HMM 기반의 한국어 음성합성에서 지속시간 모델 파라미터 제어

Control of Duration Model Parameters in HMM-based Korean Speech Synthesis

김 일 환* · 배 건 성*
Il Hwan Kim · Keun Sung Bae

ABSTRACT

Nowadays an HMM-based text-to-speech system (HTS) has been very widely studied because it needs less memory and low computation complexity and is suitable for embedded systems in comparison with a corpus-based unit concatenation text-to-speech one. It also has the advantage that voice characteristics and the speaking rate of the synthetic speech can be converted easily by modifying HMM parameters appropriately. We implemented an HMM-based Korean text-to-speech system using a small size Korean speech DB and proposes a method to increase the naturalness of the synthetic speech by controlling duration model parameters in the HMM-based Korean text-to speech system. We performed a paired comparison test to verify that theses techniques are effective. The test result with the preference scores of 73.8% has shown the improvement of the naturalness of the synthetic speech through controlling the duration model parameters.

Keywords: HMM, speech synthesis, HTS, state-duration model

1. 서 론

현재 음성합성 기술 중 음소나 음절 등의 단위로 음성 파형을 추출하여 저장한 후, 이를 이용하여 합성하는 코퍼스(corpus) 기반의 음성합성 방식이 높은 음질로 인해 최근에 가장 많이 사용되고 있다. 코퍼스 기반의 음성합성 방식은 자연스러운 운율을 표현하는 등 높은 음질의 합성음을 만들어 내지만 합성을 위해 대용량의 DB를 필요로 하고, 합성음의 발음속도나 음색 변환이 용이하지 않다. 이에 반해 최근에 많이 연구되고 있는 HMM(Hidden Markov Model) 기반의 음성합성 방식[1]은 스펙트럼과 여기신호 파라미터를 통계적인 음향모델로 표현하여 음성을 합성하므로 합성을 위한 DB 사이즈가 적을 뿐만 아니라, 화자적응과 발음속도 조절 등의 기법 적용이 용이하므로 적은 DB 로도 쉽게 합성음의 음색이나 발음속도를 변화시킬 수 있다는 장점이 있다. 이러한 HMM 기반의 음성합성 방식에 대한 연구가 일본 및 유럽에서는 활발히 진행되어왔으며 HTS(HMM-based Text-to-Speech System)의 소스도 공개되어왔지만[2] 한국어 합성에 대한 연구는 별로 활발하지 못한 형편

* 경북대학교 전자전기컴퓨터학부

이다.

본 논문에서는 HTK(HMM Tool Kit)를 기반으로 웹에 공개된 HTS 시스템을 간략히 소개하고, 한국어 DB를 이용한 음성합성에서 합성음의 자연성을 높이기 위해 음향모델의 상태지속시간 파라미터를 제어하는 방법을 제안하고, 그 실험결과를 제시한다. 한국어 HTS 시스템은 일본의 나고야 대학에서 배포한 HTS 2.0 버전을 기반으로 약 60 분 정도 녹음된 총 646 문장의 소용량 한국어 서울말 낭독체 DB를 이용하여 구현하였다. 합성음의 자연성을 개선시키기 위해 문장의 끊어 읽기 정보를 추가하도록 합성 과정에서 지속시간 모델 파라미터를 후처리 과정으로 제어하였으며, 이를 통해 좀 더 자연성이 개선된 합성음이 되도록 하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 HTS 시스템에서의 훈련부분과 합성부분에 대해서 간략하게 설명하고, 3 장에서는 한국어 음성합성 실험 환경에 대한 설명과 합성음의 자연성 증가를 위한 실험 결과를 제시한다. 마지막으로 4장에서 결론을 맺는다.

2. HMM 기반의 음성합성시스템

<그림 1>은 HMM 기반의 음성합성시스템의 전체적인 구조를 보인 것이다[3]. 시스템은 훈련부분과 합성부분으로 나누어진다. 먼저 훈련부분에서는 음성 DB로부터 스펙트럼 특성과 여기신호 특성을 나타내는 특징파라미터를 각각 추출하고, 텍스트분석기를 통해 미리 분석된 음성 DB의 레이블 정보를 이용하여 각각의 파라미터는 음성인식에서와 같은 방법으로 HMM 훈련과정을 통해 문맥종속 HMM 모델을 생성한다. 합성부분에서는 입력된 어절 또는 문장을 텍스트분석기를 통해 레이블 열로 변환하고, 각 레이블에 대응하는 문맥종속 HMM 모델을 가지고 와서 연결하게 된다. 그리고 연결된 모델에서 여기신호 파라미터와 스펙트럼 파라미터를 생성하고, 이를 합성 필터를 통해 음성 파형을 만들어 낸다.

2.1 훈련 부분

합성시스템에서 HMM의 관측 벡터는 스펙트럼 파라미터와 여기신호 파라미터로 구성된다[4]. 스펙트럼 파라미터로는 멜 켑스트럼(Mel-Cepstrum)과 이의 동적 성분들이 사용되며, 여기신호 파라미터는 로그 기본주파수(log F0)와 이의 동적 성분으로 구성된다. 각각의 HMM 모델은 음성의 시간적 변이정보를 표현해 내기 위해 상태 유지 길이에 대한 파라미터도 가지고 있다. 훈련 시, 적절한 초기 HMM 모델이 설정되어 있으면 그 다음부터는 HMM의 모든 파라미터가 임베디드 훈련 과정을 통해 자동으로 재 추정 될 수 있으므로 모든 음성 DB가 음소 단위로 레이블링 되어 있지 않아도 문맥종속 HMM 음향모델의 생성이 가능하다.

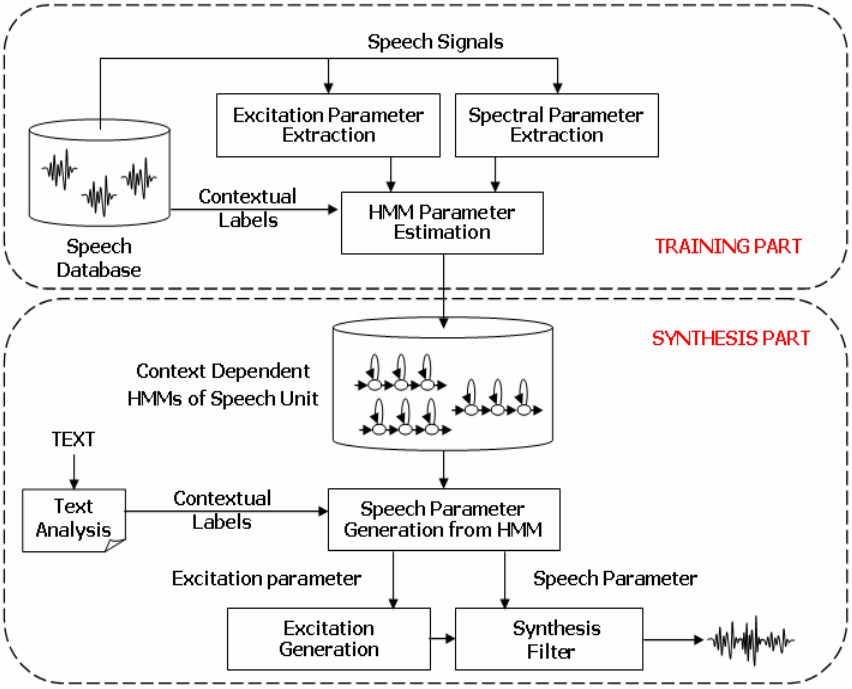


그림 1. HMM 기반의 음성합성시스템 (HTS)

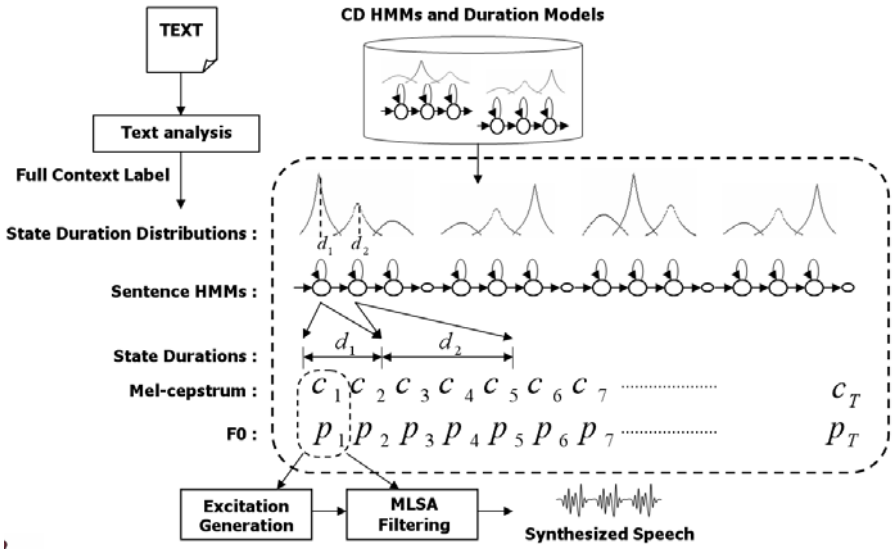


그림 2. HTS 음성합성 과정

2.2 합성 부분

<그림 2>에 HMM을 이용한 음성합성 과정을 나타내었다. 우선 주어진 텍스트를 문맥종속 레이블 열로 변환하고 레이블 열에 따라 대응하는 문맥종속 HMM 모델을 연결함으로써 문장 HMM 모델을 만든다. 그리고 문장 HMM에서 상태지속시간 모델을 기반으로 각 상태의 지속시간을 결정하고, HMM의 관측확률이 최대가 되는 스펙트럼과 여기신호 파라미터들을 결정한다. 마지막으로 생성된 스펙트럼과 여기신호 파라미터들로부터 MLSA(Mel Log Spectrum Approximation) 필터를 통해 합성음을 만들어 낸다[5].

3. 합성음의 자연성 개선을 위한 실험 및 결과

3.1 한국어 음성합성을 위한 실험 환경

본 연구에서 HMM 음향모델을 만들기 위해 훈련에 사용된 음성 DB는, 초기음향모델을 만들기 위한 ETRI 611 DB와 문맥종속 음향모델을 만들기 위한 서울말 낭독체 DB이다. ETRI 611 DB는 음소단위로 레이블링 된 단어로 구성되어 있으며, 서울말 DB는 레이블링 되지 않은 문장으로 구성되어 있다. 음성신호는 16 kHz 샘플링에 16 bit로 양자화 되었으며, 분석 프레임 크기는 25 ms이고, 블랙만 윈도우가 5 ms씩 이동하며 취해진다. 그리고 멜 켑스트럼 분석 기법을 통해 얻어지는 24 차의 멜 켑스트럼과 영차의 에너지, 그리고 이들의 차분값, 차분-차분값 등 총 75 개의 값이 특징벡터의 스펙트럼 파라미터로 사용된다. 하나의 문맥종속 HMM 모델은 5 개의 상태를 가지고, 각 상태는 싱글 가우시안 분포를 가진다. 본 논문에서 사용한 음소는 <표 1>에 주어진 것과 같이 초성음 19 개, 중성음 20 개, 종성음 7 개로 구별해 총 46 개의 유사음소로 구분되며, 묵음 모델을 포함해 총 47 개의 초기 음향모델을 만들었다. 초기 음향모델은 음소 단위로 레이블이 되어있는 ETRI 611 DB 중 rjh DB를 사용하여 만들었다[6]. 초기 모델은 문맥정보에 의해 문맥종속 모델로 확장되는데, 이때 서울말 낭독체 DB 총 646 문장을 이용하여 임베디드 훈련과 클러스터링을 통해 문맥종속 모델을 완성하였다. 이때 사용된 문맥 정보는 아래와 같다.

- {선행, 현재, 후행} 음소
- 현재 어절의 음절 수
- 현재 어절에서 음절의 위치

HTS에서는 문맥 정보가 합성음의 음질이나 운율 표현에 있어서 아주 중요한 요소라고 할 수 있다. 그러나 본 논문에서 사용한 제한된 DB와 텍스트분석기를 고려하여 선・후행 음소, 현재 어절의 음절 수 및 위치 정보까지만 확장하였다. 그리고 한국어의 모음과 자음의 음성학적 특성을 고려해 각각 몇 개의 그룹으로 구분하고, 결정-트리 기반의 문맥 클러스터링을 통해 문맥종속 모델을 완성하였다. 이러한 방법으로 생성한 음향모델을 이용하여 합성 실험을 한 결과, 충분히 인지할 정도의 양호한 음질을 갖는 합성음을 생성할 수 있었다[7].

합성음의 음질을 결정하는 요소는 크게 명료성과 자연성으로 나눌 수 있다. HMM 기반의 음성합성 방식에서 합성음의 명료성을 결정하는 파라미터는 스펙트럼 모델이고, 자연성을 결정하는 주요 파라미터로는 여기신호 모델과 지속시간 모델이다. 본 논문에서는 음향모델의 상태지속시간 파라미터를 제어하여 합성음의 자연성을 개선시키기 위한 실험을 하였다.

표 1. HTS 합성에 사용된 유사음소

b	ㅂ	0	j’	ㅉ	10	eo	ㅓ	20	yu	ㅠ	30	ne	ㄴ(종)	40
b’	ㅃ	1	ch	ㅊ	11	o	ㅗ	21	ye	ㅖ	31	de	ㄷ(종)	41
p	ㅍ	2	s	ㅅ	12	u	ㅜ	22	wa	ㅘ	32	le	ㄹ(종)	42
d	ㄷ	3	s’	ㅆ	13	eu	ㅡ	23	weo	ㅝ	33	me	ㅁ(종)	43
d’	ㄸ	4	m	ㅁ	14	i	ㅣ	24	wi	ㅟ	34	be	ㅂ(종)	44
t	ㅌ	5	n	ㄴ	15	ae	ㅐ	25	wae	ㅙ	35	ngo	ㅇ(종)	45
g	ㄱ	6	h	ㅎ	16	e	ㅔ	26	we	ㅞ	36	sil		46
g’	ㄲ	7	l	ㄹ	17	ya	ㅑ	27	oe	ㅚ	37			
k	ㅋ	8	r	ㄹ	18	yeo	ㅕ	28	eui	ㅢ	38			
j	ㅈ	9	a	ㅏ	19	yo	ㅛ	29	ge	ㄱ(종)	39			

3.2 지속시간 모델 파라미터 제어

상태지속시간 정보에 대한 밀도 함수는 N-차원 가우시안 분포 함수로 모델링된다[8]. 각 HMM 모델의 상태지속시간 밀도 함수의 차원은 HMM 상태 수와 같고, 상태 지속시간 밀도 함수의 n번째 차원은 HMM의 n번째 상태에 대응된다. 이 상태지속시간 정보에 대한 모델링은 합성음의 발음속도에 영향을 미치고, 이를 통해 쉽게 발음 속도를 조절할 수 있다. <그림 2>의 음성합성 과정에서 HMM 파라미터 발생 알고리즘에 의해 합성음의 길이 T가 주어지면 식 (1)의 값을 최대로 하는 상태열 $\mathbf{q}=q_1,q_2,...,q_T$를 얻어 최적의 상태지속시간을 결정한다.

$$\log P(\mathbf{q}|\lambda)=\sum_{k=1}^{K}\log p_k(d_k),\ \text{where}\ \ T=\sum_{k=1}^{K}d_k \tag{1}$$

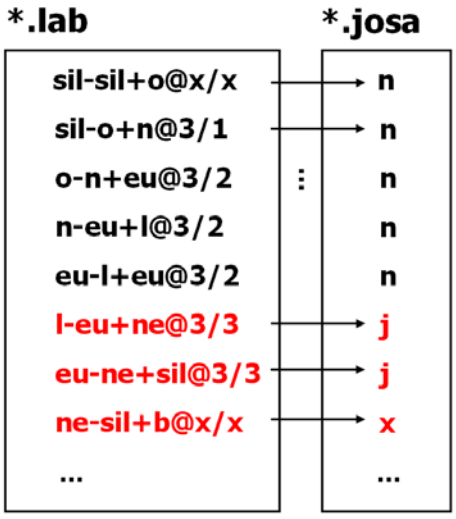
여기서 $p_k(d_k)$값은 상태 k에서 지속시간 d_k의 확률이고, K값은 HMM λ의 총 상태 수를 나타낸다. 각각의 지속시간 밀도 $p_k(d_k)$는 한 개의 가우시안 분포로 이루어져 있기 때문에 식 (1)을 최대로 하는 상태 지속시간 $(d_k)_{k=1}^{K}$는 아래의 식으로 주어진다.

$$d_k=\alpha\cdot\xi(k)+\beta\cdot\sigma^2(k) \tag{2}$$

$$T=\alpha\cdot\sum_{k=1}^{K}\xi(k)+\beta\cdot\sum_{k=1}^{K}\sigma^2(k) \tag{3}$$

여기서 $\xi(k)$와 $\sigma^2(k)$는 각각 상태 k에서 지속시간 밀도의 평균과 분산 값이다. 식 (3)에서 α와 β는 T와 연관되어 있기 때문에 발음 속도는 T 대신에 α와 β를 이용하여 제어할 수 있다. 식 (2)로부터 $\alpha=1, \beta=0$으로 설정될 때, 즉 $T=\sum_{k=1}^{K}\xi(k)$일 때 평균 발음속도로 음성이 합성된다는 것을 알 수 있고, α와 β값을 적절히 조절함으로써 발음속도가 더 빨라지거나 더 느려지게 되는 것을 알 수 있다.

운율 경계의 음성적 특징에 관한 연구결과에 따르면 악센트구와 억양구말에서는 피치 패턴과 경계성조의 변화 뿐 아니라 단위 말 음절의 길이 증가 현상이 두드러짐을 알 수 있다. 이러한 경계 앞 음절의 장음화 현상은 억양구말에서 특히 두드러진 특징으로 나타났다. 구체적으로는 악센트구 내부, 악센트구말 및 억양구말에 위치한 동일 음절들의 길이를 비교하였는데, 그 결과 악센트구말 위치에서는 1.74 배 증가하였고, 억양구말에서는 2.35 배의 증가를 보였다[9]. 이러한 연구결과를 기반으로 하여 텍스트분석기와 현재 웹에 배포되어 있는 HTS 소스를 다음과 같이 수정하였다.



(a) (b)

그림 3. /...오늘은.../ 에 대한 텍스트분석기의 출력 예
(a) 기존의 텍스트분석기 출력파일 (b) 조사 및 조사 뒤 묵음구간 검출파일

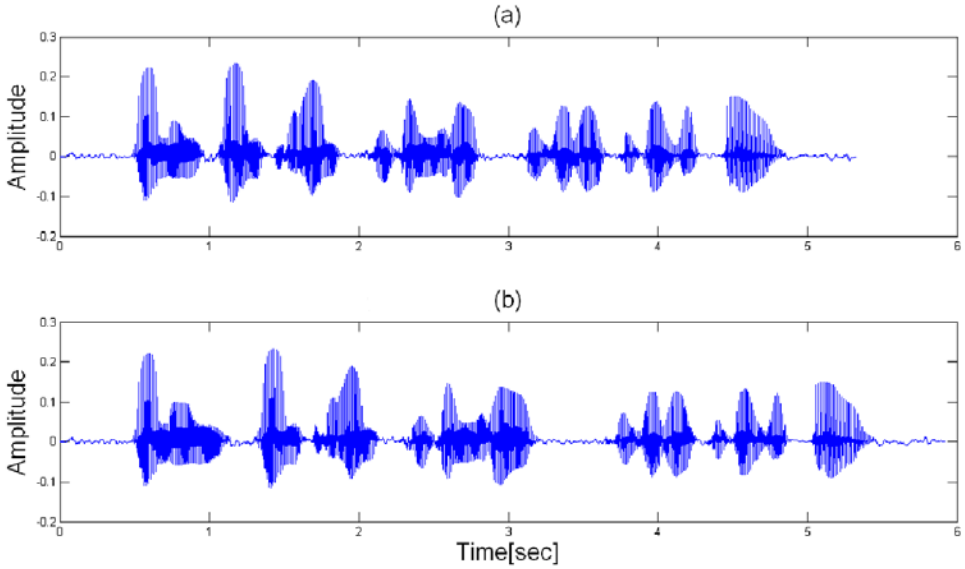


그림 4. 합성음 비교(입력문장 : /나는 난처한 표정으로 고개를 끄덕였다/)
(a) Synthetic Speech (b) Modified Synthetic Speech Signal with Duration Control

<그림 3>은 수정된 텍스트분석기의 출력 예를 보여준다. 우선 입력 문장에서 단위말에 해당하는 음절을 검출하고, 단위말 음절 및 그 음절 뒤의 묵음구간에 해당하는 모델의 지속시간 파라미터를 제어하여 문장에서 끊어 읽기 정보를 추가할 수 있도록 하였다. <그림 3(a)>는 기존의 텍스트분석기의 출력파일인데 단위말 검출은 기존의 텍스트분석기에 조사 검색 코드를 추가하여 <그림 3(b)>와 같은 조사 및 조사 뒤 묵음구간 검출파일을 생성하도록 하였다. <그림 3(b)>에서 j는 조사, x는 조사 뒤 묵음, 그리고 n은 나머지 음소를 의미한다. <그림 3>의 정보를 이용하여 각각의 음소별로 식 (3)의 지속시간 모델 파라미터(α,β)를 제어할 수 있도록 HTS 소스를 수정하였다.

문장에서 끊어 읽기 정보를 추가하여 합성음의 자연성을 개선하기 위해서는 지속시간 모델을 각각의 음소 별로 제어해야 하는데, 지속시간 모델의 분산 값은 변이가 크므로 β는 0으로 고정하고 α값만을 이용하여 지속시간 모델을 제어하는 것이 바람직하다. 실험에서 사용한 파라미터 값은 운율 경계의 음성적 특징에 관한 연구 결과를 기반으로 하여, 단위말 음절은 α=1.7, β=0, 단위말 음절 뒤 묵음구간은 α=2.0, β=0로 설정하여 합성 실험을 하였다. <그림 4>는 기존의 합성음과 지속시간 모델 파라미터 제어를 통해 얻은 합성음을 나타낸 것인데, 각 음소모델 별로 지속시간이 늘어나면서 합성음성의 전체 길이도 늘어났음을 볼 수 있다. 만약 합성음의 자연성을 개선하고자 하는 것이 아니라 단순히 문장 전체의 발음 속도를 하나의 변수로 제어하기 원한다면 분산 값의 가중치인 β값을 적절히 조절하면 된다[8].

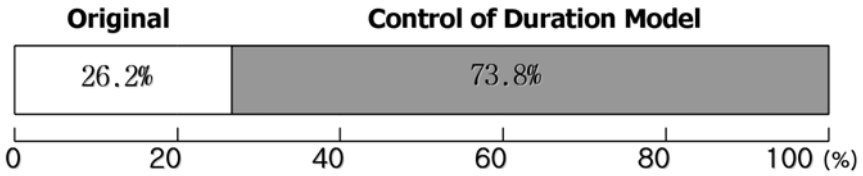


그림 5. 기존의 합성음과 지속시간 모델을 제어한 합성음의 주관적 음질 평가 결과

지속시간 모델 파라미터 제어를 통해 얻은 합성음과 기존의 합성음의 자연성을 비교해 보기 위해서 informal한 주관적 음질 평가를 수행하였다[10]. 우선 훈련 DB에 포함된 문장과 포함되지 않은 문장을 각각 20 문장 씩 총 40 문장을 선택하고, 동일한 문장에 대하여 합성음을 비교하기 위해 지속시간 모델 파라미터를 제어한 합성음과 기존의 방법으로 생성한 합성음을 각각 40 문장 씩 생성하였다. 총 10 명의 청취자들에게 40 문장 중 임의로 8 문장을 선택하도록 하여, 두 가지 합성음을 랜덤한 순서로 듣고 그중 더 자연스러운 합성음을 선택하도록 하였다. <그림 5>에 주관적 음질 평가 결과를 나타내었다. 본 논문에서 제안한 지속시간 모델 파라미터 제어를 통해 얻은 합성음이 73.8%의 득점을 얻음으로써 기존의 합성음 보다 자연성이 좀 더 개선된 것을 확인할 수 있었다.

4. 결 론

코퍼스 기반의 음성합성방식에 비해 HMM 기반의 음성합성 방식은 모델 파라미터 조절을 통하여 쉽게 음색이나 발음속도를 조절할 수 있다. 본 논문에서는 이러한 HMM 기반의 음성합성 방식의 장점을 이용하여, 지속시간 모델을 제어하여 소용량의 제한된 DB를 사용한 합성음의 자연성을 개선할 수 있는 방법을 제안하고, 합성실험을 수행하였다. 단위말 음절의 장음화 현상을 기반으로 하여 조사 및 조사 뒤의 묵음 길이를 1.7~2 배로 늘려서 문장 전체의 자연성이 개선되도록 하였는데, 실험 결과 기존의 방식에 비해 자연성이 개선된 합성음을 얻을 수 있었다. 본 논문에서는 단위말 음절을 검출하기 위해서 조사를 검색하여 실험하였는데, 향후 정확한 한국어 구문분석기를 사용하여 단위말을 검출하거나, 문맥정보에 악센트구나 억양구의 정보가 추가되어 있는 음성 DB를 사용하면 보다 자연스러운 합성음을 얻을 수 있을 것으로 사료된다.

참 고 문 헌

[1] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi T. & Kitamura, T. 1999. “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis,” *Proc. of EUROSPEECH* vol. 5, 2347-2350.

[2] http://hts.sp.nitech.ac.jp/

[3] Kim, S. J., Kim, J. J. & Hahn, M. S. 2006. “Implementation and evaluation of an HMM-based Korean speech synthesis system,” *IEICE Trans. Inf. & Syst* vol. E89-D(3),

1116-1119.
[4] Tokuda, K., Masuko, T., Miyazaki, N. & Kobayashi, T. 1999. “Hidden Markov Models Based on Multi-Space Probability Distribution of Pitch Pattern Modeling,” *Proc. of ICASSP*.
[5] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. 2000. “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. of ICASSP 2000* vol. 3, 1315-1318.
[6] 김일환, 배건성. 2008. “HMM 기반의 한국어 음성합성에서 음색변환에 관한 연구,” *한국음성과학회 · 대한음성학회 공동학술대회 논문집,* 36-39.
[7] 배재철, 배건성. 2007. “HMM 기반의 한국어 음성합성,” *신호처리합동학술대회 논문집* vol. 20, 144.
[8] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. 1998 “Duration modeling for HMM-based speech synthesis,” *Proc. of ICSLP* vol.2, 29-32.
[9] 한선희, “운율 경계의 음성적 특징 연구,” *한국음향학회지* 17(5), 12-21.
[10] Yamagishi, J., Kobayashi, T., Renals, S., King, S., Zen, H., Toda, T. & Tokuda, K. 2007. “Improved Average-Voice-based Speech Synthesis using Gender-Mixed Modeling and A Parameter Generation Algorithm considering GV,” *Proc. ISCA SSW6*.

접수일자: 2008. 10. 28
수정일자: 2008. 11. 25
게재결정: 2008. 12. 6

▲ 김일환
대구광역시 북구 산격동 1370번지
경북대학교 전자전기컴퓨터학부
Tel: +82-53-940-8627 Fax: +82-53-940-8827
E-mail: cutekih@mir.knu.ac.kr

▲ 배건성(교신저자)
대구광역시 북구 산격동 1370번지
경북대학교 전자전기컴퓨터학부
Tel: +82-53-950-5527 Fax: +82-53-940-8827
E-mail: ksbae@ee.knu.ac.kr