한국음향학회지 제28권 제2호 pp. 155~163 (2009)

가변 Break를 이용한 코퍼스 기반 일본어 음성 합성기의 성능 향상 방법

A Performance Improvement Method using Variable Break in Corpus Based Japanese Text-to-Speech System

나 덕 수*, 민 소 연**, 이 종 석*, 배 명 진***
(Deok-Su Na*, So-Yeon Min**, Jong-Seok Lee*, Myung-Jin Bae***)

*보이스웨어 기술연구소, **서일대학 정보통신과, ***숭실대학교 정보통신 전자공학부
(접수일자: 2008년 12월 11일; 채택일자: 2009년 1월 23일)

Text-to-speech 시스템에서 입력 텍스트로 부터 운율 정보를 생성하기 위해서는 운율구 경계, 음소 지속시간, 기본 주파수 포락선 설정의 3가지 기본적인 모듈이 필요하다 [1]. Break 인덱스 (BI; Break Index)는 합성기에서 운율구의 경계를 나타내고, 자연스러운 합성음을 생성하기 위해서는 BI를 정확히 예측하여야 한다. 그러나 BI는 문장의 의미나 화자의 읽기 습관 (reading style)에 따라 임의적으로 결정되는 경우가 많아 정확한 예측이 매우 어렵다. 특히 일본어 합성기에서는 악센트구 경계 (APB; Accentual Phrase Boundary)와 major phrase 경계 (MPB; Major Phrase Boundary)의 정확한 예측이 어렵다. 따라서 본 논문에서는 APB와 MPB 예측 오류를 보완할 수 있는 방법을 제안한다. BI를 고정 break (FB; Fixed Break)와 가변 break (VB; Variable Break)로 분류하여 합성단위 선택을 수행한다. 일반적으로 BI는 한번 생성되면 변하지 않는다. 따라서 BI가 잘못 생성된 경우 최적의 합성음을 생성할 수 없게 되는데, VB는 생성된 BI와 그것과 유사한 BI를 함께 이용하여 합성단위 선택을 수행함으로써 합성음의 BI가 생성된 BI와 다를 수 있는 것을 의미한다. APB와 MPB에 해당하는 BI에 대하여 VB인지 FB인지 CART (Classification and Regression Tree)를 이용하여 예측하고, VB인 경우 기본 주파수와 음소 지속시간에 대해 다중 운율 모델을 생성하여 합성단위 선택을 수행하였다. MOS 테스트 결과 원음이 4.99, 제안한 방법은 4.25, 기존의 방법은 4.01로 합성음의 자연성을 향상 시킬 수 있었다.

핵심용어: 일본어 음성합성, Break 예측, 가변 break
투고분야: 음성 처리 분야 (2.4)

In text-to-speech systems, the conversion of text into prosodic parameters is necessarily composed of three steps. These are the placement of prosodic boundaries, the determination of segmental durations, and the specification of fundamental frequency contours [1]. Prosodic boundaries, as the most important and basic parameter, affect the estimation of durations and fundamental frequency. 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. In the MOS test result, the original speech scored a 4.99, while proposed method scored a 4.25 and conventional method scored a 4.01. The experimental results show that the proposed method improves the naturalness of synthesized speech.

Keywords: Text-to-Speech system, Break prediction and variable break
ASK subject classification: Speech Signal Processing (2.4)

책임저자: 나 덕 수 (dsna@voiceware.co.kr)
133-120 서울시 성동구 성수동 2가 280-13 삼환디지털벤처타워 10층 (주)보이스웨어
(전화: 02-3016-8562; 팩스: 02-3016-8501)

I. 서 론

운율구의 구조를 나타내기 위해 break indices (BI)를 사용하는데, BI를 생성하는 방법은 규칙 기반 방법과 코퍼스 기반 방법이 있다. 규칙 기반 방법은 문장기호, 품사, 발음열 (phoneme stream) 등의 정보와 언어학적인 정보를 이용한다. 정확한 BI를 얻기 위해서는 매우 복잡하고 정교한 작업이 필요하다. 코퍼스 기반 방법에는 여러 가지 특징들을 이용하여 자동으로 decision tree를 구축하는 classification and regression trees (CART) 방법과, hidden Markov model (HMM)을 이용하는 방법이 있다 [2]. 하지만 개인마다 조금씩 다른 읽기습관 (reading style)에 의해 모델링의 정확도가 높지 못하다. 1996년 Cambell의 연구를 살펴보면, 자동으로 예측한 BI와 사람이 레이블링한 BI가 일치하는 정확도는 69% 정도이다. 그러나 예측 값을 +/- 1로 조정한다면 90%로 올라가는 것을 알 수 있다 [3-4]. 이러한 결과는 break index 사이의 불명확성 (uncertainty, subtleness)을 나타내는 것으로 BI 예측을 어렵게 하는 요인이다. 그러나 합성기에서 이러한 특징을 이용한다면 보다 자연스러운 합성음을 얻을 수 있다.

II. 일본어의 운율 구조와 가변 Break

일본어의 운율 구조에서, 하나의 문장은 몇 개의 IP로 이루어지고, IP는 몇 개의 MP (Major Phrase, intermediate Phrase)로 구성된다. 그리고 몇 개의 AP는 MP를 구성하고, AP는 몇 개의 단어, 마지막으로 음소 (phone) 또는 음절 (syllable, mora)로 이루어진다. BI는 이러한 운율 구조를 표현하기 위한 것으로 J-ToBI (Japanese Tone and Break Indices)에서는 4개의 레벨을 사용하고 있다. 그러나 운율구 경계의 모호함 또는 다양성에 의해 4개 이상의 레벨로 표현하기도 한다 [5-6]. 표 1은 J-ToBI에서 사용하는 4개의 BI의 정의를 나타내고, 표 2는 본 논문에서 사용한 6개의 BI에 대한 정의를 나타낸 것이다. 본 논문에서 사용한 BI 0과 1은 J-ToBI의 0과 1과 같고, 나머지 BI 2~5는 J-ToBI 2와 3을 세분화한 것이다.

일본어는 억양구마다 피치의 변화 범위 (pitch range)가 달라지고, 억양구의 끝에서는 피치가 변하는 몇 가지 패턴이 나타나는데 이러한 것을 BPMs (boundary pitch movements) [4][6]라고 한다. 본 논문에서의 BI 2는 AP 경계이지만 BPM이 나타나지 않고, 바로 이어서 포즈가 오지 않으며 AP의 피치 변화 범위도 이어지는 AP에 영향을 받는 것을 의미하고, BI 3은 BPM이 나타나고, 포즈가 이어지지만 AP의 피치 변화 범위가 이어지는 AP에 영향을 받는 것을 의미한다. BI 4는 BPM이 나타나고 포즈도 이어지며 AP의 피치 변화 범위도 이어지는 AP와 상관성이 없는 것을 의미하고, BI 5는 문장의 끝을 나타낸다. 본 논문에서는 텍스트 분석 정보를 이용하여 BIs을 자동 생성하고, 자동 생성된 BIs를 위와 같은 특징을 고려하여 전문가가 수정함으로써 운율 코퍼스를 구축하였다.

텍스트 분석 정보를 이용하여 BI를 자동으로 생성하는 경우, BI 0과 1은 텍스트 전처리의 결과인 음소열과 단어 분리 결과를 이용하여 생성하였고, BI 2와 3은 악센트와 문장의 구조에 따라 결정하였다. 4와 5는 쉼표 (comma)와 마침표 (period)로 구분하였다. 텍스트 분석을 통해 BI를 생성할 때 BI 2와 3을 다른 BI (0,1,4,5)와 구분하는 것은 쉬우나 서로 유사하여 둘 중 하나를 결정하기 위해서는 복잡한 규칙이 필요하다.

연속 음성에서는 표 2와 같이 BPMs와 포즈의 유무에 따라 BI 2와 BI 3을 구분할 수 있으나, BI를 레이블링한 코퍼스를 분석해 보면 문맥정보 (context)에 따라 BI 2와 3이 구분되는 경우도 있지만 그렇지 않은 경우도 많다.

표 1. J-ToBI의 Break Index
Table 1. Break indices of J-ToBI.

	4 degrees of BIs
0	Strong cohesion. Typical of fast speech or AP-medial lenition processes [4].
1	No higher-level boundary. Typical of the majority of AP-medial word boundaries [4].
2	Medium disjuncture. Typically corresponds to the tonally -defined accentual phrase boundary [4].
3	Strong disjuncture. Typically corresponds to the tonally -defined intonation phrase boundary [4].

표 2. 본 논문에서 사용한 Break Index
Table 2. Break Indices used in this paper.

	6 degrees of BIs
0	No prosodic break : same as the J_ToBI usage
1	Prosodic word boundary (WB) : same as the J_ToBI usage
2	Accentual phrase or minor tone group boundary (AP) : No BPM at the end & No followed by a pause & Without pitch range resetting
3	Major phrase (intermediate phrase) boundary (MP) : BPM at the end & Followed by a pause & Without pitch range resetting
4	Intonation phrase boundary (IP) : BPM at the end & Followed by a pause & pitch range resetting
5	Sentence boundary (SB)

심지어 같은 문맥정보 이지만 BI가 2인 데이터와 3인 데이터의 비율이 50:50인 경우도 있다. 그런데 합성기의 일반적인 break 생성 모듈에서는 하나의 경계에 한 가지의 BI만 생성하고 이것을 이용하여 합성단위 선택을 수행한다. 따라서 유사한 BI를 가지는 데이터는 사용하지 못하여 최적의 합성 음질을 얻을 수 없게 된다. 본 논문에서는 이러한 문제를 해결하기 위해 break 생성 모듈에서 생성된 BI가 2와 3인 경우 고정 break (FB; fixed break)와 가변 break (VB; variable break)로 구분하였다. FB인 경우 break 생성 모듈에서 결정된 BI (BI 2 또는 3)를 가지는 후보 합성단위 (candidate-units)만으로 합성단위 선택을 수행하고, VB인 경우 BI가 2와 3인 후보 합성단위를 모두 포함하여 합성단위 선택을 수행한다.

표 3. 가변 Break 예측 규칙 예
Table 3. Examples of variable break prediction rule.

조사 と	(と)+は+おもう (~라고 생각한다)의 경우		break 2 고정
	(と)+は+裏腹に, 裏腹で (~와는 정반대로)		
	と+は+형용사		break 3 가변
	그 외의 경우		
조사 が	ところが (~했는데, ~했더니)		break 3 고정
	조사 が가 2개 이상일 때 앞의 것		
	문장에 は 없이 が가 주격 조사일 때		
	그 외의 경우		break 2 가변
그 외 조사의 기본처리	조사 뒤에 설명 구조	조사+명사+동사	break 2 가변
		조사+동사+동사, 조동사, 보조용언	
		조사+형용사 보조동사	
	조사 뒤에 수식 구조	조사+특정 부사 (また, ...)	break 3 고정

III. 가변 break 예측

대용량 음성 코퍼스를 사용하는 경우 합성단위 선택 과정에서 목표 (target) break 뿐만 아니라 유사한 break들을 모두 이용한다면 음성 코퍼스에 포함된 다양한 break 정보와 부족한 규칙에 대한 보완을 효율적으로 수행 할 수 있고, 다양한 문맥정보를 가지는 합성단위들을 후보로 추출하여 합성음의 음질을 향상 시킬 수 있다. 본 논문에서는 먼저 규칙에 의한 가변 break의 예측이 가능한지 알아보기 위해 몇 가지 규칙들을 조사하였다. 그리고 이를 이용하여 통계적 모델링 방법인 CART [7]를 이용하여 가변 break를 예측하였다.

3.1. Break 예측 규칙

녹음된 데이터와 대본을 분석해보면, 명사와 조사, 명사와 접미사 또는 복합명사를 만드는 명사와 명사처럼 결합되면서 악센트에 영향을 주는 단어들 사이의 break는 기본적으로 고정 break일 확률이 높고, 가변 break는 의미론적 또는 형태론적인 특징으로 인해 대부분 정해진다. 따라서 가변 break 예측을 위한 규칙은 문법적 지식에 근거한 규칙과 실험적인 규칙으로 구성된다.

표 3은 가변 break예측을 위한 규칙의 예를 나타낸 것으로 언어처리 모듈에서의 결과인 품사 정보와 문자 정보 등을 이용하는 규칙들이다. 이러한 규칙의 대부분은 AP 경계에 해당하는 break 2, 3에 대하여 예측되는 것으로 가변 break를 예측하는 규칙은 인접한 품사의 종류를 검사하는 것과 같이 일반적인 규칙들로 이루어지고, 고정 break를 예측되는 규칙은 관용적 표현과 같이 특정한 단어의 조합 형태인지 검색하는 제한적인 규칙이 많다. 이러한 규칙은 일본어 문법과 그 예문을 분석하여 만들어졌다.

3.2. CART를 이용한 가변 break 모델링

규칙을 이용하여 생성된 BI 2와 3의 에러를 보완하기 위해 코퍼스를 이용하여 VB를 예측하였다. 먼저 CART를 이용하여 규칙으로 생성되는 BI 2와 3에 대해서 AP 경계인지 MP 경계인지를 결정하는 decision tree를 구성하였다. Decision tree의 출력은 확률 값으로 나타내고, 확률 값을 이용하여 VB를 결정한다 [11].

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

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

$\hat{P}_i$는 AP/MP decision tree의 출력형태를 나타낸 것으로 AP경계가 될 확률 $P_{AP}(i)$와 MP경계가 될 확률 $P_{MP}(i)$의 vector로 출력된다. 그리고 $P_{AP}(i)$와 $P_{MP}(i)$의 관계는 식 2와 같다.

그림 1은 VB의 모델링 과정이다. VB 모델링은 AP/MP decision tree를 구성하는 것과 이것을 이용하여 VB의 결정과정에서 사용될 문턱값 ($\overline{P_{AP}}$, $\overline{P_{MP}}$)를 계산하는 것으로 이루어진다. 먼저 CART의 훈련과 테스트를 위해 8,586문장의 수동으로 레이블링한 BI 정보를 이용하였다. 5,769문장으로 훈련하고 2,718문장으로 테스트하였다. BI 레이블링은 한 명의 전문가가 녹음된 음성을 듣고 수행하였다. 표 4는 CART 학습에 사용한 특징들이다. W_k는 k번째 단어이고, M_m은 W_k의 마지막 mora이다.

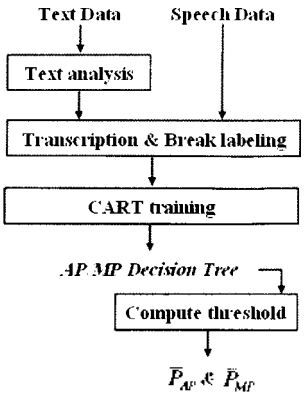


그림 1. 가변 break 모델링
Fig. 1. Modeling of varialbe break.

(mora는 시간적 개념의 음절로 박 (拍)이라고도 하고 1mora는 단모음을 포함하는 1음절을 나타낸다. 발음 (撥音)과 촉음도 1mora이고 장모음, 이중모음은 1음절 2mora이다.) 7번째 특징은 형태소에 관한 것으로, 표 5와 같이 비슷한 문법적 기능을 가지는 단어들을 분류하여 사용하였다. 예로 앞 단어와 강하게 연결되는 형식명사 (structural nouns)와 조동사 (auxiliary verbs)인 단어들을 분류하고, 관용구를 형성하는 단어들과 자주 사용되는 동사 및 조사 등을 분류하였다. 8번째 특징은 이전 (previous) BIs이다. 학습된 트리는 81개의 단말노드를 가진다.

그림 2는 실제 AP와 MP의 경계에 대한 예측 확률의 분포를 나타낸 것이다. 실제 AP 또는 MP란 성우가 발성한 break가 AP 또는 MP 경계임을 나타내는 것이고, 그림 a)는 성우가 읽은 동일한 텍스트를 이용하여 AP/MP를 예측하였을 때의 AP의 확률분포를 나타낸 것이고, b)는 MP의 확률분포를 나타낸 것이다. 실제 AP로 발성되는 경우는 예측 확률도 대부분 1에 근접한 값으로 나타났지만, MP로 발성되는 경우는 0.8~0.9의 확률 구간이 가장 많이 나타나고 30%이하였다. 이것은 AP로 읽는 패턴은 매우 일정하여 규칙적임을 나타내는 것이고 MP로 읽는 패턴은 불규칙적이고 AP로 읽을 확률도 높음을 나타내는 것이다.

$$\text{Precision} = \frac{\text{AP(또는 MP) 예측이 성공한 경우의 개수}}{\text{AP(또는 MP)로 예측된 경우의 개수}}$$

$$\text{Recall} = \frac{\text{AP(또는 MP) 예측이 성공한 경우의 개수}}{\text{실제 AP(또는 MP) 경계 개수}}$$

$$\text{F1-Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{(\text{Precision} + \text{Recall})}$$

표 6은 AP/MP decision tree의 테스트 결과이다. 실제

표 4. CART 모델링에 사용한 특징들
Table 4. Features used for CART Modeling.

No	Features
1	Number of mora in (W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2})
2	Number of mora before/after W_i within IP
3	Part-of-speech type of (W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2})
4	Tone pattern of consecutive five mora before Mn
5	Tone pattern of consecutive five mora after Mn
6	Phoneme of mora in (W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2})
7	Kind of morph (W_{i-1}, W_i, W_{i+1}, W_{i+2})
8	Break of (W_{i-2}, W_{i-1})

표 5. 표4의 7번째 특징으로 사용된 Morph의 종류
Table 5. Kind of morphs for 7th feature in table 2.

Set	Morph
0	structural nouns (ex:こと,ため,間,あいだ,上,うえ,内,うち, etc.)
1	auxiliary verbs (ex:した,た,ます,る,ない,なり,う,よう,いた,ました,ました, etc.)
2	auxiliary adjectives (ex:ない,なく,なかった,なくて,っぽい,やすい,にくい,らしい,なければ, etc.)
3	time nouns (ex: 一昨日,昨日,きのう,今日,明日,あさって,朝,昼,夜,午前,午後, etc.)
• • n	(ex:お断り,お願い,お電話,お知らせ, etc.)

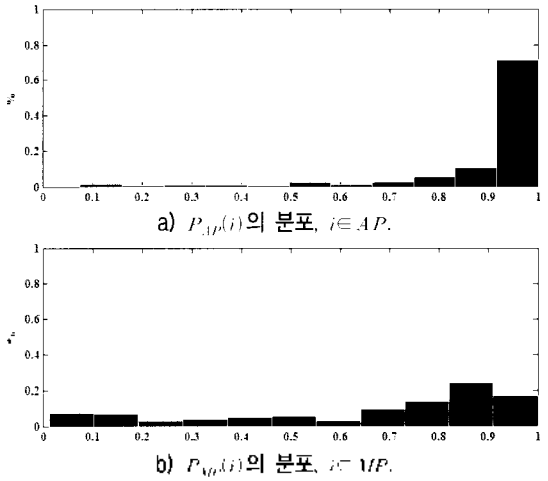


그림 2. 실제 AP, MP 경계의 예측 확률의 분포
Fig. 2. Distributions of prediction probabilities in real AP and MP boundaries.

표 6. AP/MP의 decision tree의 테스트 결과
Table 6. Performance of AP/MP prediction tree.

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

AP 경계에 대하여 $P_{AP}(i) > P_{MP}(i)$인 경우 TRUE로 그렇지 않으면 FALSE로 결정하였다. 실제 MP 경계에 대해서도 $P_{AP}(i) < P_{MP}(i)$인 경우 TRUE로 그렇지 않으면 FALSE로 결정하였다. 위와 같이 TRUE와 FALSE에 대해 confusion 행렬을 만들고 precision, recall, F1-measure를 측정하였다. AP에 대한 성능이 MP 보다 높게 나타났고, 전체 accuracy는 88.52%를 나타내었다.

VB를 결정할 때 사용되는 문턱값, P_{AP}와 P_{MP}는 다음과 같다.

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

$$\overline{P_{AP}} = \frac{1}{N_{AP}} \sum^{N_{AP}} P_{AP}(i)\delta_{AP}(i) \quad (4)$$

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

$$\overline{P_{MP}} = \frac{1}{N_{MP}} \sum^{N_{MP}} P_{MP}(i)\delta_{MP}(i) \quad (6)$$

$\overline{P_{AP}}$는 AP 경계에 대해 예측이 정확한 경우의 $P_{AP}(i)$의 평균이고, $\overline{P_{MP}}$는 MP 경계에 대해 예측이 정확한 경우의 $P_{MP}(i)$의 평균이다. N_{AP}와 N_{MP}는 AP 경계와 MP 경계의 개수이다. P_{AP}는 0.9245, P_{MP}는 0.8085로 나타났다.

3.3. 가변 Break의 결정

합성기에서 생성된 BI 중 0,1,4,5는 FB가 되고, BI 2와 3에 대해서만 VB와 FB를 결정한다. VB는 AP/MP decision tree와 문턱값를 이용하여 그림 3과 같이 결정한다. 생성된 BI가 2와 3인 경우 decision tree를 이용하여 예측확률을 구하고 이것이 문턱값 보다 높은 경우는 FB로 하고 낮은 경우에는 VB로 결정한다.

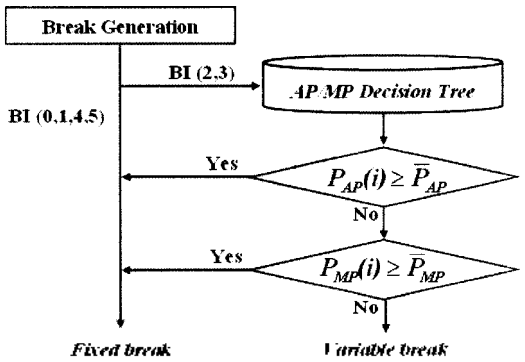


그림 3. 가변 break의 결정
Fig. 3. Decision of variable break.

$$\tilde{P}_i = \begin{cases} [1.0] & P_{AP}(i) \geq \overline{P_{AP}} \\ [0.1] & P_{MP}(i) \geq \overline{P_{MP}} \\ \hat{P}_i & otherwise \end{cases} \quad (7)$$

VB는 $P_{AP}(i)$와 $P_{MP}(i)$가 모두 각각의 평균 보다 작은 경우이고, 그렇지 않은 경우 (FB 인 경우) $P_{AP}(i)$와 $P_{MP}(i)$를 1 또는 0으로 조정한다. $\tilde{P}_i$는 VB를 적용하여 $\hat{P}_i$를 다시 나타낸 것이다.

표 7은 AP/MP decision tree의 테스트 결과를 이용하여 VB의 비율을 분석한 것이다. 예측이 정확하지 않은 경우의 VB의 비율 (b)와 (c)가 예측이 정확한 경우의 (a)와 (d)보다 높게 나타났다. 이것은 합성기에서 BI 생성 에러가 발생할 때 VB로 결정될 확률이 높은을 보여 주는 것이고 합성단위 선택 과정에 의해 BI 생성 에러를 보완할 수 있음을 나타낸다. 즉 동일하게 VB로 결정되었다고 해도 예측이 성확한 경우에는 생성된 BI를 가지는 합성단위들이 많이 포함되는 후보 합성단위 엔트리가 구성되고 예측이 정확하지 않은 경우에는 생성된 BI가 아닌 다른 BI를 가지는 합성단위들이 많이 포함되는 후보 합성단위 엔트리가 만들어진다는 것이다. 이렇게 구성된 후보 합성단위 엔트리를 이용하여 합성단위 검색을 수행하면 합성음에서의 BI의 정확도가 높아질 수 있다.

IV. Multiple Prosodic Targets

합성기에서 합성단위 선택을 위해 BI, 음소 지속시간, 기본주파수 포락선으로 구성되는 prosodic target (PT)을 생성한다. 기존의 합성기에서는 BI가 생성되면 변하지 않기 때문에 하나의 PT (single PT)만을 사용하지만 VB를 사용하게 되면 생성된 BI가 바뀔 수 있기 때문에 여러 개의 PT (multiple prosodic targets)를 생성해야만 한다.

그림 4는 제안하는 운율 생성 과정을 나타낸 것이다. 기존의 break 생성을 수행한 후 VB 예측을 수행하고, 'multiple pitch & duration generation'을 수행한다.

표 7. 가변 break 비율
Table 7. Rate of variable break.

prediction / target	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)

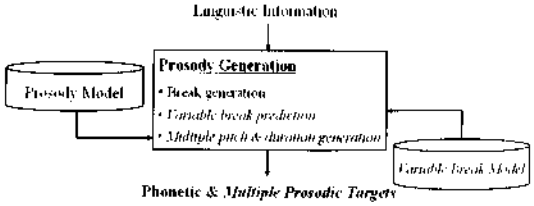
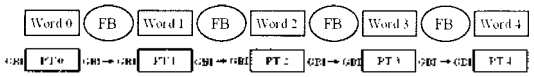
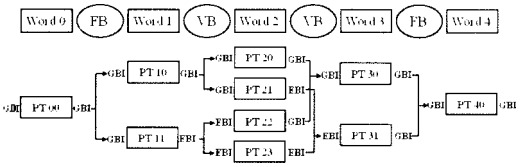


그림 4. 제안하는 운율 생성 과정
Fig. 4. Proposed Prosody Generation.



(a) Single Prosodic Target.



(b) Multiple Prosodic Targets.

그림 5. Single과 Multiple Prosodic Targets
Fig. 5. Single and Multiple Prosodic Targets (PT-Prosodic Target, GBI-Generated BI, EBI-Expanded BI).

Word의 PT는 그 word의 left BI와 right BI에 따라 달라질 수 있다. 그림 5의 (a)와 같이 일반적으로 BI가 변하지 않는 FB인 경우는 하나의 word에 하나의 PT만이 존재한다. 그러나 (b)와 같이 VB를 사용할 경우에는 각 word의 BI는 생성된 BI (GBI : generated BI)와 바뀔 수 있는 BI (EBI: expanded BI)가 될 수 있다. 따라서 하나의 word에 최대 4가지의 PT가 필요하다. BI가 2이고 VB이면 EBI는 BI 3이 되고, BI가 3이면 EBI는 2가 된다.

본 논문에서는 제안하는 합성단위 선택을 수행한 후 선택된 음성파형을 PT에 대하여 신호처리를 하지 않고 간단한 OLA (Overlab-Add)를 이용하는 연결방법만으로 합성음을 생성한다. 그러나 PT를 이용하여 기본주파수와 음소지속시간에 대해 신호처리를 수행하여 합성음을 생성하는 시스템이라면, 합성단위 선택 결과를 이용하여 multiple prosodic targets에서 single prosodic target을 재구성하여 이것을 이용하여 운율에 대한 신호처리를 수행 할 수 있다.

V. 가변 Break를 이용한 합성단위 선택

합성단위 선택 기반 음성 합성시스템의 선택 과정 자체는 동적 프로그래밍 (dynamic programming, Viterbi) 알고리즘으로 수행되지만 그 전에 이루어지는 후보 합성단위 (candidate) 선택과 비용 (cost) 계산이 합성음의 음질에 보다 많은 영향을 미친다 [8]. 일반적으로 목표 문맥 (target context) 정보로 후보 합성단위를 선택하여 이것들에 대하여 목표 비용 (target cost) 및 연결 비용 (concatenation cost, join cost)을 계산하는데, 후보 합성단위의 수가 많을수록 합성음의 음질이 좋아 질 수 있다. 자연스러운 합성음을 생성하기 위해서는 음성 코퍼스에 다양한 운율 정보가 포함되어 있어야 하고 후보 합성단위도 코퍼스의 이러한 특징이 반영되도록 가능성 있는 많은 합성단위들을 포함하여야 한다. 그러나 후보 합성단위의 수가 증가 할 경우 합성단위 검색 (Viterbi search)과정의 수행시간이 급격히 늘어나 실시간 합성이 어려워지기 때문에 후보 합성단위의 수를 무조건 늘리는 것도 효율적이지 못하다 [9].

본 논문에서는 그림 6과 같이 가변 break에 대해서 BI를 확장하여 후보 합성단위의 개수를 증가시킨다. GBI를 가지는 합성단위 뿐 아니라 EBI를 가지는 합성단위도 후보에 포함 시키는 것이다. BI의 확장으로 인하여 합성단위의 수가 많아지면 사전선택을 수행한다 [5]. 사전선택의 비율은 식 7을 사용한다. 예를 들어 생성된 BI가 2이고 $P_{AP}(i)$가 0.6이면, 합성단위 선택의 엔트리는 BI가 2인 후보 합성단위가 60%가 되고 BI가 3인 후보 합성단위가 40%가 되도록 사전선택을 수행한다.

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

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

식 8은 목표 비용함수 (target cost function)을 나타낸 것이다. TC^{p}는 운율 비용 (prosodic cost)이고 TC^{c}는 문맥정보 비용 (context cost)이다. 본 논문에서는 multiple

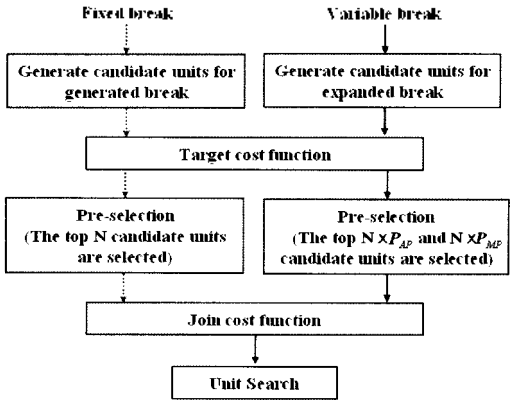


그림 6. 가변 break를 이용한 합성단위 선택
Fig. 6. Proposed Unit-Selection.

prosodic targets을 사용하므로 식 9와 같이 각 prosodic target에 대하여 비용 계산하고 그 중 최소값을 선택한다.

VI. 실험결과 및 고찰

AP/MP decision tree를 테스트하기 위한 2,817개의 문장과 녹음 데이터를 이용하여 예측 결과와 합성단위 선택 결과의 변화를 실험하였다. 녹음 음성에서의 경계정보가 정답이라고 가정하고, 가변 break를 사용하지 않았을 때와 사용했을 때 예측된 결과와 실제 선택된 합성단위의 경계정보가 어떻게 차이가 나는지 알아보았다. 합성단위 선택의 결과만 분석하기 위해 2,817개의 문장은 DB 처리 (발음전사 및 AP/MP 표기) 결과와 합성기의 언어처리 결과가 동일한 것을 사용한 것이다. 그리고 위 문장들은 합성기의 DB에 포함하여 테스트하였다. 합성기가 음질에 대해 최대의 성능을 보이는 경우는 DB에 포함된 문장과 입력 문장이 일치하는 경우 모든 합성단위를 연결하여 가져오는 경우이다. 그러나 실제 합성기에서는 언어처리의 오류나 break 예측 오류로 인해, DB에 입력문장과 일치하는 합성단위들이 존재하여도 이것들을 모두 사용하지 못하게 된다. 따라서 break예측 오류가 존재하여도 이것을 보완하기 위한 것이 가변 break이므로 테스트 문장에 대한 DB를 합성 DB에 포함하여 테스트하였다.

표 8은 실제 AP에 대한 예측 및 합성 결과를 나타낸 것이다. (표에서 AP는 AP 경계, MP는 MP의 경계이고, IP&SB는 IP와 문장경계 (Sentence Boundary)이고, WB는 단어 경계이다. '예측률'은 예측된 경우의 합이고, '합성률'은 합성된 경우의 합이다.) 관측된 AP 경계는 5,762개이다. 예측 결과만 생각하면 가변 break를 사용하지 않은 경우의 정확도가 95.45%이고, 사용한 경우 71.42%였다. (실제 AP에 대한 실험 결과이므로 예측률이 곧 예측 정확도이다. 표 (a)의 예측률 95.45%는 ①, ②, ③, ④의 합이고, 실제 AP에 대하여 ①은 AP로 예측하고 AP로 합성한 경우, ②는 AP로 예측하고 MP로 합성한 경우, ③은 AP로 예측하고 IP&SB로 합성한 경우, ④는 AP로 예측하고 WB로 합성한 경우로 모두 AP로 예측한 경우이다.) 가변 break를 사용하지 않을 때는 CART 모델링을 이용하지 않고 규칙으로 예측하기 때문에 MP 보다 AP로 예측하는 경우가 많다. 실제 발성에서도 AP 경계의 수가 MP 경계의 수보다 매우 많고, MP로 예측 하는 규칙이 제한적이므로 이와 같은 결과가 나타난다. 하지만 합성된 결과에서 정확도를 분석해 보면, 가변 break를 사용하지 않은 경우 93.72%이고, 사용한 경우 95.72%로 가변 break를 사용한 경우가 조금 좋은 결과를 나타내었다. (실제 AP에 대한 실험 결과이므로 합성률이 곧 합성의 정확도이다. 표 (a)의 합성률 93.725%는 ①,⑤,⑥,⑦의 합이고, 실제 AP에 대하여 ①은 AP로 예측하고 AP로 합성한 경우, ⑤는 MP로 예측하였지만 AP로 합성한 경우, ⑥은 IP&SB로 예측하였지만 AP로 합성한 경우, ⑦는 WB로 예측하였지만 AP로 합성한 경우로 모두 AP로 합성한 경우이다.)

가변 break를 사용하였을 때, 예측에서 실제 AP경계인 것을 MP로 예측하는 오류는 증가하였으나, 합성단위 선택에서 제안한 방법을 사용하여 이러한 오류를 줄일 수 있었던 것이다. 즉, AP경계를 MP경계로 잘못 예측했지만, 합성단위 선택과정에서 예측된 MP경계뿐만 아니라 AP 경계를 가지는 합성단위들을 포함하는 후보 합성단위 엔트리를 구성하기 때문에 오류가 수정될 수 있는 것이다. 가변 break를 사용하는 제안하는 합성단위 선택방법에서는 후보 합성단위 엔트리를 2가지 (AP, MP)의 서로 다른 경계 정보를 가지는 후보들로 구성함으로써 예측 오류의 보완이 가능한 것이다.

가변 break의 사용은 AP 경계보다 MP 경계에서 보다 그 효과가 크게 나타난다. 표 9는 실제 MP에 대한 예측

표 8. 실제 AP에 대한 예측 및 합성 결과
Table 8. The result of prediction and synthesis for real AP.

(a) 가변 Break를 사용하지 않은 경우

Synthesis / Prediction	AP	MP	IP&SB	WB	예측률
AP	93.64% ①	0.03% ②	0.45% ③	1.33% ④	95.45%
MP	0.08% ⑤	3.44%	0.43%	0.20%	4.15%
IP & SB	0.00% ⑥	0.00%	0.31%	0.00%	0.31%
WB	0.00% ⑦	0.00%	0.00%	0.09%	0.09%
합성률	93.72%	3.47%	1.19%	1.53%	

(b) 가변 Break를 사용한 경우

Synthesis / Prediction	AP	MP	IP&SB	WB	예측률
AP	69.46%	0.43%	0.47%	1.06%	71.42%
MP	26.26%	0.82%	0.23%	0.31%	27.62%
IP & SB	0.00%	0.07%	0.26%	0.00%	0.33%
WB	0.00%	0.00%	0.00%	0.63%	0.63%
합성률	95.72%	1.32	0.96	2.00%	

및 합성 결과를 나타낸 것으로 2,817개의 문장에서 3,253개의 실제 MP 경계에 대하여 합성기의 예측 결과와 합성단위 선택 결과인 합성음의 경계를 비교한 것이다. AP의 경우와 마찬가지로 2,817개의 문장은 합성 DB를 구축할 때 사용한 문장 중에 언어처리 결과와 실제 발성한 결과가 동일한 문장이다. 가변 break를 사용하지 않았을 때의 예측 정확도는 45.45%이고, 합성 결과에서도 42.16%이다. 규칙만을 이용해서는 정확한 MP 경계를 예측하고 합성할 수 없음을 보여주는 결과이다. 그러나 CART를 이용하여 MP를 모델링하여 예측하고, 가변 break를 사용하면 예측 정확도는 78.81%이고, 합성 결과에서는 96.18%로 매우 우수한 결과를 얻을 수 있었다. 비록 합성 DB에 포함된 텍스트를 이용하여 테스트 한 결과이지만 가변 break의 사용이 보다 효율적으로 DB를 사용함을 나타낸다. MP 경계에 대한 예측이 잘못되었어도 제안한 합성단위 선택 방법에 의해 합성음에서는 오류가 수정되어 보다 정확한 break를 구현할 수 있었다. 즉, 고정 break만 사용하는 경우 예측이 정확하지 않으면 합성음에서 break를 수정할 수 있는 방법이 존재하지 않으나, 가변 break를 사용하고, 제안한 방법으로 합성단위 선택을 수행하면 예측의 오류를 보완할 수 있는 것이다.

제안한 합성기의 음질을 평가하기 위해 MOS (Mean Opinion Score) 테스트를 수행하였다. 음성 코퍼스는 방음된 녹음실에서 전문 여성 아나운서에 의해 녹음된 4시간의 음성 데이터를 사용하였고, 녹음 시간은 발성의 처음과 끝의 묵음은 포함되지 않았고, 발성 중간의 묵음은 포함 되었다. 녹음을 위해 사용된 대본은 뉴스기사, 소설, 대화체 문장 및 숫자, 알파벳, 인터넷 주소 (URL) 등으로 구성하였다. 테스트는 일본인 여성 5명이 참가하였고, 테스트 문장은 JEITA 종합평가문장 [10] 중 127문장을 선택하였다. MOS 테스트는 원음 127개와 VB를 사용한 시스템1과 사용하지 않은 시스템2로 생성한 합성음 254개를 섞어 불규칙한 순서로 청취하고, 5개의 레벨 (1~5, Bad, Poor, Fair, Good, Excellent) 중 하나를 선택하도록 하였다. 테스트 결과 원음은 4.99, 시스템 1은 4.25, 시스템2는 4.01을 나타내었다.

표 9. 실제 MP에 대한 예측 및 합성결과
Table 9. The result of prediction and synthesis for real MP.

(a) 가변 Break를 사용하지 않은 경우

Synthesis / Prediction	AP	MP	IP&SB	WB	예측률
AP	43.00%	**0.00%**	0.09%	7.51%	50.6%
MP	**0.56%**	**42.16%**	0.19%	2.54%	45.45%
IP & SB	0.00%	**0.00%**	3.10%	0.00%	3.10%
WB	0.00%	**0.00%**	0.00%	0.85%	0.85%
합성률	43.56%	**42.16%**	3.38%	10.90%	

(b) 가변 Break를 사용한 경우

Synthesis / Prediction	AP	MP	IP&SB	WB	예측률
AP	0.29%	**17.21%**	0.03%	0.22%	17.75%
MP	**0.18%**	**77.77%**	**0.52%**	**0.34%**	**78.81%**
IP & SB	0.03%	**1.20%**	0.52%	0.00%	1.75%
WB	0.00%	**0.00%**	0.00%	1.69%	1.69%
합성률	0.50%	**96.18%**	1.07%	2.25%	

VII. 결 론

합성음의 자연성은 운율에 의해 결정되므로, 보다 자연스러운 운율을 구현하는 것은 모든 음성 합성기의 공통된 목표이다. 대용량 코퍼스 기반 TTS 시스템은 음성 데이터베이스에 이미 다양한 운율 정보를 저장하고 있어 이를 효율적으로 이용한다면 충분히 자연스러운 운율을 구현 할 수 있다. 하지만 합성기에서 생성하는 운율이 제한적이고 이것을 이용하여 합성단위를 선택함으로써 자연스러운 합성음을 얻기 힘들다.

본 논문에서는 합성음의 자연성을 향상시키기 위해 생성된 운율을 보다 효율적으로 이용할 수 있는 방법을 제안하였다. 운율정보의 하나인 break에 가변 break를 도입하여 음성 데이터베이스의 각 세그먼트가 가지는 다양한 운율정보를 이용할 수 있는 합성단위 선택을 가능하게 하였다.

감사의 글

본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.

참 고 문 헌

1. R. E. Donovan, "Trainable speech synthesis," PhD. Thesis, *Cambridge University, Engineering Department*, pp. 1-28, 1996.
2. X. Sun and T. H. Applebaum, "Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model," *Proc. EUROSPEECH2001*, vol. 1, pp. 537-540, Sep. 2001.
3. N. Campbell, "Autolabeling Japanese ToBI," *Proc. ICSLP'96*, vol. 4, pp. 2399-2402, Oct. 1996.
4. J. J. Venditti, "The J_ToBI model of Japanese intonation," in

Prosodic Typology: The Phonology of Intonation and Phrasing, ed. S. A., pp. 172-200, Oxford University Press, New York, 2005.
5. S. Kiriyama, S. Kitazawa, "Evaluation of a prosodic labeling system utilizing linguistic information," *Proc. INTERSPEECH 2004*, pp. 2993-2996, Oct. 2004.
6. K. Maekawa, H. Kikuchi, Y. Igarashi, J. J. Venditti, "X-JToBI: an extended J-ToBI for spontaneous speech," *Proc. ICSLP-2002*, pp. 1545-1548, Sep. 2002.
7. 이상호, 오영환, "CART를 이용한 운율구 추출 및 휴지기간 모델링," *한국음향학회 학술발표대회 논문집* 17권 1호, 81-86쪽, 1998.
8. A. Conkie, M. C. Beutnagel, A. K. Syrdal, P. E. Brown, "Preselection of candidate units in a unit selection-based text-to-speech synthesis system," *Proc. ICSLP2000*, vol. 3, pp. 314-317, Oct. 2000.
9. 나덕수, 민소연, 이광형, 이종석, 배명진, "일본어 악센트 특징을 이용한 합성단위 선택 기반 일본어 TTS의 후보 합성단위의 사전선택 방법," *한국음향학회지* 26권, 4호, 159-165쪽, 2007.
10. Technical Standardization Committee on Speech Input/Output Systems, Speech synthesis system performance evaluation methods, *JEITA IT-4001*, pp. 42-45, 2003.
11. D. S. Na and M. J. Bae, "A Variable Break Prediction Method using CART in a Japanese Text-to-Speech System," *IEICE Trans. Inf. & Syst.*, vol. E92-D, no. 2, pp. 349-352, 2009.

저자 약력

•나 덕 수 (Deok-Su Na)

2009년 2월 : 숭실대학교 정보통신공학과 (공학박사)
현재 : (주)보이스웨어 연구원
한국음향학회지 제19권 제2호 참조

•민 소 연 (So-Yeon Min)

2003년 2월 : 숭실대학교 전자공학과 (공학박사)
현재 : 서일대학 정보통신과 교수
한국음향학회지 제21권 제3호 참조

•이 종 석 (Jong-Seok Lee)

1983년 2월 : 서울대학교 전자공학과 (공학사)
1985년 2월 : 서울대학교 전자공학과 (공학석사)
1995년 2월 : 서울대학교 전자공학과 (공학박사)
1985년 ~ 2000년 : LG 중앙연구소 선임연구원
2001년 ~ 현재 : (주)보이스웨어 부사장

•배 명 진 (Myung-Jin Bae)

현재 : 숭실대학교 정보통신전자공학부 교수
한국음향학회지 제21권 제3호 참조