한국음향학회지 제26권 제8호 pp. 403~408 (2007)

일본어 합성기에서 유동 Break를 이용한 합성단위 선택 방법

A Unit Selection Methods using Flexible Break in a Japanese TTS

송 영 환*, 나 덕 수**, 김 종 국*, 배 명 진*, 이 종 석**
(Young-Hwan Song*, Deok-Su Na**, Jong-Kuk Kim*, Myung-Jin Bae*, Jong-Seok Lee**)
*숭실대학교 정보통신 전자공학부, **보이스웨어 기술연구소
(접수일자: 2007년 8월 6일; 수정일자: 2007년 8월 27일; 채택일자: 2007년 9월 5일)

대용량 코퍼스를 이용하는 합성단위 선택 (unit selection) 기반 합성기에서 break는 자연성 및 명료성에 큰 영향을 미치는 파라미터로 unit selection 과정에서 음소 정보와 함께 중요한 특징으로 사용된다. 일본어는 피치의 상대적 높낮이로 표현되는 악센트를 가지는 언어이고, 악센트의 변화에 따라 AP (Accentual Phrase)가 결정되고 AP 경계에서 break가 형성된다. break는 규칙 기반 방식이나 통계적 방식인 J-ToBI를 이용하여 예측 할 수 있으나 다양성으로 인해 정확한 예측이 어렵다. 따라서 본 논문에서는 다양한 운율 정보를 포함하고 있는 대용량 코퍼스의 장점을 이용하기 위해 break를 고정 break와 유동 break로 나누어 합성단위 검색을 수행한다. 실험 결과 제안한 합성단위 선택 방법으로 합성음의 자연성을 향상 시킬 수 있었다.

핵심용어: 일본어 음성합성, 합성단위 선택

투고분야: 음성처리 분야 (2.4)

In a large corpus-based speech synthesizer, a break, which is a parameter influencing the naturalness and intelligibility, is used as an important feature during a unit selection process. Japanese is a language having intonations, which are indicated by the relative differences in pitch heights and the APs (Accentual Phrases) are placed according to the changes of the accents while a break occurs on a boundary of the APs. Although a break can be predicted by using J-ToBI (Japanese-Tones and Break Indices), which is a rule-based or statistical approach, it is very difficult to predict a break exactly due to the flexibility. Therefore, in this paper, a method is to conduct a unit search by dividing breaks into two types, such as a fixed break and a flexible break, in order to use the advantages of a large-scale corpus, which includes various types of prosodies. As a result of an experiment, the proposed unit selection method contributed itself to enhance the naturalness of synthesized speeches.

Key words: Ource localization, Waveguide invariant parameter, Interference pattern matching, Circle

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

현재 상용화되거나 연구되고 있는 음성합성 기술 중 합성음의 음질이 가장 우수한 것은 대용량 음성 코퍼스를 이용한 합성단위 선택 기반 연결형 합성 기술이다. 이 기술의 가장 큰 장점은 기존의 규칙합성 시스템이 가지고 있는 제한적인 운율 변화에 의한 자연성 감소의 단점을 극복한 것이다. 대용량 음성 코퍼스 (speech corpus)의 구축을 통해 다양한 운율변화를 구현할 수 있게 됨으로써 사람의 목소리와 비슷한 음질의 합성음을 생성할 수 있게 된 것이다.

이러한 음성합성 시스템의 합성단위 선택 과정은 문맥정보와 운율 파라미터에 의해 결정되는데 보다 자연스러운 합성음을 얻기 위해서는 정확한 운율 모델링이 필수적이다. 입력 텍스트에서 운율 파라미터를 생성하기 위해서는 억양구 경계 결정, 음소 지속시간 결정, 기본주파수의 윤곽선 설정의 3가지의 기본적인 모듈이 필수적이다 [1].

책임저자: 송 영 환 (song0one@yahoo.co.kr)
156-743 서울시 동작구 상도동 511 숭실대학교 전자공학과
(전화: 02-824-0906; 팩스: 02-824-0906)

억양구의 경계를 결정하기 위해서는 문장에 대한 정확한 분석 (통사론적인 측면과 의미론적인 측면의 분석)이 이루어져야 하는데, 자동으로 이러한 것이 이루어지기 매우 힘들다. 따라서 이러한 억양구 경계정보의 오류는 합성음에서 매우 빨리 읽는 현상을 유발하여 의미의 혼돈을 초래하거나, 일부분의 오류가 나머지 부분에 영향을 미치기도 한다 [1]. 현재 억양구 경계를 결정하기 위해 주로 사용되는 방법에는 규칙 기반, 데이터 기반 그리고 두 가지를 혼용하는 방식이 있다. 규칙 기반 방법은 문장기호, 품사, 발음열 등의 문장 분석으로 얻어지는 정보와 언어학적인 정보를 이용하여 작성하게 되는데, 우수한 성능을 얻기 위해서는 매우 복잡하고 정교한 작업이 필요하다. 데이터 기반 방법은 여러 가지 특징들을 이용하여 자동으로 결정 트리 (decision tree)를 구축하는 CART (Classification and Regression Trees) 방식이 주로 사용되고 있다.

음성 세그먼트의 지속시간과 기본주파수의 윤곽선을 결정하는 방법은 오래전부터 연구되고 있는데, 규칙합성기에서는 주로 규칙 기반 방식이 사용되고 있으나 코퍼스를 이용하는 합성기에서는 각각 CARTs와 ToBI (Tones and Break Indices) 레이블링 시스템을 주로 사용하고 있다.

일본어의 억양구 (Intonation Phrase)는 악센트로 인해 형성되는 몇 개의 AP (Accentual Phrase)로 구성되므로 [2] 일본어 합성기에서는 AP의 경계 정보를 결정하는 것이 곧 억양구의 경계 정보를 결정하는 것이 된다. AP의 경계를 결정하기 위해서는 우선 악센트를 표현하여야 하는데, 일본어의 악센트는 강약이 아닌 고저의 악센트이고 [2], AP의 첫음절과 두 번째 음절은 반드시 악센트의 위치가 바뀌며, 하나의 AP 안에서 악센트가 한번 내려가면 (고 악센트에서 저 악센트로 바뀌면) 다시 올라가지 못하는 특징이 있다 [3]. 이러한 특징에 의해 일본어 악센트를 표현하기 위해서는 악센트가 높은 곳에서 낮은 곳으로 떨어지는 위치와 형태를 표시해야 한다. 기존의 일본어 합성기에서는 이러한 악센트 기호를 별도로 정의하여 발음기호 사이에 표시하는 방법을 사용하는데, 이러한 방법은 합성단위 선택 기반 합성기 보다 악센트를 피치 조절로 구현하는 규칙 합성기 방식에 적합한 표현 방법이다 [4]. 따라서 본 논문에서는 선행 연구로 얻어진 악센트 정보가 결합된 발음기호를 이용한 Break 예측 방법 [5]을 이용하여 BI (Break Indices)를 결정한다.

AP의 경계 정보에는 위치 정보 뿐 아니라 인접 AP와의 연결 정보도 포함된다. 이러한 AP 연결정보는 문장의 의미나 구조에 의해, 연결되는 것과 끊어지는 것으로 나누어질 수 있으며, 보다 자세하게 나눈다면, 연결되는 것 중에도 끊어져도 되는 것과 반드시 연결되어야 되는 것으로 나눌 수 있고, 반대로 끊어지는 AP 연결정보 중에는 연결되어도 되는 것과 반드시 끊어져야만 하는 것으로 나눌 수 있다. 즉, 동일하거나 비슷한 형태의 문장을 읽은 음성 데이터를 관찰해 보면, 항상 같은 연결정보가 나타나는 AP도 존재하지만 두 가지 종류의 연결정보가 모두 나타나는 AP도 존재한다.

본 논문에서는 이러한 AP 연결정보의 다양성을 이용하여 합성단위 선택과정에서 후보 합성단위 수를 증가시켜 합성음의 자연성을 향상시킬 수 있는 방법을 제안한다. 먼저 AP 연결정보의 다양성을 반영하기위해 BI와 함께 각 BI의 다양성 정도를 고정 (fixed) break 또는 유동 (flexible) break로 결정한다. 즉 항상 같은 형태로 유지되어야 하는 break를 고정 break로, 끊어지거나 연결되는 것 모두 허용되는 것을 유동 break로 결정한다. 고정 break는 기존의 방법으로 합성단위 선택과정을 수행하고, 유동 break는 예측된 BI를 변경하면서 후보 합성단위의 수를 증가시켜 BI가 합성단위 선택 과정에서 연결 코스트에 의해 변화될 수 있게 함으로써 BI 예측 오류를 보완할 수 있도록 한다.

그림 1은 본 논문에서 사용한 TTS 시스템의 구성도이다. 대부분의 합성단위 선택 기반 연결 합성기처럼 4가지의 중요한 모듈, 언어처리 모듈 (linguistic processing module), 운율처리 모듈 (prosody generation module), 합성단위 선택 모듈 (unit selection module), 음성파형 생성 모듈 (waveform generation module)로 구성되고, 합성의 기본 단위로 폰 (phone)을, 텍스트 코드로 일본어 Shift-JIS를 사용하였다 [6].

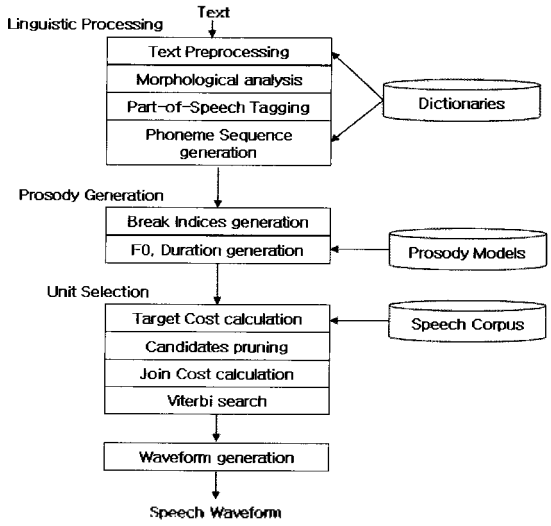


그림 1. 일본어 TTS 시스템
Fig. 1. The Japanese TTS System.

II. 고정 Break와 유동 Break

TTS 시스템에서 break를 규칙 기반 방식으로 생성하는 경우 문장의 길이나 단어들의 의미론적 또는 형태론적인 변화에 의해 발생하는 다양성을 반영하기 어렵다. 일본어도 악센트 구 사이의 break에 따라 의미나 명료성이 변화할 수 있지만, 동일한 문장을 발성하는 경우에도 서로 다른 break가 허용될 수도 있다. 특히 문장이 길어지거나 문장의 구조가 복잡할수록 이러한 break의 사용이 모호해 지는 경우가 많아진다. 그리고 합성단위 선택 과정에서 하나의 break는 인접한 합성단위 뿐 아니라 인접하지 않은 합성단위들의 선택에도 영향을 미치므로 합성음의 음질이 최적이 될 수 있도록 break를 결정하여야한다.

대용량 음성 코퍼스를 사용하는 경우 합성단위 선택 과정에서 목표 (target) break 뿐만 아니라 타당한 break들을 모두 이용한다면 음성 코퍼스에 포함된 다양한 break 정보와 부족한 규칙에 대한 보완을 효율적으로 수행 할 수 있고, 다양한 문맥 (context) 정보를 가지는 합성단위들을 후보로 추출하여 합성음의 음질을 향상 시킬 수 있다. 따라서 본 논문에서는 변경되었을 경우 합성음의 음질 열화를 일으킬 수 있는 break를 고정 break로 설정하고, 그 외의 것들에 대해 유동 break로 설정하여 합성단위 선택을 수행하도록 하는 방법을 제안한다. 명사와 조사, 명사와 접미사 또는 복합명사를 만드는 명사와 명사처럼 결합되면서 악센트에 영향을 주는 단어들 사이의 break는 기본적으로 고정 break로 설정하고, 그 외에 의미론적 또는 형태론적인 규칙을 적용하여 고정 또는 유동 break를 설정한다. 이러한 것은 문법적 지식에 근거한 규칙과 실험적인 규칙으로 구성된다.

표 2는 유동 break예측의 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 유동 break를 예측하는 규칙은 인접한 품사의 종류를 검사하는 것과 같이 보다 일반적인 규칙들로 이루어지고, 고정 break를 예측되는

표 1. Break 인덱스
Table 1. Break Index.

0	(하나의 단어)
1	(하나의 AP안의 단어의 경계)
2	(포즈 없이 연결되는 AP 경계)
3	(포즈로 연결되는 AP 경계)
4	(IP와 IP 경계)
5	(문장 경계)

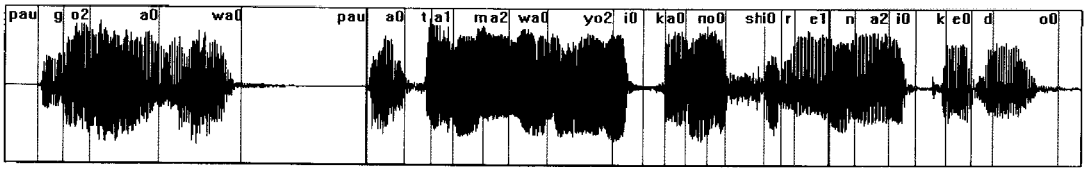
표 2. 유동 break의 예측 예
Table 1. Example of flexible break prediction.

조사 は	(と)+は+おもう (~라고 생각한다)의 경우		break 2 고정
	(と)+は+裏腹に, 裏腹て (~와는 정반대로)		
	と+は+형용사		
	그 외의 경우		break 3 유동
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 유동
그 외 조사의 기본처리	조사 뒤에 설명 구조	조사+명사+동사	break 2 유동
		조사+동사+동사, 조동사, 보조용언	
		조사+형용사 보조동사	
	조사 뒤에 수식 구조	조사+특정 부사(また, ...)	break 3 고정
특수 상황 처리	を가 타동사의 목적격 조사인 경우		break 2 유동
	を뒤에 숫자+접미사가 오는 경우		break 3 고정
	조사 の뒤는 기본적으로		break 2 유동
	조사 の뒤에 대명사가 오는 경우		break 3 고정
	조사 の가 숫자를 읽을 때 사용되는 경우		break 3 고정
	조사 に뒤에는 기본적으로		break 2 유동
	조사 に의 관용적 표현에 대한 처리:間(あいだ)+に		break 3 고정
	に가 시간을 나타낼 때		break 3 유동

규칙은 관용적 표현과 같이 특정한 단어의 조합 형태인지 검색하는 제한적인 규칙이 많다. 이러한 규칙은 일본어 문법과 그 예문을 분석하여 만들어졌다.

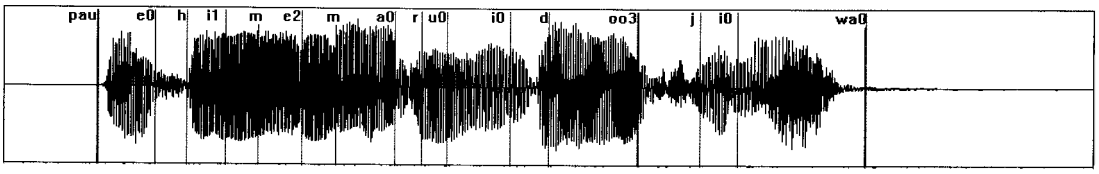
III. 유동 Break를 이용한 합성단위 선택방법

합성단위 선택 기반 음성 합성시스템의 선택 과정 자체는 dynamic programming 알고리즘으로 수행되지만 그 전에 이루어지는 후보 합성단위 (candidate) 선택과 비용 (cost) 계산이 합성음의 음질에 보다 많은 영향을 미친다. 일반적으로 목표 문맥 (target context) 정보로 후보 합성단위를 선택하여 이것들에 대하여 목표 비용 (target cost) 및 연결 비용 (join cost)를 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 증



(a) 조사 は[wa0]가 AP 경계(break 2와 break3)인 예문

텍스트 : ゴアは頭はよいかもしれないけど、
발 음 : # g o2 a0 % wa0 # a0 t a1 m a2 % wa0 / yo2 i0 % k a0 m o0 / sh i0 r e1 % n a2 i0 % k e0 d o0 #



(b) 조사 は[wa0]가 IP 경계(break 4)인 예문

텍스트 : えひめ丸移動時は、
발 음 : # e0 h i1 m e2 % m a0 r u0 / i0 d oo3 % j i0 % wa0 #

그림 3. break의 다양성의 예
Fig. 3. Example of break flexibility.

가 할 경우 합성단위 검색 (Viterbi search)과정의 수행시간이 급격히 늘어나 실시간 합성이 어려워지기 때문에 후보 합성단위의 수를 무조건 늘리는 것도 효율적이지 못하다 [4]. 따라서 본 논문에서는 break를 유동 break에 대해서만 문맥 정보를 변경하여 후보 합성단위의 수를 늘리는 방법을 제안한다. 또, 기본적으로 본 논문에 사용된 합성기는 실시간 합성을 위해 일본어에 적합한 악센트 구 매칭 방법을 사용하여 후보 합성단위에 대해 사전선택(preselection)을 수행하는데 [6], 유동 Break를 사용하는 장점이 줄어들지 않도록 수정하였다.

그림 2는 본 논문에서 사용한 합성단위 선택과정의 순서도이다. 먼저 합성단위는 기본적으로 폰 (phone) 단위이고, 후보 합성단위의 수가 문턱치 이하인 경우만 반음소(half-phone)단위로 합성단위 선택과정을 수행한다. 문턱치는 3으로 사용하였다.

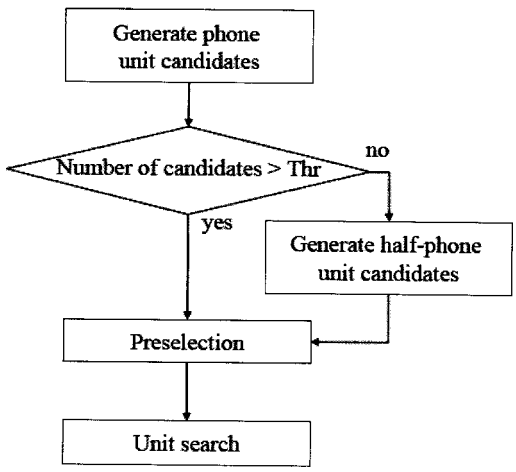


그림 2. 합성단위 선택과정
Fig. 2. Block diagram of unit selection.

그림 3은 음성 코퍼스에 존재하는 break 다양성을 나타낸 것으로 음성파형과 phone의 위치, 텍스트, 발음전사(transcription) 결과이다. 발음전사결과에서 %는 break 1, /는 break2, #은 break3(처음과 마지막 #은 제외)이다. 조사 'は'는 발음기호로 'wa0'이고, (a)의 4번째와 10번째, (b)의 마지막에 각각 나타나고, 이것의 break는 (a)에서는 AP의 경계인 2와 3이고, (b)에서는 IP의 경계인 4이다. 이처럼 조사 'wa0'는 코퍼스 안에서 다양한 break환경을 가질 수 있다. 그리고 break 3과 break4의 'wa0'는 포즈와 연결되는 매우 유사한 음운 환경을 가지고 있어 합성단위 선택과정에서 break 확장을 고려할 때 우선적으로 고려될 수 있다. 이처럼 음운 환경이 유사한 break는 0과 1, 3과 4, 4와 5이다.

표 3. 유동 break의 확장
Table 3. Expanding of flexible break.

break 인덱스	확장된 break
0	0, 1
1	1, 0
2	2, 3, 4
3	3, 4, 2
4	4, 5, 3
5	5, 4, 3

표 3은 본 논문에서 사용한 break 종류와 유동 break의 확장을 나타낸 것이다. break는 단어와 단어 사이의 경계 정보이고 이것은 합성단위 선택 과정에서 음소 (phone)의 경계 정보로 변환된다. 즉, 앞 단어의 마지막 음소와 뒤 단어의 첫 음소 경계 정보는 BI에 의해 결정되고, 그 외의 나머지 음소의 경계정보는 0이 된다. 각 break의 확장은 우선순위를 정하여 이루어지는데, 변하여도 영향이 적은 것

으로 먼저 확장한다. 즉 음운환경이 비슷한 것으로 우선 확장한다. break 0과 1은 동일한 AP를 형성하는 것으로 0은 1로, 1은 0으로 각각 확장한다. break 2이상은 모두 AP 또는 그 이상의 운율 경계를 나타내는 것으로 break 2는 포즈 (pause)와 연결되지 않고, 3, 4, 5는 포즈와 연결되는 특징을 가지는 것으로 서로 확장하면서 경계 정보 및 인접한 좌측 음소와 우측 음소의 문맥 정보도 변경한다. 그림 3 (a)의 텍스트가 입력으로 들어오고 처음으로 사용된 조사 'は'의 break가 BI는 2인 유동 break로 분석되었다면, tri-phone 기반의 문맥 정보는 [a0 % wa0 / a0]와 같다. 이것을 break 3으로 확장하려면 [a0 % wa0 #]로 변경하여야 한다. break 3인 경우 포즈와 연결되므로 tri-phone의 오른쪽 음소는 고려하지 않는다.

위와 같이 유동 break를 이용하면 기존의 문맥정보에 해당하는 후보 합성단위에 변경된 문맥에 의한 후보 합성단위가 추가되어 코퍼스에 포함된 다양한 운율을 이용 할 수 있을 뿐 아니라, 특히 후보 합성단위가 없거나 부족한 경우에도 안정된 합성단위 선택 결과를 얻을 수 있다. 그리고 후보 합성단위가 많아 사전선택이 필요한 경우 우선순위가 높은 경계 정보를 가지는 합성단위들과 그 외의 합성단위들이 적절한 비율을 유지하도록 후보 합성단위 수를 줄이면 위의 장점을 유지할 수 있다.

IV. 실험결과 및 고찰

일본어 합성기를 개발하기 위해 구축된 음성 코퍼스는 방음된 녹음실에서 전문 여성 아나운서에 의해 녹음되었고, 녹음을 위해 사용된 대본은 뉴스기사, 소설, 대화체문장 및 숫자, 알파벳, 인터넷 주소 (URL) 등으로 구성하였다. 녹음된 음성 코퍼스는 표 4와 같다.

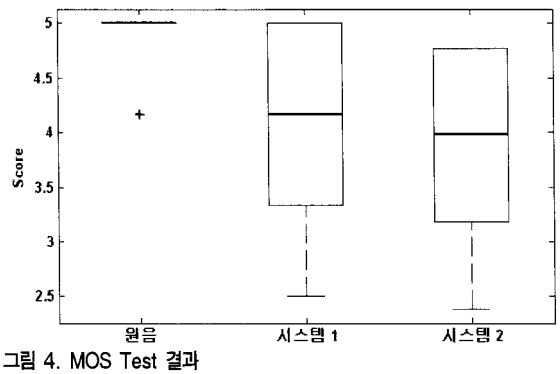


그림 4. MOS Test 결과
Fig. 4. Result of MOS Test.

표 4. 음성 코퍼스
Table 4. Speech corpus.

성별	녹음시간	개수			
		문장	IP	AP	음소
여성	41.04	17230	35871	142061	1104450

시스템의 성능을 평가하기 위해 합성음의 MOS (Mean Opinion Score) 테스트를 수행하였다. 테스트는 일본인 여성 5명이 참가하였고, 테스트 문장은 JEITA 종합평가문장 [7] 중 127문장을 선택하여 실험하였다. MOS 테스트는 원음 127개와 유동 break를 사용한 시스템 1과 사용하지 않은 시스템 2로 생성한 합성음 254개를 섞어 불규칙한 순서로 청취하고 5개의 레벨 (1~5, Bad, Poor, Fair, Good, Excellent) 중 하나를 선택하도록 하였다.

V. 결 론

보다 자연스러운 운율을 구현하는 것은 모든 음성 합성기의 공통된 목표로 합성음의 자연성이 운율에 의해 결정되기 때문이다. 코퍼스 기반 합성기는 음성 DB에 이미 다양한 운율 정보를 저장하고 있어 이를 효율적으로 이용한다면 충분히 자연스러운 운율을 구현할 수 있지만 합성기에서 생성하는 운율이 제한적이고 이것을 이용하여 합성단위를 선택함으로써 자연스러운 합성음을 얻기 힘들어진다.

본 논문에서는 합성음의 자연성을 향상시키기 위해 코퍼스 기반 일본어 합성기에서 생성된 운율을 보다 효율적으로 이용하여 합성단위를 선택하는 방법을 제안하였다. 운율정보의 하나인 break를 고정 break와 유동 break로 나누어 음성 DB의 각 세그먼트가 가지는 다양한 운율정보를 이용할 수 있는 합성단위 검색을 수행하였다. 실험 결과 제안한 방법으로 보다 자연스러운 합성음을 얻을 수 있었다.

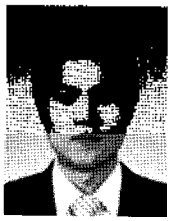
참 고 문 헌

1. R. E. Donovan, *Trainable speech synthesis, PhD. Thesis*, (Cambridge University, Engineering Department, 1996) pp.1-28.
2. J. Venditti, *Japanese ToBI labeling guidelines*, (OSU Working Papers in Linguistics,1997) pp. 127-162
3. 전성용, *일본어의 발음과 악센트*, (1st ED, Japanese Technical

Publishing Company, 2002) pp. 5-11
4. 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. ICSLP, 3, 314-317, 2000.
5. 나덕수, 이종석, 김종국, 배명진, "일본어 합성기에서 악센트 정보가 결합된 발음기호를 이용한 Break 예측 방법," 대한음성학회, 말소리, 62, pp.69-84, 2007.
6. 나덕수, 민소연, 이광형, 이종석, 배명진, "일본어 악센트 특징을 이용한 합성단위 선택 기반 일본어 TTS의 후보 합성단위의 사전선택 방법," 한국음향학회지 26-4, pp.159-165, 2007.
7. Technical Standardization Committee on Speech Input/Output Systems, "Speech Synthesis System Performance Evaluation Methods," JEITA IT-4001, 42-45, 2003

저자 약력

• 송 영 환 (YoungHwan Song)



2007년 2월: 숭실대학교 정보통신전자공학부 (공학사)
2007년 3월~현재: 숭실대학교 전자공학과 (석사과정)

• 나 덕 수 (Deok-Su Na)

현재: (주)보이스웨어 연구원
한국음향학회지 제26권 제4호 참조

• 김 종 국 (Jong-Kuk Kim)

현재: 숭실대학교 정보통신전자공학부 소리공학연구소
한국음향학회지 제23권 제1호 참조

• 배 명 진 (Myung-Jin Bae)

현재: 숭실대학교 정보통신전자공학부 교수
한국음향학회지 제21권 제3호 참조

• 이 종 석 (Jong-seok Lee)

현재: (주)보이스웨어 부사장
한국음향학회지 제26권 제4호 참조