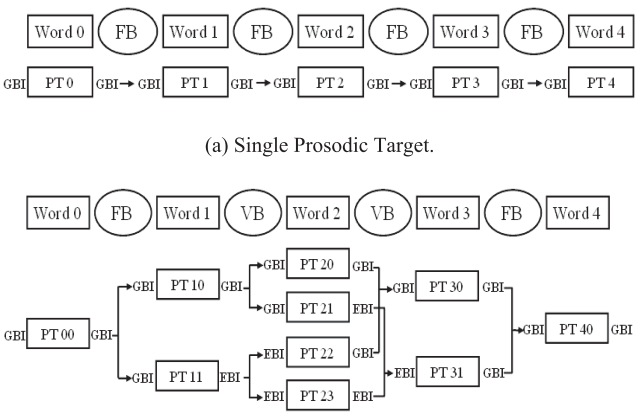


Fig. 3 Proposed prosody generation.



(a) Single Prosodic Target.

(b) Multiple Prosodic Targets.

Fig. 4 Single and multiple prosodic targets. (PT - Prosodic Target, GBI - Generated BI, EBI - Expanded BI)

The PT of one sentence is the connected form of the sub-PTs of words, and the sub-PTs are generated according to the words' left BIs and right BIs. Figure 4 describes a single PT and multiple PTs. In general, if the type of FB is used in a conventional synthesizer, only one sub-PT is needed for a word such as in Fig. 4 (a). However, if the VB is used as in Fig. 4 (b), multiple sub-PTs are needed for a word because the left and right BIs of each word can become a generated BI (GBI) and an expanded BI (EBI). Thus, at most, 4 kinds of sub-PTs are needed for a word. If the GBI is 2 and a VB, the EBI becomes BI 3. And if the GBI is 3 and a VB, the EBI is 2.

4. Unit Selection by Using Variable Break

The unit selection process itself is done using a dynamic programming (Viterbi) algorithm, and it uses target-costs and join-costs [3]. Generally, the quality of synthesized speech can be improved if the number of candidate-units is increased during the unit-selection process. To generate naturally synthesized speech, various prosodic patterns must be included in the corpus and unit-selection inventory, which is the entry of candidate-units.

Figure 5 shows the proposed unit-selection. In the proposed method, the number of candidate-units is increased by expanding the BI for a VB. Not only units with a GBI but also units with an EBI are included in the unit-selection inventory. If the number of candidate-units is increased over the reasonable threshold (N), pre-selection is done [4]. The rate of pre-selection for a GBI and EBI is chosen using Eq. (7). For example, if the GBI is 2 and $P_{AP}(i)$ is 0.6, the pre-selection is conducted to create a unit-selection inven-

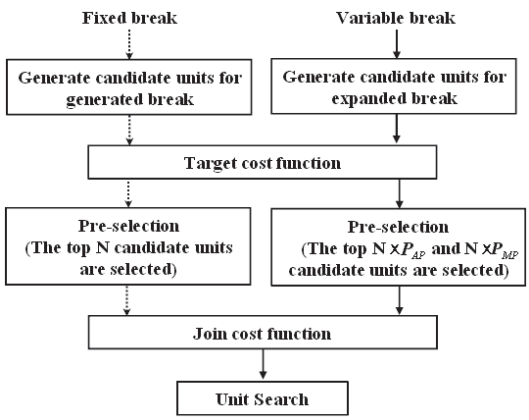


Fig. 5 Proposed unit-selection.

tory in which BI 2 is 60% and BI 3 is 40%. Equation (8) shows the target cost function. Term TC^p is the prosodic-target cost and TC^c is the context-target cost. In Eq. (9), k represents the sub-PT. We calculate the costs using sub-PTs and then select the minimum value.

$$TC(t_i, u_i) = TC^p(t_i, u_i) + TC^c(t_i, u_i) \quad (8)$$

$$TC^p(t_i, u_i) = \arg\min_k TC_k^p\left(t_i^k, u_i\right) \quad (9)$$

5. Experimental Results

A single-speaker speech corpus was collected. The gender and size are female and 41-hours in length. The contents of the corpus include news, novels, and conversations. The sizes in hours do not include silences at initial and final utterances, but include medial utterance pauses. The speaker was professional narrator. The speech was recorded in a sound-proof room. A mean opinion score (MOS) was used to evaluate the sound quality of the synthesized speech. An opinion score was set on a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Five Japanese females participated and, 127 sentences from the JEITA Overall Evaluation Sentences [5] were selected. 127 original speeches and 254 synthesized speeches generated by system 1, which used a VB, and system 2, which did not use a VB, were mixed in order to be listened to in irregular order. The results are shown in Fig. 6. The original speech scored a 4.99, while system 1 scored a 4.25 and system 2 a 4.01.

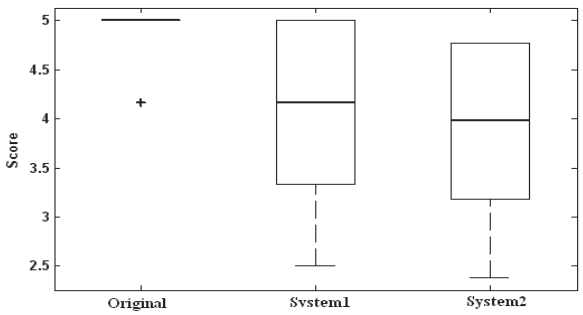


Fig. 6 Results of MOS test.

6. Conclusion

The naturalness of synthesized speech is decided upon by the prosody used. Thus, realizing a natural prosody is the common goal of every synthesizer. Since a large corpus-based TTS system already has various prosodic patterns in the speech corpus, it is possible to realize a natural prosody if the patterns are efficiently used. However, it is not easy to generate natural synthesized speech by unit-selection using a restricted prosodic target. In this paper, we have proposed a method of using the generated prosody more efficiently. By introducing a variable break, we could complement the restricted break prediction and use various prosodic patterns of a speech corpus in the unit-selection process.

References

[1] X. Sun and T.H. Applebaum, "Intonational phrase break prediction using decision tree and N-gram model," Proc. EUROSPEECH2001, vol.1, pp.537–540, 2001.
[2] J. Venditti, "Japanese ToBI labeling guidelines," OSU Working Papers in Linguistics, pp.127–162, 1997.
[3] A. Conkie, M.C. Beutnagel, A.K. Syrdal, and P.E. Brown, "Preselection of candidate units in a unit selection-based text-to-speech synthesis system," Proc. ICSLP2000, vol.3, pp.314–317, 2000.
[4] D.S. Na, S.Y. Min, K.H. Lee, J.S. Lee, and M.J. Bae, "A pre-selection of candidate units using accentual characteristic in a unit selection based Japanese TTS system," J. Acoustical Society of Korea, vol.26, no.4, pp.159–165, 2007.
[5] Technical Standardization Committee on Speech Input/Output Systems, "Speech synthesis system performance evaluation methods," JEITA IT-4001, pp.42–45, 2003.