일본어 합성기에서 악센트 정보가 결합된 발음기호를 이용한 Break 예측 방법

나덕수(보이스웨어), 이종석(보이스웨어),
김종국(숭실대), 배명진(숭실대)

<차 례>

< Abstract >

Break Predicting Methods Using Phonetic Symbols Combined with Accents Information in a Japanese Speech Synthesizer

Deok-Su Na, Jong-Seok Lee, Jong-Kuk Kim, Myung-Jin Bae

Japanese is a language having intonations, which are indicated by the relative differences in pitch heights and the accentual phrases (APs) 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, 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 which can enhance the quality of synthesized speech by reducing the errors in predicting break indices (BI), are proposed. The method is to use a new definition for the phonetic symbols, which combine the phonetic values of Japanese words with the accents information. Since a stream of defined phonetic symbols includes the information on the changes in intonations, the BI can be easily predicted by dividing the intonation phrase (IP) into several APs. As a result of an experiment, the accuracy of break generations was 98% and the proposed method contributed itself to enhance the naturalness of synthesized speeches.

* Keywords: Phonetic symbols, Break indices (BI), Unit selection, TTS system.

1. 서 론

현재 상용화되거나 연구되고 있는 음성합성 기술 중 합성음의 음질이 가장 우수한 것은 대용량 음성 코퍼스를 이용한 합성단위 선택(unit selection) 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.

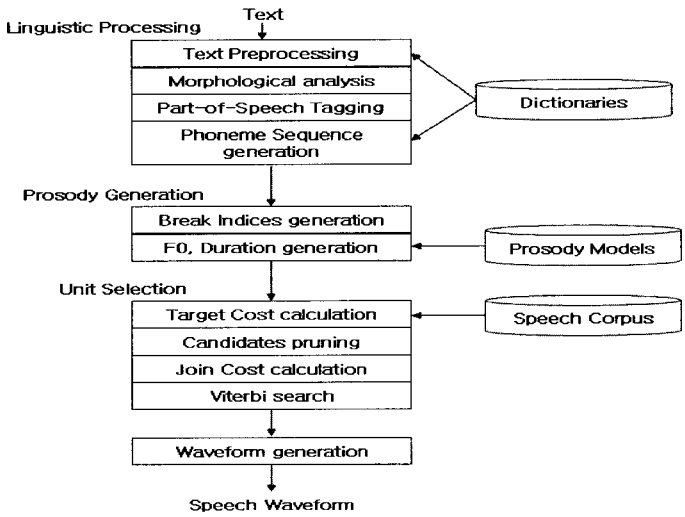
이러한 음성합성 시스템의 합성단위 선택 과정은 문맥 정보와 운율 파라미터에 의해 결정되는데 보다 자연스러운 합성음을 얻기 위해서는 정확한 운율 모델링이 필수적이다. 입력 텍스트에서 운율 파라미터를 생성하기 위해서는 억양구 경계 결정, 음소 지속시간 결정, 기본주파수의 윤곽선 설정의 3가지의 기본적인 모듈이 필수적이다[1].

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 Tree) 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 CART와 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 경계 정보에는 위치 정보 뿐 아니라 인접 AP와의 연결 정보도 포함된다. 일본어의 이러한 AP 연결정보는 문장의 의미나 구조에 의해, 연결되는 것과 끊어지는 것으로 나누어질 수 있고, 이것들의 정확한 구현에 의해 합성음의 자연성 및 명료성을 향상시킬 수 있다.

AP의 경계를 결정하기 위해서는 우선 악센트를 표현하여야 하는데, 일본어의



<그림 1> 일본어 TTS 시스템

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 규칙합성기 방식에 적합한 표현 방법이다.

본 논문에서는 일본어의 악센트 정보를 각 음절마다 아라비아 숫자로 부여하여 음소열로 악센트의 높낮이 변화를 나타낼 수 있는 새로운 발음기호를 정의하고, 이것을 이용하여 AP 경계 정보를 포함한 간단한 BI(Break Indices)를 결정하는 방법을 제안한다.

2장에서는 제안한 방법으로 구현한 음성합성 시스템에 대하여 설명하고, 3장에서는 제안한 발음기호의 구성에 대해 나타내고 4장에서는 BI 생성 방법에 대해 설명한다. 5장에서는 제안한 방법에 대한 실험 및 결과를 보이고, 6장에서 결론을 내린다.

2. 일본어 TTS 시스템

<그림 1>은 제안한 TTS 시스템의 구성도이다. 대부분의 합성단위 선택 기반 연결 합성기처럼 제안한 시스템도 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.

2.1 언어 처리

일본어 TTS의 언어 처리 모듈은 기호, 숫자 등을 변형하는 텍스트 전처리와 문장 분석 및 품사(Part of Speech, POS) 태깅을 이용하여 본 논문에서 제안하는 발음기호로 변환하는 태거/발음 변환으로 구성된다.

전처리에서는 일반 숫자, 소수, 분수, 음수, 날짜, 시간, 통화, 전화번호, 스코어, 주소, 수식, URL, e-mail, 기호, 영어 단어 등을 처리한다. 또한, 일본어에서 숫자나 기호의 형태론적 패턴만을 고려하지 않고, 앞 뒤 문장의 의미론적 패턴을 고려하여 모호한 문장을 처리한다. 예를 들어, "10:30で勝利した。(10대30으로 승리하였다)"와 "10:30に退勤した。(10시30분에 퇴근하였다)"를 구별하여 각각 스코어와 시간으로 처리한다. 알파벳 처리의 경우도 약어 사전과 영일(English to Japanese) 변환 사전을 구성하여 사용하고, 사전에 없는 단어 중 일본어 가타카나(Katakana)로 대응 가능한 단어는 가타카나로 변형하고, 나머지 단어는 오토마타(Automata)를 사용하여 변환한다.

발음 변환에 사용되는 사전은 임의의 글자로부터 시작하는 모든 단어를 한 번에 검색할 수 있는 형태로 구성하였는데, 그 이유는 어절의 구분이 없이 한 문장이 하나의 단어처럼 모두 붙어서 입력되는 일본어의 특징 때문이다. 그리고 태깅 알고리즘은 확률모델을 사용하고, Viterbi 검색 알고리즘으로 최적의 결과를 찾는다. 동사나 형용사의 경우 사전에는 어근만 등록하고 태깅 이전에 활용규칙에 따라 어미 활용을 한다. 또한, 일본어는 같은 한자(kanji)라도 품사도 다르고 읽는 방법도 다른 경우가 많아, 형태론적으로 구별이 가능한 경우에는 예외 규칙을 사용하여 확률 모델의 오류를 보정하였다. 발음 및 악센트를 생성하기 위해 기본적으로 모든 단어의 발음과 악센트는 사전에 등록하지만, 같은 한자라도 의미나 문장 형태에 따라 다르게 읽는 경우가 많고, 특히 악센트의 경우 사전에 등록된 형태보다 앞 뒤 단어에 따라 바뀌는 경우가 많이 나타나기 때문에 분석에 의해 발음과 악센트는 바뀔 수 있다. 명사+접미사의 경우는 접미사의 악센트에 따라 앞 명사의 악센트가 바뀌고, 수사+수단위 접미사의 경우도 수사의 발음 및 악센트가 수단위 접미사에 의해 바뀐다. 복합명사의 경우는 보통 한 명사처럼 악센트가 바뀌지만, 고유명사나 특정 명사의 경우는 악센트를 바꾸지 않는 경우도 있다. 동사나 형용사의 경우는 몇 가지 악센트 유형으로 나뉘고 각 악센트 유형별로 악센트가 바뀌게 된다. 동사+동사의 경우도 하나의 동사처럼 악센트가 바뀌는 경우도 있다. 사전에 없는 가타카나의 경우 보통 고유명사나 외래어의 표현일 가능성이 많은데,

이 경우는 가타카나의 형태에 따른 규칙을 사용한다.

본 논문에 사용된 발음 변환 모듈은 20만 단어 이상이 등록된 사전을 이용하여 단어의 발음과 악센트 정확도를 측정하였을 때 98% 이상의 정확도를 얻을 수 있었다.

2.2 운율 생성

운율 생성 모듈은 break 인덱스, 기본주파수, 음소 지속시간 생성으로 구성된다. Break는 형태소 단위의 단어들의 경계정보로 정의하여 6가지 종류로 구분하였고, 언어처리의 결과인 발음기호 열과 품사정보 등을 이용하여 생성한다. 보다 자세한 내용은 3장에서 다루기로 한다.

기본주파수 및 음소 지속시간 파라미터는 5시간 정도의 음성 코퍼스를 기반으로 제작된 CART를 사용하여 생성된다. 기본주파수 파라미터 생성을 위한 CART는 음절의 경계 형태별로 구성하였고, 훈련(training)에 사용된 feature는 인접한 3개 모음의 종류와 각각의 음절 형태(syllable type), 톤 형태(tone type), 단어에서 음절의 위치, 그리고 IP에서의 단어 위치 등이다. 특히 톤 형태는 ToBI 시스템에서와 같이 실제 음성데이터에서 추출한 피치정보를 사용한 것이 아니라 제안하는 발음기호로 표현된 발음기호 열로부터 추출하여 사용하였다. 음소지속시간 파라미터 생성을 위한 CART는 음소의 형태에 따라 구성하였는데, 모음에 대해서는 장음과 단음을 구분하였다. 훈련에 사용된 feature는 인접한 3개 음소의 종류, AP에서 음소의 위치, 단어를 구성하는 음절의 수, 단어에서의 음절 위치, 그리고 IP에서 단어의 위치 등이다.

TTS 시스템에서는 위의 CART를 이용하여 각 음소의 목표 경계 피치와 음소 지속시간을 생성하여 합성단위 선택 과정에서 이용한다.

2.3 합성단위 선택

제안한 시스템의 음성 코퍼스는 문맥 기반(context-based) clustered tree로 구성된 음성 세그먼트로 구성되어있고, 각 음성 세그먼트들은 음성파형과 경계 피치, 에너지, 스펙트럼 정보로 이루어진다. 합성단위 선택 과정은 음성 코퍼스와 언어처리 및 운율처리에서 얻어진 문맥정보, 피치, 지속시간 정보 등을 이용하여 수행되어진다. 먼저 문맥 정보를 이용하여 후보(candidate) 합성단위들을 추출하게 된다. 그리고 이렇게 추출된 후보들에 대해 목표 비용(target cost)을 계산하여 후보 합성단위들의 사전선택(pre-selection)을 수행한다. 대용량 코퍼스를 이용하는 합성기에서는 문맥 정보가 동일한 후보가 많아 모든 후보로 Viterbi 검색을 수행한다면 실시간 합성이 힘들어지기 때문에 합성음의 음질 열화를 최소로 할 수 있는 효율적

인 사전선택방법이 반드시 필요하다[4]. 제안한 합성기에서는 사전선택에 의한 음질 열화를 최소로 하기위해 CCL(Connected Context Length)과, 일본어의 특징을 고려한 방법을 수행한다[5]. 사전선택된 후보들에 대해서는 연결비용(join cost)을 계산하여 위에서 계산된 목표 비용과 합하여 Viterbi 검색을 수행한다.

2.4 파형 생성

합성단위 선택 과정에서 Viterbi 검색을 수행하면 입력된 텍스트에 대한 최적의 합성단위들을 순서적으로 얻을 수 있는데, 합성음은 선택된 각각의 합성단위의 음성파형을 연결하여 생성한다.

3. 제안하는 일본어 발음기호

본 논문에서 제안하는 발음기호는 음가(phonetic value)를 표현하는 알파벳과 악센트를 표현하는 아라비아 숫자의 조합으로 구성된다. 음가를 나타내는 알파벳은 일본어 합성기 성능 평가에 대한 표준인 JEITA IT-4001(구 JEIDA G24 2000)[6][7]을 참고로 하여 제안한 시스템의 운율 생성 및 합성단위 선택에 쉽게 적용할 수 있도록 변경하였다. 본 논문에서 사용한 합성기는 기본적으로 폰을 합성단위로 사용하고, tri-phone을 기반으로 합성단위 선택을 수행한다. 따라서 제안하는 기호는 최대한 음소의 기본 정보를 많이 표현할 수 있으면서도 인접한 좌, 우 음소와 서로 정보를 융합하고 분리하는 것이 용이하도록 설계하였다.

3.1 모음과 반모음

제안하는 발음기호의 모음 표현에 대한 큰 특징은 장음과 반모음이다. 일본어는 장음과 단음이 확실히 구분되는 특징이 있고, 장음은 단음에 비해 평균적으로 2배의 지속시간을 나타낸다. 연결형 합성기에서 합성 단위의 불연속(discontinuity)에 의해 합성음의 음질 열화가 발생하는데, 특히 모음과 모음이 연결될 때 자주 발생한다. 또 음성 코퍼스를 제작할 때 음성의 각 세그먼트를 분리하는데 일반적으로 음성인식에서 사용하는 labeler를 사용한다. 그런데, 이러한 labeler의 경우 동일한 모음이 연속되는 부분에서 오류가 많이 발생하는 단점이 존재한다. 일본어의 장음도 같은 모음이 반복되는 형태이므로 이것을 2개의 단음으로 분리하여 처리한다면 많은 labeling 오류가 발생할 수 있을 뿐 아니라 그러한 오류의 수정이 없다면 합성음에서 불연속에 의해 문제가 발생할 수 있다. 따라서 제안하는 발음기호에서는 장음에 대해서 하나의 기호를 할당하였다. 일본어의 모음에는 /a/, /i/, /u/,

/e/, /o/가 있는데, 이것들의 장음으로 /aa/, /ii/, /uu/, /ee/, /oo/를 추가 하였다.

일본어의 반모음은 /w/, /y/가 있는데, 이것들은 뒤에 모음과 결합하여 사용되지만 평균적으로 모음의 길이보다 짧은 지속시간을 가지는 특징이 있다. 반모음을 하나의 독립된 폰으로 사용할 경우 짧은 지속시간으로 정확한 labeling이 어려울 뿐만 아니라 합성단위 선택 과정에서도 합성단위 수가 증가하고 합성음에서도 연결되는 부분이 늘어나게 된다. 따라서 이러한 문제를 줄이기 위해 반모음은 뒤에 오는 모음과 결합하여 하나의 폰으로 나타내었다. 즉 반모음에 대해서는 음절 형태를 합성단위로 사용하여 /wa/, /wi/, /we/, /wo/과 /ya/, /yu/, /ye/, /yo/로 나타내고, 여기에 위에서 정의한 장음에 대한 /waa/, /wii/, /wee/, /woo/ 및 /yaa/, /yuu/, /yee/, /yoo/로의 확장을 추가하였다. (/w/와 모음 /u/의 결합, /y/와 /i/의 결합은 사용하지 않았다.)

3.2 자음

일본어의 자음에 대한 표현은 기존의 다른 발음기호들과 큰 차이는 없고, 16개로 표현하였다.

<표 1> 정의된 자음

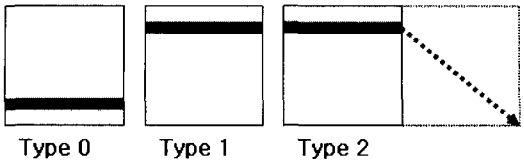
voice stop	/b/, /d/, /g/
unvoice stop	/p/, /t/, /k/
voiced fricative/affricate	/z/, /j/
unvoiced fricative	/h/, /s/, /sh/
unvoice affricate	/ch/, /ts/
liquid	/r/
nasal	/m/, /n/

3.3 일본어 특수 Phoneme (ん과 っ)

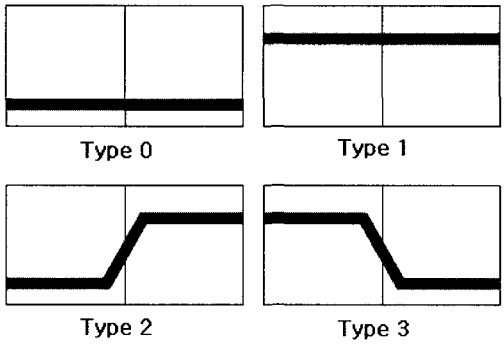
일본어에서만 나타나는 특별한 발음에는 ん(발음; 撥音, syllabic nasal)과 っ(촉음; 促音, choked sound)이 있는데[8], ん는 /xN/으로 っ는 /lt/로 나타내었다. /xN/은 단독으로 하나의 발음기호를 형성하므로 악센트 정보를 가질 수 있도록 하였으나, 장음에 대한 확장은 적용하지 않았다.

3.4 악센트 정보가 결합된 발음기호

제안한 발음기호는 위에서 표현한 모음들과 ん의 음가에 악센트 정보를 결합한 형태인데, 악센트 표현은 단음과 장음을 구분하였다. 새로 정의된 장음은 하나의 폰에서 악센트의 높낮이 변화가 발생할 수 있으므로 변화가 발생하지 않는 단



<그림 2> 단음의 악센트 형태



<그림 3> 장음의 악센트 형태

음과 차이를 두었다. 단음의 악센트 정보는 3가지, 장음은 4가지로 구분하였다.

<그림 2>는 단음에 대한 악센트 형태를 나타낸 것으로 사각형은 단음의 길이 또는 한 개의 박(mora, 모음의 길이)을 나타내고, 굵은 선은 악센트(피치레벨)의 높이를 나타낸다. Type 0은 낮은 악센트, type 1은 높은 악센트를 나타내고, type 2는 type 1과 같이 높은 악센트이지만 뒤의 모음에서 악센트가 낮아지는 특성을 가지는 형태이다. 단음에 대한 악센트 type은 /a/, /i/, /u/, /e/, /o/, /wa/, /wi/, /we/, /wo/, /ya/, /yu/, /ye/, /yo/, /xN/와 결합한다. 즉, 단음 /a/는 악센트 정보가 결합되면 /a0/, /a1/, /a2/로 확장된다.

<그림 3>은 장음에 대한 악센트 type의 형태를 나타낸 것이고, <그림 2>와 마찬가지로 하나의 사각형은 하나의 박을 나타내는데, 장음은 단음에 비해 2배의 박으로 나타낼 수 있다. Type 0과 1은 단음과 마찬가지로 각각 낮은 악센트와 높은 악센트를 표현하고, type 2와 3은 악센트의 변화가 나타나는 형태를 뜻한다. Type 2는 낮은 악센트에서 높은 악센트로 변화하는 장음이고, type 3은 높은 악센트에서 낮은 악센트로 변화하는 장음이다. 장음에 대한 악센트 type은 /aa/, /ii/, /uu/, /ee/, /oo/, /waa/, /wii/, /wee/, /woo/, /yaa/, /yuu/, /yee/, /yoo/와 결합되는데, 예를 들어 장음 /aa/는 악센트 정보가 결합되면 /aa0/, /aa1/, /aa2/, /aa3/로 확장된다.

제안하는 악센트 정보를 결합한 전체 발음기호는 211개의 폰으로 구성되는데, 7가지 악센트 형태와 13가지의 단음 및 장음이 결합한 91개의 모음, 악센트 정보

가 결합된 3개의 ん에 대한 발음, 16개의 자음 그리고 っ에 대한 발음으로 이루어진다.

4. Break 예측 방법

코퍼스 기반 연결 합성기에서 break는 합성단위 선택 과정에서 문맥정보와 목표 기본주파수 및 목표 음소 지속시간 등과 함께 중요하게 사용되는 운율 정보로, 텍스트를 읽을 때 발생하는 발성 패턴에 관한 정보이다. 즉, 여러 문장을 읽은 녹음 데이터를 살펴보면, 의미나 문장구조 또는 문장 길이에 의해 단숨에 읽어내는 정도가 각각 다르게 나타나고, 어느 부분은 텍스트에 쉼표(comma)가 없어도 포즈가 나타나고, 그 포즈의 길이도 경우에 따라 변하는 것을 알 수 있다. 이처럼 자연스러운 발성에서는 다양한 break가 존재하고 그로 인해 정보의 명확한 전달이 이루어진다.

일본어 합성기의 break는 악센트에 의해 영향을 받기 때문에 정확한 악센트 추정이 필수적이고, 악센트 정보만으로 간단한 예측이 가능하다. 일본어는 하나의 IP가 몇 개의 악센트 구로 구성되는데[2], 악센트 구는 악센트가 형성되는 단위로, 하나의 악센트 구에는 하나 이하의 악센트 핵(높은 악센트에서 낮은 악센트로 변화가 발생하는 부분)이 반드시 나타난다[6]. 그리고 하나의 악센트 구는 단 숨에 읽는 단위이므로 일본어 합성기에서는 악센트 구를 분리함으로써 간단한 break를 구현할 수 있다.

4.1 Break의 정의

본 논문에서는 악센트 정보가 포함된 발음기호를 사용하여 간단히 악센트 구를 분리하고 품사 및 의미론적인 규칙을 사용하여 break를 세분화하는 방법을 제안한다. 먼저, 다음부터 사용할 break의 의미를 단어와 단어 사이의 연결 정보로 제한하고, 단어 또한 언어처리의 POS(part of speech) 태깅 결과에 의한 분리 가능한 최소단위로 그 의미를 한정한다.

제안한 합성기의 break 인덱스는 단어와 단어 사이의 연결 정보로써 앞 단어의 마지막 폰의 연결 정보이고, 0~5까지 6종류로 표현하였다. 0은 break가 없는 것으로 2개의 단어를 하나의 단어로 연결하고자 할 때 사용하고, 1은 동일한 AP를 구성하는 단어와 단어 사이의 경계를 나타낸다. 2와 3은 AP가 분리된다는 것을 나타내는 것으로, AP와 AP 사이에 포즈 없이 연결되는 경우를 2로 나타내고, 그렇지 않고 두 AP사이에 포즈가 존재하는 경우를 3으로 나타낸다. 여기서 포즈는 IP와 IP사이에 나타나는 포즈보다 짧은 것으로, 합성음 생성에서는 50 msec의 포즈

를 사용하였다. 4와 5는 각각 IP 또는 문장이 분리되는 것을 나타낸다.

<표 2>는 break와 AP 경계에 대한 관계를 나타낸 것이다. Break 0과 1은 AP 경계가 아니고, 2는 포즈가 없이 두 개의 AP가 연결되는 것을, 3은 두 개의 AP 사이에 포즈가 존재하는 것을 각각 나타내는 것으로, AP측면에서 break 2는 두 개의 AP가 서로 연결되어 있지만, 3은 서로 연결되어 있지 않음을 의미한다. 4와 5도 3과 마찬가지로 AP가 서로 연결되어 있지 않은 상태를 나타낸다.

<표 2> Break 인덱스와 AP 경계 정보

break 인덱스	AP 경계
0 (하나의 단어)	하나의 AP
1 (하나의 AP안의 단어의 경계)	하나의 AP
2 (포즈없이 연결되는 AP 경계)	연결되는 AP와 AP 경계
3 (포즈로 연결되는 AP 경계)	분리되는 AP와 AP 경계
4 (IP와 IP 경계)	분리되는 AP와 AP 경계
5 (문장 경계)	분리되는 AP와 AP 경계

4.2 Break 예측

제안한 합성기에서는 먼저, 3장에서 설명한 악센트 정보가 포함된 새로운 발음기호와 품사 및 문장기호 정보로 break를 생성하고, 후처리를 통해 이를 변경한다. (Break 0은 POS 태깅 에러를 위한 것이므로 따로 설명하지 않는다.) Break 1과 2를 결정하기 위해서는 우선 AP를 결정해야 하는데, 이를 위해서는 2가지 규칙이 필요하다. 첫째는 하나의 AP에서 2개 이상의 악센트 핵이 존재할 수 없다는 것이고, 둘째는 조사나 접미사와 같이 상호작용에 의해 악센트가 결정되는 품사의 단어사이에는 AP의 경계가 형성되지 않는다는 것이다.

<표 3>은 실제 입력텍스트 [富士山は日本で一番高い山です。](후지산은 일본에서 가장 높은 산이다.)의 break의 분석을 나타낸 것이다. 첫 번째 열의 "텍스트"는 언어처리 결과에 의한 단어를 나타낸 것이고, 2번째 열은 각 단어의 발음을 3장에서 설명한 새로운 발음기호로 표현한 것이다. 우선 이 문장에서 /は/, /で/, /です/는 앞 단어와 결합하면서 악센트가 형성되는 품사의 단어이므로 이 단어들 앞에는 AP의 경계가 올 수 없다. 따라서 /は/는 /富士山/과 결합되어 뒤의 /日本/과의 사이가 AP 경계를 이루는지 분석되어진다. 앞 단어들, /富士山は/의 발음은 [hu2ji0sa0xN0wa0]이고, /日本/의 발음은 [ni0ho2xn0]인데, 이 경우 앞 단어들에도 악센트 핵이 존재하고 뒤 단어에도 악센트 핵이 존재하므로 두 단어 사이에는 AP의 경계가 형성된다. /日本で/와 /一番/ 사이에도 마찬가지로 한 번 내려간 악센트는 동일한 AP에서는 다시 올라가지 않는 특징을 적용하면 두 단어 사이도 AP경계가 형성된다. 위와 같은 방법으로 /一番/과 /高い/, 그리고 /高い/와 /山です/도 서로 다

른 AP로 분석된다. 그리고 /。/는 문장종료 기호이므로 /です/의 break는 5가 되고, 쉼표 또는 다른 문장 기호가 존재한다면 이에 해당하는 break로 설정한다.

<표 3> 입력텍스트의 분석 결과 (입력텍스트: [富士山は日本で一番高い山です。])

텍스트	발음기호	예측된 break	후처리된 break
富士山	[hu2ji0sa0xN0]	1	1
は	[wa0]	2	3
日本	[ni0ho2xn0]	1	1
で	[de0]	2	3
一番	[i0chi1ba1xn1]	2	2
高い	[ta0ka2i0]	2	2
山	[ya0ma2]	1	1
です。	[de0su0]	5	5

<표 4> Break 2와 3을 결정하는 후처리 규칙 예

조사 は	(と)+は+おもう (~라고 생각한다)의 경우		break 2
	(と)+は+裏腹に, 裏腹で (~와는 정반대로)		
	と+は+형용사		
	그 외의 경우		break 3
조사 が	ところが (~했는데, ~했더니)		break 3
	조사 が가 2개 이상일 때 앞의 것		
	문장에 は 없이 が가 주격 조사일 때		
	그 외의 경우		break 2
그 외 조사의 기본처리	조사 뒤에 설명 구조	조사+명사+동사	break 2
		조사+동사+동사, 조동사, 보조용언	
		조사+형용사 보조동사	
	조사 뒤에 수식 구조	조사+특정 부사(また, ...)	break 3
특수 상황 처리	を가 타동사의 목적격 조사인 경우		break 2
	を뒤에 숫자+접미사가 오는 경우		break 3
	조사 の뒤는 기본적으로		break 2
	조사 の뒤에 대명사가 오는 경우		break 3
	조사 の가 숫자를 읽을 때 사용되는 경우		break 3
	조사 に뒤에는 기본적으로		break 2
	조사 に의 관용적 표현에 대한 처리:間(あいだ)+に		break 3
	に가 시간을 나타낼 때		break 3

위의 AP분석을 수행하면 <표 3>의 3번째 열과 같이 AP 경계로 분석된 단어의 break는 2가 된다. break 2는 의미론적 또는 형태론적인 규칙을 적용하는 후처리를 통해 break 3으로 수정될 수 있다. 이러한 규칙에 대해 예를 들면, 조사 /は/가 문장의 첫 단어와 결합하면 break 3으로 변경하는 규칙이 있다. Break 3은 break 2를 변경하여 생성하는 경우가 대부분이나 1을 변경하거나, 문장기호에 의해 생성될 수도 있다. <표 3>의 4번째 열은 후처리된 break를 나타낸다.

후처리 규칙은 문법적 지식이나 관용어적인 사용 등을 고려한 것으로 매우 다양하고 복잡해 질 수 있다. <표 4>는 본 논문에서 사용한 몇 가지의 후처리 규칙의 구체적 예를 나타낸 것이다.

5. 실험 및 결과

5.1 제안한 발음 기호를 이용한 Break 예측 에러

제안한 break 인덱스 생성 방법의 성능을 평가하기 위해 JEITA 종합평가문장 [6] 중 107문장을 선택하여 실험하였다. JEITA 종합평가문장은 단문과 복문, 의문문과 평서문 그리고 다양한 수식 구조를 가지도록 구성된 문장으로 이루어져 있는데, 그 중 break 2 또는 break 3이 최소한 1번 이상 발생할 수 있는 텍스트를 선택하였다. 실험은 악센트 정보가 포함된 발음기호로 정확한 break를 생성할 수 있음을 알아보기 위한 것으로 부정확한 발음에 의한 영향을 제거하기 위해 실험 텍스트에 대한 발음변환 모듈의 오류를 보정하였다. 선택된 문장에 대해 언어처리를 수행하여 각 문장을 단어 별로 나누고 단어와 단어 사이에 적합한 break 정보들을 제안한 방법으로 생성하고, 이것을 수동으로 수정하여 오류를 측정하였다.

Break 인덱스를 생성한 결과, 실험에 사용된 텍스트는 1848개의 단어로 분석되었고, break 1이 50.7%로 가장 많이 나타나고, 그 다음이 break 2, break 3 순서로 많이 생성되었다. 그리고 수동으로 정확한 break 정보를 수집하기 위해 2년간 발음변환 모듈 개발에 참가한 일본인 여성 1인이 자동 생성된 break 인덱스를 수정하였다.

<표 5>는 break 생성에 대한 성능을 나타낸 것이다. Break 2와 3을 제외한 다른 break에서는 오류가 발행하지 않았는데, 이것은 발음과 언어처리의 오류를 보정함으로써 얻어진 결과로 해석되었다. 즉 정확한 발음 변환이 이루어진다면 break 2와 3을 제외한 다른 break는 아주 정확한 예측이 가능함을 알 수 있었다. 그리고 대부분의 에러는 break 2와 3에서 발생했는데 이것은 두 종류의 break를 정확히 예측할 수 있는 문장의 구조 및 의미론적인 분석이 포함된 규칙이 부족함을 나타낸다.

<표 5> Break 예측 오류(단어의 수)

Break 인덱스	0	1	2	3	4	5	전체
생성된 단어의 수	11	937	501	187	105	107	1848
오류 단어의 수	0	0	2	26	0	0	28(1.52%)

5.2 제안한 일본어 음성합성기의 Corpus 및 성능 평가

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 6>과 같다. 녹음시간은 발성의 중간 포즈만 남기고 처음과 끝의 포즈를 제거한 것이다[9].

<표 6> 전체 음성 코퍼스

성별	녹음시간 (hour)	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[6] 중 127문장을 선택하여 실험하였다. 합성음은 테스트 문장이 포함되지 않은, 3개의 서로 다른 크기의 코퍼스를 이용하는 시스템들을 구성하여 생성하였다. <표 7>에서 시스템 1은 <표 6>의 전체 음성 코퍼스에서 테스트 문장들을 녹음한 부분을 제외한 코퍼스를 사용한 합성기이고, 시스템 2와 3은 임베디드(embedded) 시스템을 위해 작은 크기로 제작한 코퍼스를 사용한 합성기이다. MOS 테스트는 원음 127개와 각각의 시스템으로 생성한 합성음 381개를 섞어 불규칙한 순서로 청취하고 5개의 레벨(1~5, Bad-Poor-Fair-Good-Excellent) 중 하나를 선택하도록 하였다.

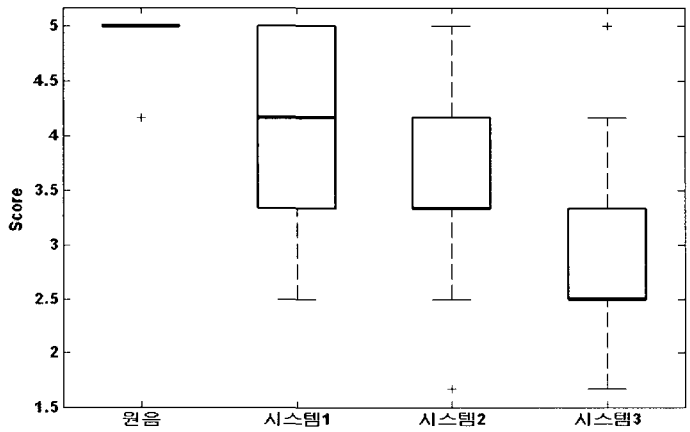
<표 7>은 각 시스템에서 사용된 코퍼스 크기를 나타낸 것으로, 시스템 1은 39.4시간, 시스템 2는 2.8시간, 시스템 3은 0.59시간의 녹음데이터를 사용하였다. <표 8> 및 <그림 4>를 보면, 테스트 결과 원음이 4.99의 MOS를 나타낼 때 시스템 1의 합성음은 4.2의 좋은 결과를 얻을 수 있었다. 그리고 시스템 1의 합성속도는 3.06 GHz Xeon CPU를 가진 PC에서 실시간의 약 150배 이상으로, 500 바이트 텍스트를 합성할 때 0.29초가 소요되는 성능을 나타내었다.

<표 7> 합성음 생성에 사용한 음성 코퍼스의 크기

	음성 코퍼스 크기		
	시스템1	시스템2	시스템 3
녹음시간 (hour)	39.4	2.8	0.59

<표 8> MOS 결과

원음	시스템1	시스템2	시스템3
4.99	4.20	3.54	2.93



<그림 4> MOS 테스트 결과

6. 결 론

보다 자연스러운 운율을 구현하는 것은 모든 음성 합성기의 공통된 목표로 합성음의 자연성이 운율에 의해 결정되기 때문이다. 코퍼스 기반 합성기는 음성 코퍼스에 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 합성기에 보다 적합한 발음기호를 정의하고 이것을 이용하여 운율을 생성하는 방법

및 음성합성 시스템을 제안하였다. 먼저 일본어 악센트의 변화 정보를 효율적으로 이용할 수 있도록 음가에 악센트 정보를 결합한 새로운 발음기호를 정의하였고, 이것을 이용하여 운율정보의 하나인 break를 생성하였다. 제안한 break 생성 방법으로 매우 정확한 break 정보를 얻을 수 있었고 우수한 성능의 일본어 음성합성 시스템을 구축할 수 있었다.

참 고 문 헌

[1] R. E. Donovan, *Trainable speech synthesis*, Ph.D. Thesis, Cambridge University, Engineering Department, pp. 1-28, 1996.

[2] J. Venditti, "Japanese ToBI labeling guidelines", *OSU Working Papers in Linguistics*, pp. 127-162, 1997.

[3] 전성용, *일본어의 발음과 악센트*, 1st Ed., pp. 5-11, Japanese Technical Publishing Company, 2002.

[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*, Vol. 3, pp. 314-317, 2000.

[5] D. S. Na, W. S. Jun, J. W. Lee, M. H. Cho, J. S. Lee, M. J. Bae, "A preselection method using accentual phrase matching in unit selection-based Japanese text to speech", *Proc. WESPAC IX*, pp. 124, 2006.

[6] Technical Standardization Committee on Speech Input/Output Systems, "Speech synthesis system performance evaluation methods", *JEITA IT-4001*, pp. 42-45, 2003.

[7] T. Kazuyo, A. Makoto, M. Toshimitsu, I. Shuichi, "JEIDA standard of symbols for Japanese text-to-speech synthesizers", *Proc. 3rd Oriental COCOSDA Workshop*, pp. 27-32, 2000.

[8] S. Narayanan, A. Alwan, *Text to speech synthesis: New paradigms and advances*, 1st Ed., pp. 155-173, New Jersey: Prentice Hall Professional Technical Reference, 2004.

[9] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, K. Tokuda, "Ximera: A new TTS from ATR based on corpus-based technologies", *Proc. ISCA 5th Speech Synthesis Workshop*, pp. 179-184, 2004.

[10] H. Fujisaki, S. Ohno, "Prosodic parameterization of spoken Japanese based on a model of the generation process of F0 contours", *Proc. ICSLP*, pp. 2439-2442, 1996.

[11] Y. Kitahara, "Development of high quality and intelligent speech synthesis technology", *Hitachi-Criticism(日立評論)*, Vol. 88, No. 6, pp. 60-65, 2006.

[12] http://www.oki.com/jp/RDG/JIS/oto/tts/

[13] http://www.cjk.org/cjk/

접수일자: 2007년 4월 9일
게재결정: 2007년 6월 21일

▶ 나덕수(Deok-Su Na) : 교신저자
주소: 133-120 서울시 성동구 성수동 2가 280-13 삼환디지털벤처타워 10층
소속: ㈜보이스웨어
전화: 02) 3016-8562
E-mail: dsna@voiceware.co.kr

▶ 이종석(Jong-Seok Lee)
주소: 133-120 서울시 성동구 성수동 2가 280-13 삼환디지털벤처타워 10층
소속: ㈜보이스웨어
전화: 02) 3016-8503
E-mail: jslee@voiceware.co.kr

▶ 김종국(Jong-Kuk Kim)
주소: 156-733 서울시 동작구 상도동 511 숭실대학교
소속: 숭실대학교 정보통신 전자공학부 소리공학연구소
전화: 02) 824-0906
E-mail: kokjk@ssu.ac.kr

▶ 배명진(Myung-Jin Bae)
주소: 156-733 서울시 동작구 상도동 511 숭실대학교
소속: 숭실대학교 정보통신 전자공학부 소리공학연구소
전화: 02) 824-0906
E-mail: mjbae@ssu.ac.kr