한국음향학회지 제26권 제4호 pp. 159~165 (2007)

일본어 악센트 특징을 이용한 합성단위 선택 기반 일본어 TTS의 후보 합성단위의 사전선택 방법

A Pre-Selection of Candidate Units Using Accentual Characteristic In a Unit Selection Based Japanese TTS System

나 덕 수*, 민 소 연**, 이 광 형**, 이 종 석*, 배 명 진***
(Deok-Su Na*, So-Yeon Min**, Kwang-Hyoung Lee**,
Jong-Seok Lee*, Myung-Jin Bae***)
*보이스웨어 기술연구소, **서일대학, ***숭실대학교 정보통신 전자공학부
(접수일자: 2007년 4월 6일, 수정일자: 2007년 4월 30일, 채택일자: 2007년 5월 16일)

본 논문에서는 합성단위 선택 (unit selection) 기반 일본어 합성기에 필요한 후보 합성단위들에 대한 사전선택 (pre-selection)의 새로운 방법을 제안한다. 일반적인 사전선택 방법은 하나의 억양구에서 음소 열에 대한 비용을 계산하여 이용하는 방법이다. 그런데, 일본어는 다른 언어와는 다르게 상대적인 피치의 높낮이로 나타나는 악센트를 가지는 언어이고, 몇 개의 단어가 하나의 악센트 구를 형성하는 특징이 있다. 또한 일본어의 운율은 악센트 구를 기본 단위로 하여 변화하는 특징이 있어서, 사전선택에서 이러한 악센트 구 단위의 운율 변화를 반영함으로써 음질을 향상시킬 수 있고, 악센트 구에서 음소 열에 대한 비용을 계산하여 억양구에서 하는 것보다 계산량을 줄일 수 있다. 제안한 방법은 일본어의 악센트 구를 정의하여 음소 열에서 이것을 분석하고, 각 악센트 구에서 합성 할 음소의 각 후보에 대해 CCL (Connected Context Length)을 구하는 악센트 구 매칭을 이용하여 사전선택을 수행하는 방법이다. 제안한 방법은 Voiceware의 합성기인 VoiceText를 baseline 시스템으로 사용하여 구현하였고, 인지적 에러 (억양 에러, 연결 에러)와 합성시간에 대해 평가 하였다. 실험 결과, 제안한 방법은 합성 음질을 보다 자연스럽게 향상시켰고, 합성 속도를 개선하였다.

핵심용어: 일본어 음성합성, 합성단위 선택, 사전선택

투고분야: 음성 처리 분야 (2.4)

In this paper, we propose a new pre-selection of candidate units that is suitable for the unit selection based Japanese TTS system. General pre-selection method performed by calculating a context-dependent cost within IP (Intonation Phrase). Different from other languages, however, Japanese has an accent represented as the height of a relative pitch, and several words form a single accentual phrase. Also, the prosody in Japanese changes in accentual phrase units. By reflecting such prosodic change in pre-selection, the quality of synthesized speech can be improved. Furthermore, by calculating a context-dependent cost within accentual phrase, synthesis speed can be improved than calculating within intonation phrase. The proposed method defines AP, analyzes AP in context and performs pre-selection using accentual phrase matching which calculates CCL (connected context length) of the phoneme's candidates that should be synthesized in each accentual phrase. The baseline system used in the proposed method is VoiceText, which is a synthesizer of Voiceware. Evaluations were made on perceptual error (intonation error, concatenation mismatch error) and synthesis time. Experimental result showed that the proposed method improved the quality of synthesized speech, as well as shortened the synthesis time.

Key words: Japanese Speech synthesis, Unit selection, Pre-selection

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

현재 널리 사용되고 있는 합성 기술로는 대용량 코퍼스 기반 합성 시스템의 하나인 합성단위 선택 (unit selection) 기반 합성 시스템이 있다. 고음질의 합성음을 생성할 수 있는 합성단위 선택 기반 음성 합성 시스템은 자연스러운 합성음을 얻기 위해서 보다 큰 음성 코퍼스가 필요하다 [1].

합성단위 선택 과정은 동적 프로그래밍 (Viterbi) 알고리듬을 사용하여 수행되는데, 먼저 전체 합성음에서 특정 위치의 후보로 선택된 합성단위에 대한 목표 비용과, 각각의 후보 합성단위 사이의 접합에 대한 비용을 계산하여 후보들의 최적의 순서를 찾고, 찾아진 최적의 후보들은 합성 과정에서 이용된다 [2].

이러한 음성 합성 시스템에서는 음절, 음소나 변이음과 같은 작은 합성 단위를 사용하는데, 합성의 단위가 작아짐으로써 합성단위 선택 시 후보의 수가 크게 증가되어 후보의 수를 줄여 주는 사전선택 과정이 필요하게 된다 [3].

사전선택은 복잡한 코스트를 계산하기 전에, 합성단위 후보들 전체에 대해 간단하고 빠른 코스트 계산을 수행하여 그 수를 최적의 n개로 줄여주는 과정이다 [2]. 일반적인 사전선택 방법은 억양구 (intonation phrase, IP) 내에서 음소 열에 관한 코스트를 계산하여 수행된다. 그러나, 일본어는 악센트가 존재하고, 악센트 구 (accentual phrase, AP) 단위로 운율이 변화하는 특징을 가진 언어이므로 일본어 합성기는 이러한 특징을 반영할 수 있는 사전선택 방법이 필요하다. 본 논문에서는 IP가 아닌 AP 단위로 수행하는 사전선택 방법을 제안한다.

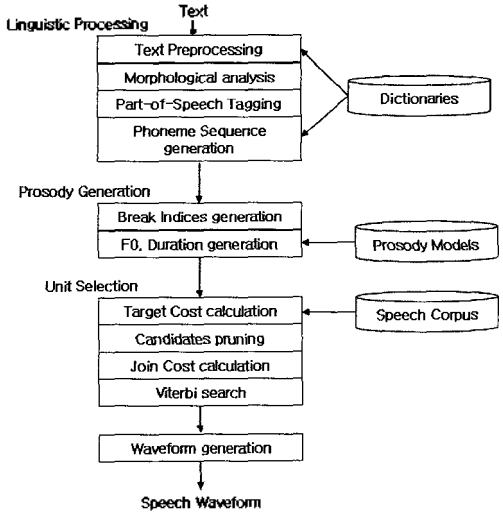

그림 1. 일본어 VoiceText 시스템
Fig. 1. The Japanese VoiceText System.

제안한 방법을 실험하기 위해서, Voiceware의 TTS 시스템인 VoiceText를 이용하였다. 이 시스템은 IP 내에서 음소의 연결 정도를 이용하여 합성단위 후보들에 대한 사전선택을 수행하는 상용 소프트웨어로써, 영어 (U.S), 중국어 (Mandarin), 한국어, 일본어의 4개 언어를 지원한다. II장에서 VoiceText의 일본어 합성기의 특징을 설명한다.

II. 일본어 VoiceText System

그림 1은 일본어 VoiceText 시스템의 구성도이다. 대부분의 합성단위 선택 기반 연결 합성기처럼 4가지의 부분 (linguistic processing, prosody generation, unit selection, waveform generation)으로 구성되고, 합성의 기본 단위로 phone을, text code로 일본어 Shift-JIS를 사용하였다.

2.1. 언어 처리

일본어 TTS의 언어 처리 모듈은 기호, 숫자 등을 변형하는 텍스트 전처리와 문장 분석 및 품사 태깅을 이용하여 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.

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.

2.2. 운율 생성

운율 정보로는 break, 기본주파수 (F_0), 음소 지속시간을 사용한다. break는 형태소 단위의 단어들의 경계정보로 정의하고 6가지 종류로 구분하여, 언어처리의 결과인 발음기호와 품사정보 등을 이용하여 생성한다.

기본주파수 및 음소 지속시간 정보는 5시간 정도의 음성 코퍼스를 이용하여 CARTs (Classification and Regression Trees)를 구성하고, TTS 시스템에서는 이것을 이용하여 목표 세그먼트의 경계 피치와 지속시간을 생성하여 합성단위 선택 과정에서 이용한다.

2.3. 합성단위 선택 (Unit Selection)

VoiceText 시스템의 음성 코퍼스는 컨텍스트 기반 clustered tree 형태의 음성 세그먼트로 구성되어있고, 각 음성 세그먼트들은 파형과 경계 피치, 에너지, 스펙트럼으로 이루어진다. 합성단위 선택 과정은 음성 코퍼스와 언어처리 및 운율 생성부에서 얻어진 문맥 (context) 정보, 피치, 음소 지속시간 등을 이용하여 수행되어진다. 먼저 tri-phone 문맥 정보들을 이용하여 후보 합성단위들을 추출하는데, tri-phone 문맥 정보들이 일치하는 후보가 없는 경우 문맥 정보의 종류를 줄여가면서 후보 합성단위를 추출하고, 추출된 후보들에 대해 목표 비용을 계산하여 사전선택을 수행한다. 대용량 코퍼스를 이용하는 합성기에서는 문맥 정보가 동일한 후보가 많아 모든 후보로 Viterbi 검색을 수행한다면 실시간 합성이 힘들어지기 때문에 합성음의 음질 열하를 최소로 할 수 있는 효율적인 합성후보에 대한 사전선택 방법이 반드시 필요하다 [2].

대용량 음성 코퍼스를 이용하는 합성기에서 자연스러운 합성음을 얻기 위해서는 각 합성 단위에서의 후보들도 DB내에서 연속되어야 하는데, 사전선택에서 이러한 후보들이 우선적으로 선택되도록 목표비용을 조절하여야 한다. 이를 위해서 VoiceText에서 중요하게 사용하는 것이 음소 열에 대한 비용이고, 이것을 위해 각 후보들의 CCL (connected context length)를 구한다. CCL은 합성하고자 하는 음소의 전, 후 음소 열과, 후보의 DB내 전, 후 음소 열을 비교하여 일치하는 음소 열의 최대 길이를 의미한다. 계산 과정은 다음 장에서 설명한다.

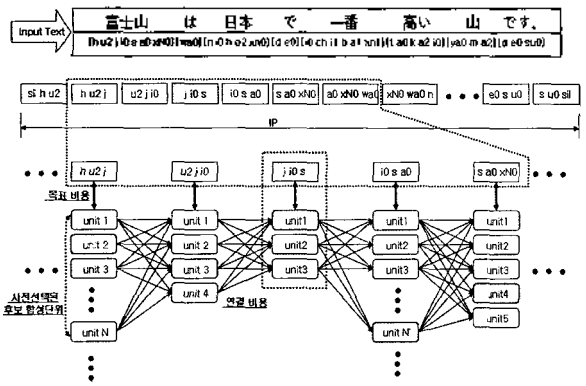

그림 2. VoiceText의 합성단위 선택
Fig. 2. The unit selection of Japanese VoiceText System.

본 논문의 baseline 시스템으로 사용된 VoiceText에서는 Viterbi 검색이 하나의 IP에서 이루어지므로, CCL도 IP 범위에서 음소 열을 비교하여 구한다. VoiceText의 사전선택은 각 후보의 전, 후 음소 열과 합성하고자 하는 음소의 전후 음소 열을 IP 전체에 대해 비교하여 CCL을 구하고, 그 값이 큰 순으로 후보를 정렬한다. 정렬된 후보들의 피치, 음소 지속시간 등 나머지 파라미터의 목표비용을 계산하여 최소 목표비용을 가지는 N (N은 상수)개 이하의 후보를 남긴다. 이때, CCL과 반비례하는 가중치를 적용하여, CCL 값이 큰 후보들의 선택될 가능성을 높인다. 그러나 일본어는 AP 안에서 운율의 변화가 크게 형성되어, AP가 다른 음소들 사이의 상관관계가 크지 않게 된다. 따라서 사전선택에서도 이러한 특징을 이용하는 것이 바람직하므로 본 논문에서는 VoiceText의 사전선택 방법을 IP에 대해 적용하는 것이 아니라 AP에 적용하는 악센트 구 매칭 방법을 제안한다.

사전선택을 통과한 후보들에 대해서는 연결 비용을 계산하여 위에서 계산된 목표 비용과 합하여 Viterbi 검색을 수행한다.

2.4. 파형 생성

합성단위 선택 과정에서 Viterbi 검색을 수행하면 입력된 텍스트에 대한 최적의 합성단위 열을 얻을 수 있는데, 합성음은 선택된 각각의 합성단위에 대한 파형을 연결하여 생성한다.

III. 제안한 방법

본 논문에서 제안하는 악센트 구 매칭 방법은 위와 같은 VoiceText의 사전선택 방법을 AP에 대해 적용하기위해 CCL을 AP 범위에서 계산하는데, 이를 위해서는 먼저 입력 텍스트에서 분석된 하나의 IP에 해당하는 음소 열을 break정보를 이용하여 하나 이상의 AP들로 분리한다.

3.1. Break 정보와 악센트 구의 경계 정보

일본어의 악센트는 인접한 음절의 상대적인 피치의 높낮이로 나타나는 것으로, 강약이 아닌 고저 악센트이고, 하나의 AP에서 악센트가 한번 내려가면 다시 올라오지 않는 특징이 있다 [4]. 그리고 VoiceText의 일본어 합성

표 1. Break 인덱스와 AP 경계 정보
Table 1. Break Index and Accentual Phrase Boundary.

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

표 2. 입력 텍스트의 분석 결과
Table 2. The analysis result of the input text. [富士山は日本で一番高い山です。].

텍스트	발음기호	Break	APB
富士山	[hu2ji0sa0xN0]	1	0
は	[wa0]	3	2
日本	[ni0ho2xn0]	1	0
で	[de0]	3	2
一番	[i0chi1ba1xn1]	2	1
高い	[ta0ka2i0]	2	1
山	[ya0ma2]	1	0
です。	[de0su0]	5	2

기는 일본어 발음을 표현하기 위해 악센트 정보가 포함된 기호를 정의하여 사용한다. 예를 들어, 모음 [ア]를 [a0], [a1], [a2]로 표현한다. 여기서 [a]는 음가이고, 아라비아 숫자는 악센트 정보이다.

VoiceText의 일본어 합성기에서 사용하는 break 인덱스는 단어와 단어 사이의 연결 정보로써 앞 단어의 마지막 음소의 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, AP 사이에 포즈 없이 연결되는 경우를 2로 나타내고, 그렇지 않고 두 AP 사이에 포즈가 존재하는 경우를 3으로 나타낸다. 여기서 포즈는 IP와 IP사이에 나타나는 포즈보다는 짧은 것으로, 합성음 생성에서는 50msec의 포즈를 사용한다. 4와 5는 각각 IP, 문장이 분리되는 것을 나타낸다.

표 1은 break와 AP 경계정보의 관계를 나타낸 것이다. break 0과 1은 AP의 경계가 아니고, 2는 포즈가 없이 두 개의 AP가 연결되는 것을, 3은 두 개의 AP 사이에 포즈가 존재하는 것을 각각 나타내는 것으로, AP 측면에서 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 2는 실제 입력텍스트 [富士山は日本で一番高い山です] (후지산은 일본에서 가장 높은 산이다.)의 break와 AP 경계 정보를 나타낸 것이다. 입력 텍스트는 5개의 AP로 분석되는데, 첫 번째와 두 번째는 포즈가 연결되는 AP이고, 세 번째, 네 번째 그리고 다섯 번째 사이에는 포즈가 연결되지 않는 것을 나타낸다.

3.2. CCL과 악센트 구 매칭 방법

현재 합성하고자 하는 음소를 p[i], p[i]의 인접한 전,후 음소 열을 {、、、p[i-2], p[i-1], p[i+1], p[i+2], 、、、}라 하고, p[i]의 후보를 u[i], u[i]의 DB내 인접한 전, 후 음소 열을 {、、、u[i-2], u[i-1], u[i+1], u[i+2] 、、、}라고 한다면, tri-Phone으로 합성단위 선택을 수행하는 경우 합성하고자 하는 음소 열, p[i-1]-p[i]- p[i+1]와 u[i-1]-u[i]-u[i+1]는 기본적으로 일치하고, CCL은 1으로 한다. CCL 값은 전방(forward), 후방(backward) CCL로 나누어 계산되는데, p[i-2]-p[i-1]-p[i]와 u[i-2]-u[i-1]-u[i]가 일치하면, 전방 CCL 값이 1 증가하고, p[i]-p[i+1]-p[i+2]와 u[i]-u[i+1]-u [i+2]가 일치하면 후방 CCL 값이 1 증가한다. 각 후보에 대해 위와 같은 과정을 악센트 경계까지 계속하여 CCL 값을 구한다.

표 3과 그림 3은 입력텍스트 [富士山は日本で一番高い山です]의 첫 번째 AP ([hu2ji0sa0xN0]-[wa0]) 중, 음소 [i0]에 대한 후보들 (C[0], C[1], C[2])과 그 것들의 녹음 대본, DB 내 음소 열, 그리고 CCL 값을 나타낸 것이다. ([j-i0-s]는 [i0]의 tri-phone이다.) DB내 음소 열은 하나의 AP만 나타내었다. C[0]를 포함하는 AP의 음소 열은 합성하고자 하는 AP의 음소 열과 완전히 일치하고, C[1]과 C[2]는 부분적으로 일치하는 것을 알 수 있다. 합성하고자 하는 AP는 8개의 음소로 되어있고, i[0]는 4번째 음소이다. CCL은 전방, 후방으로 합성하고자 하는 AP의 크기만큼 후보의 DB내 음소 열과 비교하여 얻어지는데, AP의 경계에 위치하는 음소에 대해서는 AP의 경계 정보도 일치해야 한다. AP의 경계 정보가 일치하지 않으면 CCL을 증가시키지 않는다.

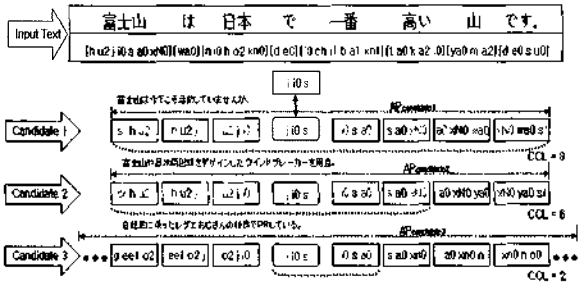

그림 3. 선택된 후보들의 CCL
Fig. 3. The CCL of candidate units.

표 3. (hu2ji0sa0xN0) 중 i0에 대해 선택된 후보
Table 3. Selected candidates of i0 in (hu2ji0sa0xN0).

후보 합성단위	텍스트 데이터	음소열	CCL (전방, 후방)
C(0)	富士山は今でこそ活動していませんが、	hu2ji0sa0xN 0 wa0	8 (4,3)
C(1)	富士山や日米兩國旗をデザインしたウインドブレ"[カ"[を用意。	hu2ji0sa0xN 0 ya0	6 (3,2)
C(2)	自轉車に乘ったレゲエおじさんの映像でPRしている。	re0gee1o2ji0 sa0xn0 no0	2 (0,1)

그림 4는 실험에 사용된 VoiceText의 사전선택 순서도이고, 표시된 부분이 제안한 악센트 구 매칭 방법이다. VoiceText의 사전선택은 IP에서 각 후보의 CCL을 계산하고 정렬한다. 정렬된 후보 중 가장 큰 CCL 값을 갖는 후보의 음소 열이 합성하고자 하는 IP 전체의 음소 열과 일치한다면, 가장 큰 CCL 값을 가진 n개의 후보만 남기게 되고, 그렇지 않은 경우 CCL의 크기와 반비례하는 가중치를 적용하여 음소 열 이외의 목표 비용을 계산한다. 후보 수가 N (N은 상수)개 이상인 경우, 최소 목표비용을 가지는 N개만 선택한다. 제안한 방법은 VoiceText에서 IP 전체가 일치하는 음소 열이 없는 경우, AP에서 CCL을 다시 계산한다. 계산된 CCL로 후보를 정렬한 뒤 AP전체가 일치하는 n개의 후보들이 있다면 그것들만 남기고, 그렇지 않은 경우, AP에서 계산된 CCL을 적용하여 가중치를 계산하고, VoiceText와 같은 과정을 통하여 사전선택을 수행한다. 만일, AP의 전체 음소 열이 일치하는 후보의 수가 N개를 넘는 경우, 목표비용으로 N개의 후보만 남기게 된다. 제안한 방법은 CCL 계산을 IP보다 작은 AP에서 수행함으로써 계산시간을 줄일 수 있을 뿐 아니라, 일본어의 운율구조에 보다 적합한 검색을 수행할 수 있다. 일본어에서는 서로 다른 AP에 존재하는 음소들의 상관관계가 다른 언어에 비해 적고, 독립적인 경계의 AP사이에는 상관관계가 거의 없기 때문에, 사전선택도 이러한 특징을 반영함으로써 합성음의 음질을 향상시킬 수 있다. VoiceText에서 사용하는 CCL은 IP 전체에서 계산되어지기 때문에 AP 경계를 넘어서는 경우가 존재한다. 예를 들어 표 3에서 [i0]의 최대 후방 CCL은 AP 범위에서 3이지만 IP 내에서는 이보다 클 수가 있다. 그러나 CCL이 3보다 크게 되면 AP의 경계를 넘어서기 때문에 실제 음소 사이의 상관관계는 크지 않으나, CCL값이 커짐으로 인해 합성단위 선택 과정에서 선택될 가능성이 커지게 된다. 특히 [i0]가 포함된 AP의 경계는 다음 AP와의 사이에 포즈가 존재하는 독립적인 경계이므로 AP의 경계를 넘어서는 음소와는 상관관계가 거의 없게 된다. 따라서, 일본어 합성기에서는 AP의 경계를 넘어서는 CCL값의 중요도가 실제 크기에 비례하여 커지지 않으므로, CCL의 계산을 AP 내에서 수행해야만 한다. 이것이 보다 일본어의 실제 운율 변화 특성을 잘 반영하는 방법이다.

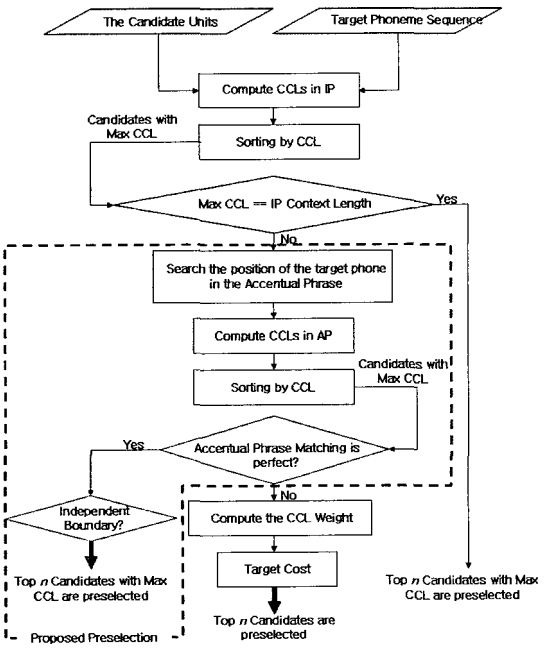

그림 4. 제안한 사전선택 방법
Fig. 4. The proposed pre-selection method.

IV. 실험결과 및 고찰

실험에 사용된 음성 코퍼스는 방음된 녹음실에서 전문 여성 아나운서에 의해 녹음되었고, 녹음을 위해 사용된 대본은 뉴스기사, 소설, 대화문장 및 숫자, 알파벳, 인터넷 주소 (URL), 음절 등으로 구성하였다. 녹음된 음성 코퍼스는 표 4와 같다. 녹음시간은 발성의 중간 포즈만 남기고 처음과 끝의 포즈를 제거한 것이다 [5]. 제안한 방법의 성능을 평가하기 위해, 인지적 에러와 합성시간을 baseline 시스템인 VoiceText와 비교하였다. 인지적 에러는 억양 에러와 연결 에러 (concatenation mismatch error)로 나누어 분석하였다. 인지적 에러 평가는 훈련된 일본인 여성이 JEITA [6, 7] 평가문장 (120 문장)을 이용하여 합성된 합성음을 반복 청취하여, 억양이 이상하거나, 접합부분의 불일치로 인해 잡음이 발생하는 부분을 표시하였다. 억양은 주로 악센트, 음의 크기와 발성 속도 등을 고려하였다. (평가를 담당한 일본인은 2년 이상 합성음 평가 및 악센트 규칙을 개발한 경력을 가지고 있다.) 에러율은 전체 AP 중에 에러가 발생한 AP의 백분율로 나

표 4. 음성 코퍼스
Table 4. Speech corpus.

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

표 5. 인지 에러비교
에러율 = (에러가 발생한 AP의 개수 / 전체 AP의 개수)
Table 5. Comparing perceptual error rate.
Error rate = (number of error AP / total number of AP).

	억양 에러	연결 에러
Baseline	4.20%	2.77%
Proposed Method	3.79%	2.15%

표 6. 합성속도 비교 (sec/500byte)
Table 6. Comparing synthesis time (sec/500byte).

	ARS 대본	뉴스 기사
Baseline	0.402756	0.598359
Proposed Method	0.329666	0.547159

타내었다. 합성 시간은 자동응답시스템 (automatic response system, ARS) 대본과 뉴스 텍스트를 이용하여 측정하였다. ARS 대본은 도메인이 지정된 텍스트에 대한 합성 결과를 나타내고, 뉴스 텍스트는 일반적인 합성 결과를 나타낸다. 합성은 3.06GHz Xeon CPU를 가진 워크스테이션을 이용하여 수행하였다.

표 5는 baseline 시스템과 제안한 방법의 인지적 에러율을 나타낸다. 전체 테스트 문장은 976개의 AP와 2002개의 단어로 분석되고, 에러율은 제안한 방법에서 억양 에러가 0.41%, 연결 에러가 0.62% 줄어들었다. 표 6은 합성시간에 대한 비교이다. ARS 대본에 대해서는 18.15%, 뉴스 텍스트에 대해서는 8.56%의 합성시간이 줄어들었다. 특히 ARS 대본처럼 DB에 포함된 도메인의 텍스트에서 보다 큰 효과가 나타났는데, 이것은 DB가 최적화 될수록 제안한 방법의 합성속도가 개선 될 수 있음을 보여주는 것이다. 이러한 결과에 의해 제안한 방법이 일본어 합성기에서 합성속도를 개선하면서 합성음의 음질을 향상시키는 것을 확인 할 수 있었다.

V. 결 론

현재 널리 사용되는 합성단위 선택 기반 음성 합성 시스템은 고음질의 합성음을 얻기 위해 대용량의 음성 데이터베이스를 사용하고 있다. 그러나 이러한 대용량 데이터베이스의 사용은 합성단위 선택에서의 계산량을 증가시켜 실시간 합성이 어려워지는 문제를 야기하고, 이러한 문제를 해결하기 위해 합성단위 선택의 계산량을 줄여주는 사전선택을 수행한다. 하지만, 사전선택에 의해 음질 저하가 발생 할 수 있으므로, 음질 저하를 피하면서 계산량이 적은 방법이 필요하다.

본 논문에서는 Voiceware의 합성기인 VoiceText에서 사용하는 CCL을 일본어의 악센트 구에 적용하여 합성단위 후보의 사전선택을 수행하는 방법을 제안하였다. 일본어는 상대적인 피치의 높낮이로 나타나는 악센트를 가진 언어로, 운율의 변화가 악센트 구 단위로 나타난다. 따라서 합성단위 선택에서 사용하는 목표비용 계산 시 이러한 악센트 구 단위의 운율변화를 반영하면 일본어에 적합한 비용계산을 할 수 있다. 제안한 방법은 VoiceText를 baseline 시스템으로 하여 구현하고, 인지적 에러와 합성시간에 대해 비교하여 실험하였다. 실험 결과, 인지적 에러가 감소하면서, 합성속도도 좋아지는 결과를 얻을 수 있었다. 앞으로 악센트 구와 같은 일본어에서 나타나는 특징들을 분석하고 이용하여, 일본어 합성기의 합성단위 선택을 지속적으로 개선해야 한다.

참 고 문 헌

1. H. Segi, T. Takagi and T. Ito, "A Concatenative Speech Synthesis Method Using Context Dependent Phoneme Sequences with Variable Length as Search Units", Proc. 5th ISCA Speech Synthesis Workshop, 115-120, Pittsburgh, June, 2004.
2. 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. ICSLP-2000, 3, 314-317, Beijing, Oct, 2000.
3. T. Mizutani and T. Kagosima, "Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method", IEICE Trans. Inf. & Syst., E88-D, (11) 2565-2572, 2005.
4. J. Venditti, "Japanese ToBI Labeling Guidelines.", OSU Working Papers in Linguistics, 127-162, 1997.
5. H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda: "Ximera: A New TTS from ATR Based on Corpus-Based Technologies," Proc. ISCA 5th Speech Synthesis Workshop, 179-184, Pittsburgh, June, 2004.
6. T. Kazuyo, A. Makoto, M. Toshimitsu and I. Shuichi, "JEIDA Standard of Symbols for Japanese Text-to-Speech Synthesizers", Proc. 3rd Oriental COCOSDA Workshop, 27-32, Beijing, Oct, 2000.
7. Technical Standardization Committee on Speech Input/Output Systems, "Speech Synthesis System Performance Evaluation Methods", JEITA IT-4001, 42-45, April, 2003.

저자 약력

• 나 덕 수 (Deok-Su Na)

2007년 2월: 숭실대학교 정보통신공학과 박사수료
현재: (주)보이스웨어 연구원
한국음향학회지 제19권 제2호 참조

• 민 소 연 (So-Yeon Min)

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

• 이 광 형 (Kwang-Hyoung Lee)

1998년 2월: 광주대학교 전자공학과 (공학사)
2002년 2월: 숭실대학교 전자공학과 (공학석사)
2005년 2월: 숭실대학교 전자공학과 (공학박사)
현재: 서일대학교 인터넷정보과 교수

• 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