HMM 기반의 한국어 음성합성에서 음색변환에 관한 연구

김일환(경북대), 배건성(경북대)

<차 례>

1. 서론
2. 화자적응 기법이 추가된 HMM 기반의 음성합성시스템
 2.1. 훈련 부분
 2.2. 적응 부분
 2.3. 합성 부분
3. 음색변환 실험 및 결과
 3.1. HTS 데모 프로그램
 3.2. 한국어 합성음 음색변환 실험 및 결과
4. 결론

<Abstract>

A Study on the Voice Conversion with HMM-based Korean Speech Synthesis

Il Hwan Kim, Keun Sung Bae

A statistical parametric speech synthesis system based on the hidden Markov models (HMMs) has grown in popularity over the last few years, because it needs less memory and low computation complexity and is suitable for the embedded system in comparison with a corpus-based unit concatenation text-to-speech (TTS) system. It also has the advantage that voice characteristics of the synthetic speech can be modified easily by transforming HMM parameters appropriately. In this paper, we present experimental results of voice characteristics conversion using the HMM-based Korean speech synthesis system. The results have shown that conversion of voice characteristics could be achieved using a few sentences uttered by a target speaker. Synthetic speech generated from adapted models with only ten sentences was very close to that from the speaker dependent models trained using 646 sentences.

* Keywords: HMM, Speech synthesis, HTS, Voice conversion, Speaker adaptation.

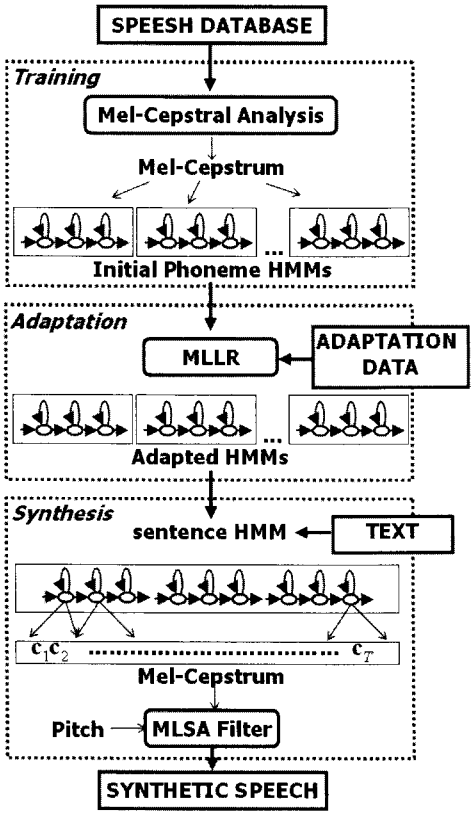
1. 서 론

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[1]. 코퍼스 기반의 음성합성 방식은 자연스러운 운율을 표현하는 등 높은 음질의 합성음을 만들어 내지만 합성을 위해 대용량의 DB를 필요로 하고, 음색을 변화시키기 위해서는 상당한 양의 새로운 DB를 필요로 하므로 음색변환이 용이하지 않다. 이에 반해 최근에 많이 연구되고 있는 hidden Markov model(HMM) 기반의 음성합성 방식[2]은 스펙트럼과 여기신호 파라미터를 통계적인 음향모델로 표현하여 음성을 합성하므로 합성을 위한 DB 사이즈가 적을 뿐만 아니라, 화자적응과 발음속도 조절[3] 등의 기법 적용이 용이하므로 적은 DB로도 쉽게 합성음의 음색이나 발음속도를 변화시킬 수 있다는 장점이 있다. 이러한 HMM 기반의 음성합성 방식에 대한 연구가 일본 및 유럽에서는 활발히 진행되어왔으며 HMM-based text-to-speech system(HTS)의 소스도 공개되어 왔지만 한국어 합성에 대한 연구는 별로 활발하지 못한 형편이다[3][10]-[12]. 본 논문에서는 HMM tool kit(HTK)를 기반으로 웹에 공개된 HTS 시스템을 간략히 소개하고, 화자적응 기법을 이용한 한국어 합성음의 음색변환에 대한 실험 결과를 제시한다. 본 논문에서 사용한 HTS 시스템은 공개된 소프트웨어로 HMM 기반의 음성인식 시스템인 HTK 프로그램을 합성시스템에 적용하기 위해 HTK 소스를 수정한 프로그램이다[4].

본 논문의 구성은 다음과 같다. 2장에서는 화자적응 기법이 추가된 HTS 시스템에서의 훈련부분과 적응부분, 합성부분에 대해서 간략하게 설명하고, 3장에서는 한국어 합성음의 음색변환 실험을 위한 환경과 실험 결과를 제시하고, 4장에서 결론을 맺는다.

2. 화자적응 기법이 추가된 HMM 기반의 음성합성시스템

<그림 1>은 기본적인 HMM 기반의 음성합성시스템에서 화자적응 기법이 추가된 구조를 보인 것이다[5]. 시스템은 훈련 부분, 적응 부분, 그리고 합성 부분으로 나누어진다. 화자적응 기법이 추가된 HMM 기반의 음성합성시스템의 전체적인 과정은 다음과 같다. 일반적인 음성합성 시스템에서는 특정 화자 한 사람의 음성 DB만을 이용하여 화자종속 모델(speaker dependent model)을 생성하지만, HMM 기반의 음색변환 기능을 갖는 음성합성시스템에서는 여러 화자의 음성 DB를 사용하여 화자독립 모델(speaker independent model)을 생성한다. 이렇게 생성된 화자독립 HMM 음향모델은 목표음성(target speech) DB를 이용하여 목표음성에 가깝게 모델



<그림 1> HMM 기반의 음성합성시스템 (HTS)

을 변화시키는 적응과정을 거치게 되고, 적응된 모델을 이용하여 음성합성 과정을 거쳐 음색이 변화된 음성을 얻게 된다.

2.1. 훈련 부분

합성시스템에서 HMM의 관측 벡터는 스펙트럼 파라미터와 여기신호 파라미터로 구성된다[6]. 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 변이 정보를 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 레이블링되어 있지 않아도 문맥종속 HMM 음향모델의 생성이 가능하다.

화자독립 모델을 생성하기 위해서는 먼저 음성인식에서의 음향모델 생성 과정

처럼 여러 화자의 음성 DB를 사용하여 모델을 만든다. 이렇게 생성된 화자독립 모델을 사용하여 음성을 합성하게 되면 훈련에 사용된 개개인 화자의 특성이 나타나지 않는 평균음성(average voice)을 얻을 수 있다.

2.2. 적응 부분

적응 부분은 훈련 부분에서 생성한 화자독립 모델을 이용하여 목표음성에 가깝게 모델을 적응시키는 과정이다. 목표음성 DB는 보통 몇 개의 문장에서 수십 개의 문장을 사용할 수 있으며, 본 논문의 실험에 사용한 HTS-2.0 버전에서는 변환기반 적응기법 중의 하나인 maximum likelihood linear regression (MLLR) 방식을 사용하고 있다[7]. 일반적으로 MLLR은 임의의 특징벡터 또는 행렬을 선형변환 행렬(linear transformation matrix)을 사용하여 변환시키는 기법인데, 본 논문에서는 HTS-2.0에서 제시하는 기본 값으로 연속분포 HMM의 평균 벡터에 대한 적응 변환을 수행하였다. HTS 시스템에서 생성되는 모델은 스펙트럼, 여기신호, 지속시간 모델의 세 가지로 구성되는데, 이 중 단일 가우시안 분포 함수로 모델링 되는 지속시간 모델을 제외하고, HMM으로 모델링 되는 스펙트럼과 여기신호 모델에 대해 MLLR을 적용하였다[8]. MLLR을 이용한 평균벡터의 변환은 식 (1)로 표현된다.

$$\hat{\mu} = \mathrm{A}\mu + \mathrm{b} = \mathrm{W}\xi \tag{1}$$

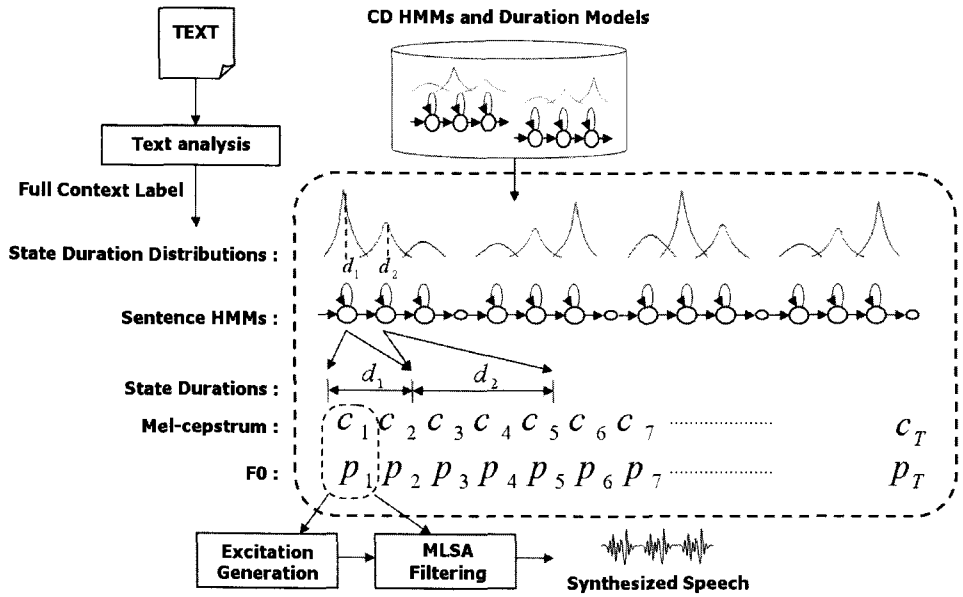
여기서 A는 $d \times d$행렬이고 b는 d차원 벡터이며, d는 관측 벡터의 차수이다. 또한 $\mathrm{W} = [\mathrm{A}, \mathrm{b}]$는 $d \times (d+1)$ regression 행렬이고 ξ는 확장된 평균벡터 $\xi = [\mu_1, \cdots \mu_d, 1]^T$이다. MLLR에서는 regression 행렬 W는 모델에 대해 적응 데이터가 최대우도가 되도록 식 (2)와 같이 추정된다.

$$\hat{\mathrm{W}}_{MLLR} = \arg\max_{\mathrm{W}} f(X \mid W, \lambda) \tag{2}$$

여기서 λ는 HMM 모델을 의미하고, 위의 식은 expectation maximization (EM) 알고리즘을 통해 W의 대한 해를 구할 수 있다.

2.3. 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 상태 지속시간 모델을 기반으로 각 상태의 지속시간을 결정하고, HMM의 관측확률이 최



<그림 2> HTS의 음성합성 과정

대가 되는 스펙트럼(멜-켑스트럼)과 여기신호(F0) 파라미터들을 결정한다. 마지막으로 발생된 스펙트럼과 여기신호 파라미터들로부터 mel log spectrum approximation (MLSA) 필터를 통해 합성음을 만들어 낸다[9].

3. 음색변환 실험 및 결과

3.1. HTS 데모 프로그램[7]

HMM 기반의 음성합성시스템은 HTS working 그룹에서 개발해 왔으며, 여러 버전의 프로그램을 공개하고 있다. HTS의 소스 코드는 HTK의 패치 파일로 배포되고 있으며, 각 버전마다 데모를 위한 스크립트와 DB도 함께 제공하고 있다. HTS-2.0에서는 HTS의 API를 이용해 영어와 일본어 합성을 위한 데모 스크립트와 화자적응 실험을 위한 데모 스크립트도 함께 배포하고 있다. 그리고 1.1 버전 이후부터는 hts_engine이라는 실시간 합성 엔진도 함께 제공하고 있으며, HTS의 인터페이스와 API의 사용 방법은 HTK와 유사하여 HTK 사용자는 쉽게 HTS를 사용할 수 있다.

3.2. 한국어 합성음 음색변환 실험 및 결과

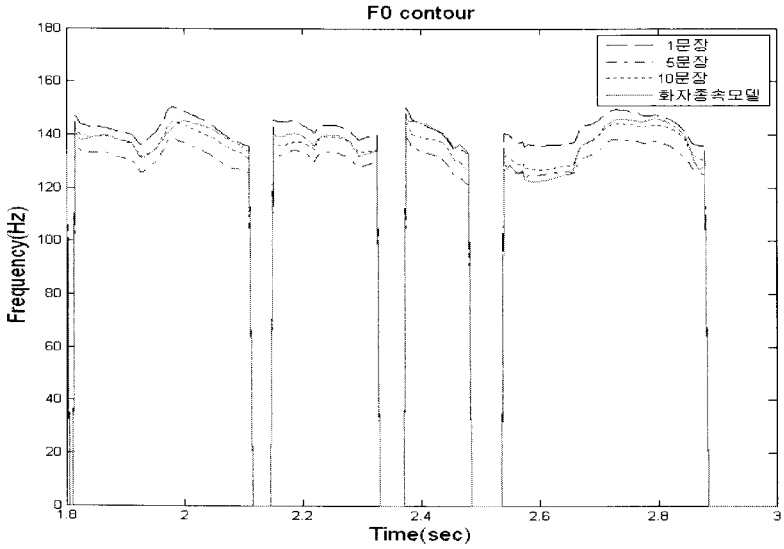
본 연구에서 HMM 음향모델을 만들기 위한 훈련에 사용된 음성 DB는 음소 단위로 레이블링 된 ETRI 611DB와 레이블링이 되지 않은 서울말 낭독체 발화 말뭉치 DB이다. 음성 신호는 16 kHz 샘플링에 16bit로 양자화되어 있다. 분석프레임 크기는 25ms이고, 블랙만 윈도우가 5ms씩 이동하며 취해진다. 그리고 멜-켑스트럼 분석 기법을 통해 얻어지는 24차의 멜-켑스트럼과 영차의 에너지, 그리고 이들의 차분값, 차분-차분값 등 총 75개의 값이 특징벡터의 스펙트럼 파라미터로 사용된다. 하나의 문맥종속 HMM 모델은 5개의 상태를 가지고, 각 상태는 싱글 가우시안 분포를 가진다. 본 논문에서 사용한 음소는 초성음(19), 중성음(20), 종성음(7)을 구별해 총 46개의 유사음소로 구분되고, 묵음 모델을 포함해 총 47개의 초기음향모델을 만들었다. 초기음향모델은 음소 단위로 레이블이 되어있는 ETRI 611DB중 rjh DB를 사용하여 만들었다. 초기음향모델은 문맥 정보에 의해 문맥종속 모델로 확장되는데, 이때 서울말 낭독체 발화 말뭉치 DB를 이용한 임베디드 훈련과 클러스터링을 통해 문맥 종속 모델을 완성하였다. 본 논문에서 사용한 문맥 정보는 다음과 같다[10].

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

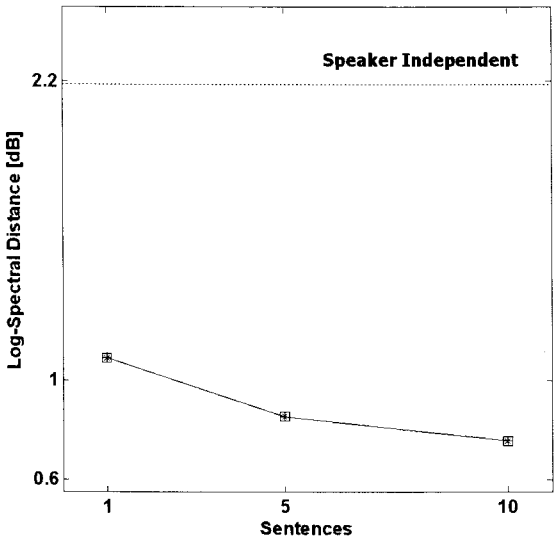
HTS 시스템에서는 문맥 정보가 합성음의 음질을 결정하는 중요한 요소라고 할 수 있다. 그러나 본 연구실에서 보유하고 있는 제한된 DB를 이용한 실험 결과, 위의 정보에 다른 문맥 정보를 추가하더라도 합성음의 음질이 크게 개선되지 않는 것을 확인하고 위의 정보까지만 확장하였다.

화자적응 실험을 위한 화자독립 모델을 만들기 위해서 서울말 낭독체 발화 말뭉치 DB 중 남성화자 3명(mv01, nw05, mw15)과 여성화자 3명(fv16, fv18, fx07), 총 6명의 DB에서 얻은 3563 문장을 사용하여 화자독립 모델을 만들었다. 목표음성 DB는 남성화자(mw06) DB를 사용하였으며 화자종속 모델 훈련에 사용된 646 문장과는 별개의 1개, 5개, 10개 문장으로 각각 화자적응 실험을 하였다.

<그림 3>은 훈련 문장에 포함되지 않은 새로운 문장으로 /여기는 경북대학교 신호처리 연구실입니다/라는 문장의 /신호처리 연구실/ 부분에 대해 각각 1개, 5개, 10개의 목표음성 문장으로 각각 적응시킨 모델과 목표음성 화자의 화자종속모델의 피치궤적의 예를 보인 것이다. 화자종속 모델은 목표음성 DB(mw06)의 646문장으로 모델을 생성하였다. <그림 3>에서 볼 수 있듯이 문장 수가 증가할수록 적응된 모델에 의해 합성된 음성은 화자종속 모델을 사용하여 합성한 음성에 가까워



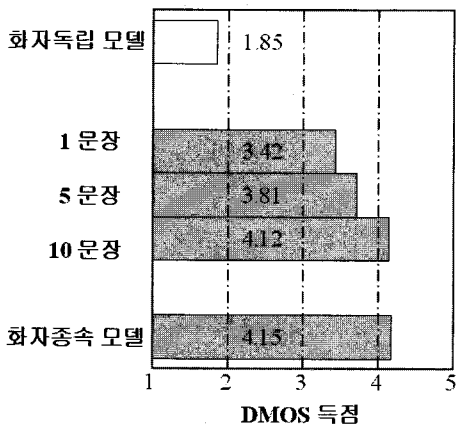
<그림 3> 화자종속 모델 합성음과 적응변환 음성의 피치궤적 비교



<그림 4> 화자종속 모델 합성음과 적응변환 음성 사이의 log-spectral distance

지는 것을 볼 수 있다.

<그림 4>는 화자종속 모델과 적응변환 음성 사이의 log-spectral distance를 나타낸 것이다. <그림 4>에서 점선은 화자독립 모델과 화자종속 모델의 log-spectral distance를 나타내고 x축은 적응에 사용된 목표음성 문장 수를 나타낸다. 그림에서



<그림 5> 음색변환 합성음에 대한 DMOS 평가

<표 1> 구현한 HTS의 파일 크기

Module		Size (kbyte)
Trained HMM data (Speaker Independent Model)	Spectrum	1,562
	Excitation	127
	Duration	120
Decision tree data	Spectrum	146
	Excitation	123
	Duration	57
Synthesis engine		115
Total		2,250

볼 수 있듯이 적응된 모델로 합성한 음성이 화자독립 모델로 합성한 음성에 비해 distance가 크게 줄어들었음을 알 수 있고, 또한 적응에 사용된 문장 수가 증가할수록 distance가 점점 줄어들어 화자종속 모델을 사용하여 합성한 음성에 가까워지는 것을 확인할 수 있다.

본 논문에서는 합성음의 음색 유사도를 평가하기 위하여 DB에 포함된 실제 음성과의 informal한 주관적 음질 평가를 수행하였다. 총 8명의 청취자들에게 평가를 수행하였고, <그림 5>에 각각 화자독립 모델, 화자적응 모델, 화자종속 모델에 의해 합성된 음성의 differential mean opinion score (DMOS) 득점을 나타내었다. <그림 5>에서 확인할 수 있듯이 적응 문장 수가 10 문장일 때에는 화자종속 모델과의 득점차이가 거의 없어 화자종속 모델로 합성한 음성과 큰 차이가 없음을 확인할 수 있다.

<표 1>은 본 논문에서 구현한 한국어 음성합성시스템의 파일 크기를 보인 것

이다. 사용된 메모리 용량은 프로그램과 화자독립 음향모델 데이터 모두를 합해 약 2,250Kbytes 정도로 비교적 적은 메모리를 차지하고 있어 휴대단말 시스템에 임베디드 형태로 적용이 가능함을 알 수 있다.

5. 결 론

본 논문에서는 HMM 기반의 음성합성 방식인 HTS 시스템과 화자적응 기법 중 MLLR을 이용한 한국어 합성음 음색변환 실험 결과를 제시하였다. 비교적 적은 양의 음성 DB를 사용하여 목표음성에 가깝게 음색이 변하는 결과를 얻을 수 있었으며, 충분히 인지 가능한 양호한 음질을 갖는 합성음을 생성할 수 있었다. 음성합성을 위한 음향모델 및 여기신호 모델의 데이터와 합성엔진을 포함하여 전체 메모리 크기가 약 2,250Kbytes 정도를 차지하였다.

목표음성 DB는 음색의 변화에만 쓰여 질 뿐 합성음의 음질(자연성, 명료도)을 결정하는 것은 아니며, 이를 결정하는 것은 화자독립 모델 생성에 사용된 음성 DB이다. 본 연구에서 사용한 음성 DB는 합성용으로 만들어진 DB가 아니므로, 향후 합성음의 음질을 향상시키기 위해서는 합성용 음성 DB를 사용하여 적절한 한국어 문맥 정보를 HMM 음향모델에 추가로 반영시키고, 휴지 및 운율정보를 여기신호에 적절히 반영하는 방법에 대한 연구가 진행되어야 한다.

참 고 문 헌

[1] A. J. Hunt, A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *Proc. ICASSP*, Vol. 1, pp. 373-376, 1996.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", *Proc. Eurospeech*, pp. 2347-2350, 1999.

[3] 김일환, 배건성, "HMM 기반의 한국어 음성합성에서 지속시간 모델 파라미터 제어", *신호처리합동학술대회 논문집*, Vol. 21, p. 125, 2008.

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

[5] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR", *Proc ICASSP*, Vol. 1, pp. 805-808, 2001.

[6] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, "Hidden Markov models based on multi-space probability distribution of pitch pattern modeling", *Proc. ICASSP*, pp. 229-232, 1999.

[7] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, K. Tokuda, "The HMM-based speech synthesis system version 2.0", *Proc. Sixth ISCA Workshop on*

Speech Synthesis, pp. 294-299, 2007.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Duration modeling for HMM-based speech synthesis", *Proc. ICSLP* Vol. 2, pp. 29-32, 1998.

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", *Proc. ICASSP*, Vol. 3, pp. 1315-1318, 2000.

[10] 배재철, 배건성, "HMM 기반의 한국어 음성합성", *신호처리합동학술대회 논문집*, Vol. 20, p. 144, 2007.

[11] S. J. Kim, J. J. Kim, M. S. Hahn, "HMM-based Korean speech synthesis system for hand-held devices", *IEEE Transactions on Consumer Electronics*, Vol. 52, No. 4, pp. 1384-1390, 2006.

[12] S. J. Kim, M. S. Hahn, "Two-band excitation for HMM-based speech synthesis", *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 1, pp. 378-381, 2007.

접수일자: 2008년 11월 10일
게재결정: 2008년 12월 24일

▶ 김일환(Il Hwan Kim)
주소: 702-701 대구광역시 북구 산격동 1370번지
소속: 경북대학교 전자전기컴퓨터학부
전화: 053) 940-8627
E-mail: cutekih@mir.knu.ac.kr

▶ 배건성(Keun Sung Bae) : 교신저자
주소: 702-701 대구광역시 북구 산격동 1370번지
소속: 경북대학교 전자전기컴퓨터학부
전화: 053) 950-5527
E-mail: ksbae@ee.knu.ac.kr