말소리와 음성과학 제4권 제3호 (2012)
pp. 111~117

HMM 기반 한국어 음성합성에서의 화자적응 방식 성능비교 및 지속시간 모델 개선

Performance Comparison and Duration Model Improvement of Speaker Adaptation Methods in HMM-based Korean Speech Synthesis

이 혜 민[1] · 김 형 순[2]
Lee, Heamin · Kim, Hyung Soon

ABSTRACT

In this paper, we compare the performance of several speaker adaptation methods for a HMM-based Korean speech synthesis system with small amounts of adaptation data. According to objective and subjective evaluations, a hybrid method of constrained structural maximum a posteriori linear regression (CSMAPLR) and maximum a posteriori (MAP) adaptation shows better performance than other methods, when only five minutes of adaptation data are available for the target speaker. During the objective evaluation, we find that the duration models are insufficiently adapted to the target speaker as the spectral envelope and pitch models. To alleviate the problem, we propose the duration rectification method and the duration interpolation method. Both the objective and subjective evaluations reveal that the incorporation of the proposed two methods into the conventional speaker adaptation method is effective in improving the performance of the duration model adaptation.

Keywords: Speech synthesis, HTS, speaker adaptation, duration model

1. 서 론

음성합성은 문자를 음성으로 변환하는 기술로, 현재 대용량의 음성 데이터베이스(DB)를 사용하여 높은 음질의 합성음을 생성할 수 있는 코퍼스 기반의 음성합성방식이 주로 사용되고 있다. 이러한 음성합성기술은 우리 생활의 다양한 분야에 사용되고 있으며, 그에 따라 사용자가 원하는 음색의 음성합성기에 대한 요구도 증가되고 있다. 그러나 다양한 음색을 보유한 코퍼스 기반의 음성합성기를 구현하기 위해서는 각 음색에 대한 대용량 DB 작업이 필요하기 때문에, 실제 사용자가 원하는 음색의 합성기를 제공하기는 쉽지 않다.

최근 연구되고 있는 Hidden Markov Model(HMM) 기반의 음성합성(HTS)[1] 기술은 화자적응 기법을 이용하여 이러한 문제점을 해결한다. 코퍼스 기반의 음성합성은 대량의 음성파형을 적절히 가공하여 사용하는 방식이지만, HTS는 음성의 파라미터를 추출하여 이를 통계학적 모델로 변환하는 방식이기 때문에 적은 DB로도 합성기를 만들 수 있으며, 음성 파라미터를 변경함으로써 다양한 형태로 음성 변환이 가능하다는 장점을 가진다. 특히 화자적응을 이용하면 사용자가 원하는 특정 음성에 대한 적은 음성 DB만으로도 다수의 화자로부터 구한 음성 모델을 특정음성 모델로 변환할 수 있다. 이와 같은 장점 때문에 비록 HTS가 코퍼스 기반의 음성합성에 비해 상대적으로 음질이 저하됨에도 불구하고 이에 대한 요구가 늘어나고 있으며 활발한 연구가 진행되고 있다[2].

HTS에서는 음색변환을 위해 음성인식에서 사용되는 화자적응 방식을 확장해서 이용한다. 화자적응을 위한 방법으로는

1) 부산대학교, oasistony@pusan.ac.kr
2) 부산대학교, kimhs@pusan.ac.kr, 교신저자
이 논문은 지식경제부 및 한국산업기술평가관리원의 QoLT 기술개발사업의 일환으로 수행하였으며(과제번호: 10036438), 논문의 일부는 2012년 한국음성학회 봄 학술대회에서 발표된 바 있다[10].

접수일자: 2012년 7월 25일
수정일자: 2012년 9월 21일
게재결정: 2012년 9월 21일

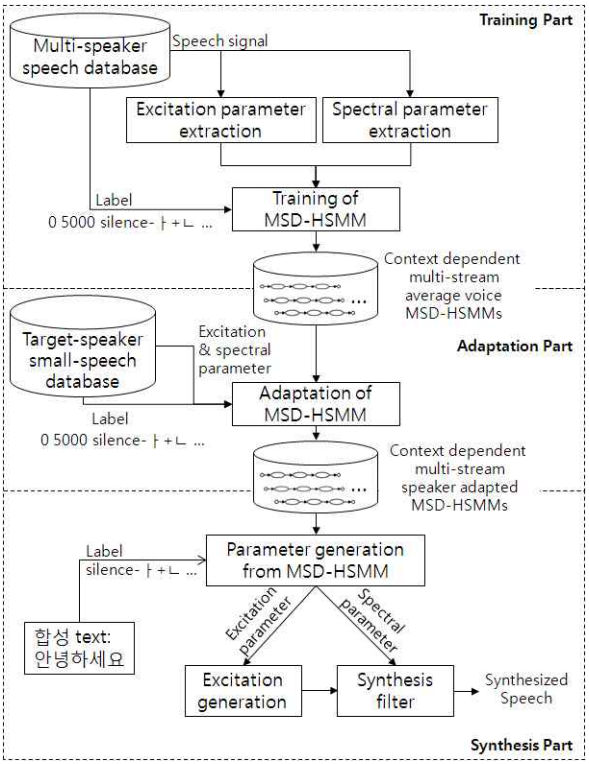


그림 1. 화자적응을 이용한 HTS [4]
Fig. 1. HTS using speaker adaptation [4]

Maximum Likelihood Linear Regression(MLLR) [3], Constrained MLLR(CMLLR)[4][5], Constrained Structural Maximum A Posteriori Linear Regression(CSMAPLR)[4], 그리고 MAP 결합 방식[4] 등이 있으며, 소량의 적응 데이터만으로도 높은 화자적응 성능을 가지도록 하는 것이 HTS 화자적응의 중요한 목표이다.

본 논문에서는 5분 정도의 적은 적응 데이터에 대해 한국어 HTS의 화자적응 성능을 향상시키기 위해서 먼저 기존 화자적응 방식들을 객관적 및 주관적 성능평가를 통해 비교하였다. 그리고 화자적응 과정에서 지속시간 모델 적응의 문제점을 완화시키기 위해 지속시간 교정(rectification)및 보간(interpolation) 방식을 제안하고, 객관적 및 주관적 성능평가 결과 음소 지속시간의 화자적응 성능이 향상됨을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 HMM 기반의 음성합성에서의 화자적응 방식에 대해서 간략한 설명과 객관적 및 주관적 성능비교를 하고, 3장에서는 화자적응 시 지속시간 모델을 개선하는 방식을 제안한다. 4장에서는 제안한 방식에 대한 실험 및 결과를 보이고, 5장에서 결론을 맺는다.

2. HMM 기반의 한국어 음성합성에서의 화자적응

<그림 1>은 화자적응을 이용한 HTS 시스템 구성도를 나타낸 것이다. 이 시스템은 평균음성모델 생성부분(training part), 화자적응 부분(adaptation part), 합성부분(synthesis part)의 세 단계로 구성된다. 평균음성모델 생성방법으로는 화자독립(Speaker Independent(SI)) 방식보다 화자적응훈련(Speaker Adaptive Training(SAT)) 방식[4][8]이 화자적응 시 높은 성능을 나타낸다고 알려져 있어서[4], 본 논문에서도 SAT 모델로 실험을 수행하였다. 또한 HTS에서는 상태 지속시간 정보가 중요하기 때문에 상태출력 확률과 상태 지속시간 확률을 함께 훈련하는 Hidden Semi-Markov Model (HSMM)[6]을 이용하여 모델을 생성하고 화자적응을 수행하였다. HSMM에서 상태 i에서의 출력 확률분포 $b_i(\boldsymbol{o})$와 지속시간 확률분포 $p_i(d)$는

$$b_i(\boldsymbol{o}) = N(\boldsymbol{o};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \tag{1}$$

$$p_i(d) = N(d;m_i,\sigma_i^2) \tag{2}$$

와 같이 표현되며, 각각 평균벡터 $\boldsymbol{\mu}_i$와 대각 공분산행렬 $\boldsymbol{\Sigma}_i$, 스칼라 평균 m_i와 분산 σ_i^2로 정의되는 Gaussian 분포라 가정한다. 여기서 $\boldsymbol{o}$는 관측벡터이고 d는 지속시간이다.

2.1 HSMM 기반의 화자적응 방식

2.1.1 Maximum Likelihood Linear Regression (MLLR)

MLLR은 다수 화자들의 평균모델과 대상(target) 화자간의 차이를 선형변환 행렬로 표현하고 최대 우도(likelihood)를 이용하여 추정하는 방식으로, 상태출력과 지속시간 분포의 평균벡터는 다음 식과 같이 변환된다.

$$\begin{aligned}\overline{\boldsymbol{\mu}}_i &= \boldsymbol{\zeta}\boldsymbol{\mu}_i + \boldsymbol{\epsilon} \\ \overline{m}_i &= \chi m_i + \nu\end{aligned} \tag{3}$$

여기서 행렬 $\boldsymbol{\zeta}$와 상수 χ는 상태 i에서의 출력 및 지속시간 확률 분포의 평균변환의 스케일(scale) 값 이며 벡터 $\boldsymbol{\epsilon}$와 상수 ν는 바이어스(bias)다.

2.1.2 Constrained Maximum Likelihood Linear Regression (CMLLR)

Constrained MLLR(CMLLR)은 상태출력과 지속시간 분포의 평균벡터와 공분산행렬을 동일한 변환행렬을 통해 함께 변환시키는 방법이다. 평균벡터와 공분산행렬의 변환 수식은 다음과 같다.

$$\begin{aligned}\overline{\boldsymbol{\mu}}_i &= \boldsymbol{\zeta}'\boldsymbol{\mu}_i - \boldsymbol{\epsilon}', \quad \overline{\boldsymbol{\Sigma}}_i = \boldsymbol{\zeta}'\boldsymbol{\Sigma}_i\boldsymbol{\zeta}'^T \\ \overline{m}_i &= \chi' m_i - \nu', \quad \overline{\sigma}_i^2 = \chi'\sigma_i^2\chi'\end{aligned} \tag{4}$$

여기서 행렬 $\boldsymbol{\zeta}'$와 상수 χ'는 상태 i에서의 출력 및 지속시간 확률분포의 평균과 공분산을 함께 변환하는 스케일 값이며 벡터 $\boldsymbol{\epsilon}'$와 상수 ν'는 바이어스다.

2.1.3 Structural Constrained Maximum *a posteriori* Linear Regression (CSMAPLR)

CSMAPLR은 상태출력과 지속시간 분포에 대한 평균벡터와 공분산행렬의 변환행렬을 Structural MAP[7]을 이용하여 추정한다. 이 방법은 회귀 트리(regression tree)의 root 노드(node)로부터 최하위 노드까지 모든 노드에 대해 부모 노드로부터의 사전정보를 이용하여 자녀 노드의 변환행렬을 구함으로써 트리구조에서 각 분포간의 유사 정도와 연결성을 반영하며, CMLLR보다 트리구조를 효율적으로 이용할 수 있다.

2.1.4 CSMAPLR+MAP 결합방식

선형 회귀(linear regression)를 이용한 화자적응 방식들은 평균모델로부터 대상화자 모델로의 선형변환을 가정한 것인데, 선형변환의 제약성을 MAP를 이용해 보완할 수 있다. MAP 결합 알고리즘은 먼저 선형변환을 통해 화자적응을 하고 충분한 적응 데이터가 있는 모델에 한하여 MAP를 이용하여 재추정하는 방법이다. 본 논문에서는 [4]에서와 같이 CSMAPLR에 MAP을 결합한 방식을 사용하였다.

2.2 한국어 HTS에서의 화자적응 성능비교

본 논문에서는 HMM 기반의 한국어 음성합성에서의 화자적응 성능을 비교하기 위하여 객관적 및 주관적 실험을 수행하였다. 음성 DB로는 (주)보이스웨어에서 제공한 화자 당 약 2시간 30분 정도 분량의 남성 5명과 여성 5명의 데이터 중 각각 4명은 평균음성모델의 훈련에, 그리고 나머지 1명씩을 적응 실험에 사용하였다. 샘플링 주파수는 16 kHz이며, log F0, 0차를 포함한 39차 mel-generalized 켑스트럼과 5대역 비주기성(aperiodicity) 및 이들의 차분값, 차분-차분값을 이용하였다. 훈련 및 적응 과정은 HTS-2.2를 이용한 STRAIGHT 버전 적응 데모 시나리오[9]를 한국어 음소 특성에 맞추어 question set 및 label 등을 수정하여 사용하였고, 전역 분산(global variance)은 mel-generalized 켑스트럼에 대해서만 적용하였다. 레이블 정보로는 강제정렬(forced alignment)를 이용한 자동음소분할 정보와 문맥정보(full-context) 20개를 사용하였다. 회귀 트리는 SAT 모델 생성에 사용한 문맥정보 및 음성학적 음소분류가 포함된 결정트리를 사용하였으며, 이 때 문턱치는 mel-generalized 켑스트럼, log F0, 지속시간에 대해 각각 100, 1000, 100으로 실험하였다. 본 논문에서는 5분 정도의 적은 적응 데이터에 대해 우수한 화자적응 성능을 얻는 것을 목표로 하여 주관적 평가는 5분의 적응 데이터에 대해서만 평가를 하였으나, 객관적 평가의 경우 5분 이외에 10분 및 30분의 적응 데이터에 대해서도 참고용으로 실험결과를 제시하였다.

2.2.1 객관적 성능평가

객관적 평가에서 원음성과 합성음의 mel-generalized 켑스트럼과 log F0를 비교하기 위하여 평균 mel-generalized 켑스트럼 거리와 log F0의 root mean square error(RMSE)를 이용하였다. 이를 위해 원음성을 화자종속 모델로 상태 정렬(state alignment) 한 결과를 사용하였다. Mel-generalized 켑스트럼 평가에서 silence와 short pause 구간은 제외시켰고, log F0는 원음성과 합성음 모두 유성음 구간에서만 계산 하였다. 지속시간 평가의 경우에도 RMSE를 이용하되, 원음성의 지속시간을 수작업으로 정확하게 구하기 어려운 점을 고려하여 화자종속 모델로부터 구한 지속시간과 비교하였으며, 역시 silence와 short pause 구간은 제외하였다. 테스트 문장으로는 훈련과 적응에 사용되지 않은 30문장을 사용하였다.

<그림 2>는 객관적 평가결과를 나타낸 것이다. 5분 길이의 적응 데이터에 대해 mel-generalized 켑스트럼과 log F0는 CSMAPLR+MAP 방식이 가장 좋은 성능을 나타내었고, 지속시간은 MLLR 방식에서 가장 좋은 성능을 나타내는 것을 확인할 수 있다.

2.2.2 주관적 성능평가

객관적 평가에서 CSMAPLR+MAP와 MLLR 방식이 5분 길이의 적응 데이터에 대해 상대적으로 좋은 성능을 나타내었기 때문에, 이들 두 방식에 대해 주관적 평가를 수행하였다. 주관적 평가를 위해 ABX text를 수행하였으며, 객관적 성능평가에 사용한 30문장 중 10문장을 평가에 사용하였고 청취 평가자의 수는 10명이다.

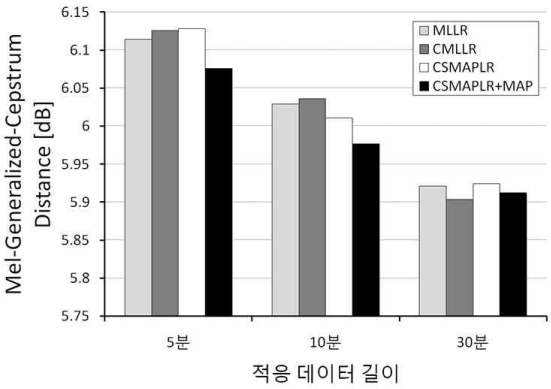
<그림 3>은 주관적 성능 평가 결과를 나타낸 것이다. 객관적 성능평가 결과에서 켑스트럼 및 F0 특성에서 우수한 성능을 나타내었던 CSMAPLR+MAP 방식이 60% 이상의 결과를 얻어 MLLR 방식보다 더 높은 성능을 나타내었다.

3. 화자적응에서 지속시간 모델 개선 방식

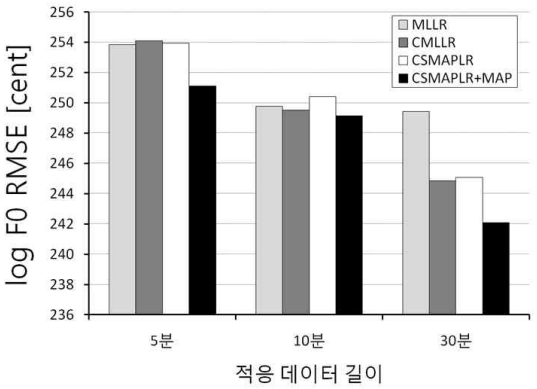
<그림 2>의 객관적 성능평가와 관련하여 SAT 방식에 의한 평균음성모델의 경우 평균 mel-generalized 켑스트럼 거리가 6.96 dB, log F0 RMSE가 257.7 cent, 그리고 지속시간 RMSE가 3.92 frame으로 얻어졌다. 이 결과를 <그림 2>와 비교해 보면 켑스트럼과 log F0 관점에서는 화자적응을 통해 성능향상이 이루어진 반면, 지속시간에 대해서는 오히려 대부분의 화자적응 방식이 평균음성모델보다 낮은 성능을 보였으며, 적응 데이터가 늘어나더라도 성능향상이 관찰되지 않았다.

이러한 현상의 원인을 살펴보는 과정에서 양수 값만 가지는 지속시간이 화자적응 이후 간혹 음수로 나타나는 것을 발견하였다. 이는 지속시간 모델이 실수 전체범위에서 정의되는 Gaussian 모델을 사용하기 때문이며[3], 소량의 적응 데이터로 화자적응을 수행하면 회귀 트리로 구한 모델 클러스터에서의 특정 모델에 대해서는 변환행렬이 잘못 추정되기 때문이다.

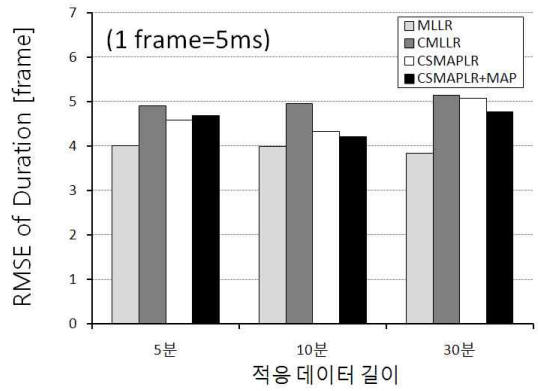
특히 캡스트럼과 log F0의 경우 매 frame 마다 파라미터를 얻는 반면 지속시간은 매 음소 마다 파라미터를 얻기 때문에 동일한 크기의 적응 데이터가 주어지더라도 지속시간 모델에 대한 적응 데이터가 더 적게 되고, 그 결과 화자적응 성능에도 부정적인 영향을 준 것으로 추정된다.



(a)



(b)



(c)

그림 2. 화자적응 방식들의 객관적 성능비교
(a) 평균 mel-generalized 캡스트럼 거리
(b) log F0의 RMSE (c) 지속시간의 RMSE

Figure 2. Objective evaluation of speaker adaptation methods. (a) average mel-generalized-cepstral distance (b) RMSE of log F0 (c) RMSE of duration

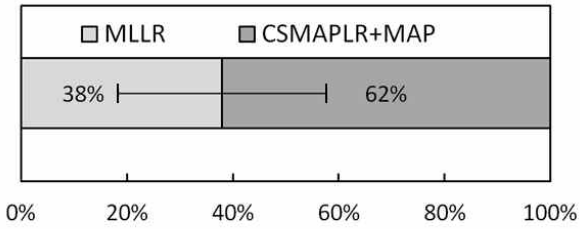


그림 3. MLLR과 CSMAPLR+MAP의 주관적 성능비교 (적응 데이터 길이: 5분)

Figure 3. Subjective evaluation of MLLR and CSMAPLR+MAP (Adaptation data: 5 min.)

본 논문에서는 잘못된 지속시간 모델로 인한 성능저하 현상을 줄이기 위하여 지속시간 교정(duration rectification(DR)) 과 지속시간 보간(duration interpolation(DI)) 방식을 제안한다.

3.1 지속시간 교정 방식

HTS toolkit에서는 지속시간을 생성할 때 지속시간 모델로부터 구한 값을 반올림하여 frame 개수인 양의 정수로 변환해 준다. 반올림을 함으로써 나타나는 원래 지속시간과의 차이는 다음 상태의 지속시간에서 빼줌으로써 보정한다.

지속시간 모델의 평균값은 상태의 frame 개수를 나타내기 때문에 양수여야 하지만 음수가 나타날 경우 HTS toolkit에서는 지속시간을 1로 고정한다. 그리고 반올림과 마찬가지로 1로 고정함으로써 나타나는 원래 지속시간과의 차이를 다음 상태의 지속시간을 구할 때 빼줌으로써 보정하는데, 음수 지속시간이 큰 음수로 나타날 경우 그 다음 상태부터 지속시간이 모두 1로 변환되어 특정구간의 음성이 매우 빠르게 발음되는 현상이 나타난다.

지속시간 교정(DR) 방식은 화자적응모델에서 모델의 평균이 음수인 모델을 평균이 0인 모델로 교정하는 방식으로써 HTS toolkit의 한계를 보상하여 음수모델로 인해 특정 구간에서 지속시간이 매우 빠르게 나타나는 현상을 줄일 수 있다.

3.2 지속시간 보간 방식

지속시간 교정 방식은 음수로 나타나는 모델을 교정함으로써 큰 오류를 줄여줄 수 있지만, 그 외에 대상화자 모델과 다른 방향으로 적응 된 모델은 보정해 줄 수 없다.

지속시간 모델의 경우 SAT 방식에 의한 평균음성모델이 화자적응모델보다 더 좋은 성능을 나타낸다. 지속시간 보간(DI) 방식은 이러한 결과를 바탕으로 SAT 방식에 의한 평균음성모델과 화자적응모델을 다음 식과 같이 가중치 α 를 이용하여 지속시간 모델을 구함으로써 지속시간 교정 방식의 부족한 부분을 보정해 줄 수 있다.

$$\hat{\lambda} = \alpha\lambda_{AV} + (1-\alpha)\lambda_{SA} \quad (5)$$

여기서 $\hat{\lambda}$ 은 지속시간 보간 방식을 이용하여 구한 화자적응 모델이고 λ_{AV}는 평균음성모델, 그리고 λ_{SA}는 화자적응모델이다.

4. 실험 및 결과

본 논문에서 제안한 지속시간 교정 및 보간 방식의 성능을 평가하기 위해서 객관적 성능평가와 주관적 성능평가를 수행하였으며 실험환경은 2.2절과 동일하다.

4.1 객관적 성능평가

본 논문에서 제안한 방식의 성능을 평가하기 위해서 기존의 화자적응 방식에 지속시간 교정만을 적용한 것과 지속시간 교정과 보간을 함께 적용한 것(DR+DI)에 대해 객관적 성능평가를 수행하였다. 지속시간 보간에서 α 값은 별도의 음성 DB(development set)을 통해 최적화를 해야 하나 본 연구에서는 이러한 DB를 구할 수 없어서 일단 0.5를 사용하였다.

<그림 4>는 화자적응 방식 별 성능평가 결과를 나타낸 것이다. 모든 화자적응 방식에서 지속시간 교정과 보간을 함께 적용한 것(DR+DI)이 가장 좋은 성능을 나타냈다.

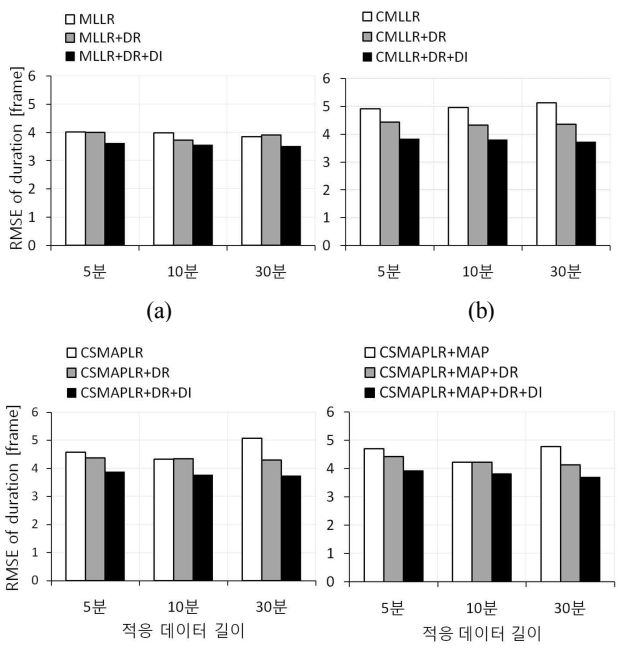


그림 4. 기존의 화자적응 방식과 제안된 방식의 객관적 성능비교 (a) MLLR (b) CMLLR (c) CSMAPLR (d) CSMAPLR+MAP

Figure 4. Objective evaluation of the conventional speaker adaptation methods and proposed methods. (a) MLLR (b) CMLLR (c) CSMAPLR (d) CSMAPLR+MAP

성능평가에서 단순히 대상화자 모델과 화자적응모델 간의 지속시간 오차의 평균이 줄어드는 것도 중요하지만 지속시간 오차가 크게 나타나는 outlier들이 줄어드는 것이 더욱 중요하기 때문에 지속시간 오차의 히스토그램을 구하여 비교하였다. <그림 5>는 CSMAPLR+MAP방식에 지속시간 교정과 보간을 적용한 결과를 지속시간 오차의 히스토그램으로 나타낸 것이다. 그림에서 8 frame 이상 차이나는 outlier의 크기가 지속시간 교정과 보간 방식을 함께 적용했을 때 현저히 줄어드는 것을 확인할 수 있다.

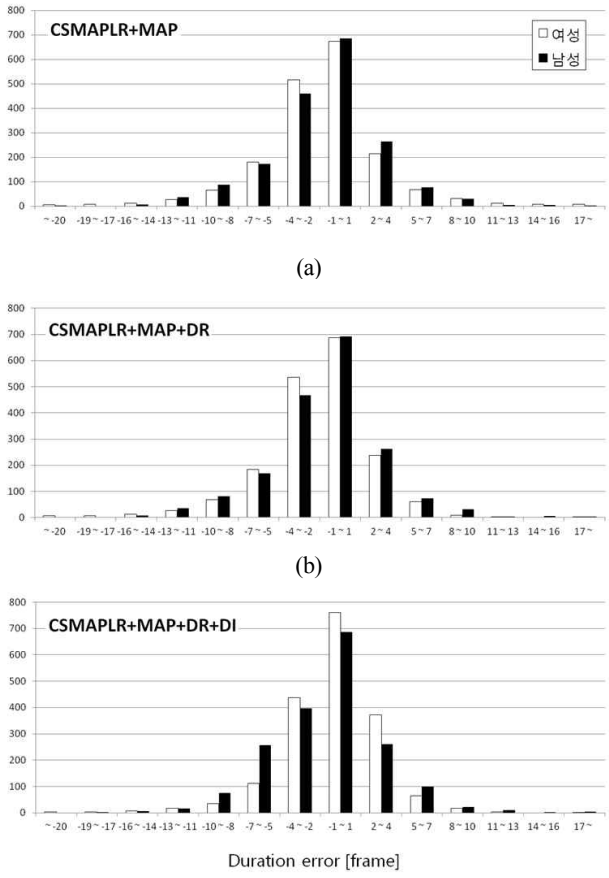
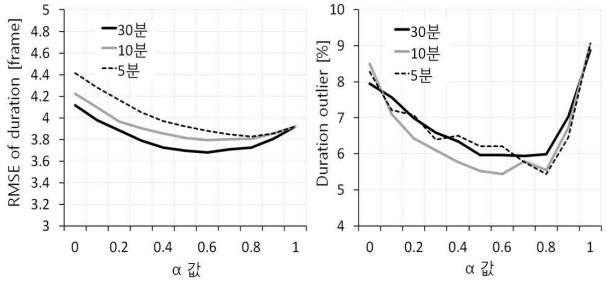


그림 5. 제안한 방식의 적용 여부에 의한 CSMAPLR+MAP의 지속시간 오차 히스토그램 (적응데이터: 5분)
(a) CSMAPLR+MAP (b) CSMAPLR+MAP+DR
(c) CSMAPLR+MAP+DR+DI

Figure 5. Duration error histograms of CSMAPLR+MAP with and without the proposed methods (Adaptation data: 5 min.)
(a) CSMAPLR+MAP (b) CSMAPLR+MAP+DR
(c) CSMAPLR+MAP+DR+DI

지속시간 보간 방식에서 α 값에 따른 성능 특성을 보기 위해 α 값을 변화 시켜가며 지속시간 RMSE와 지속시간 outlier의 비율을 통해 성능을 비교해 보았다. <그림 6>은 대표로 CSMAPLR+MAP에 지속시간 교정과 보간을 함께 적용한 방식에 대해서 α 값을 변화 시켜가며 성능을 비교한 것이다. 여기서 α 가 0일 때는 CSMAPLR+MAP에 지속시간 교정

을 적용한 모델이며, α 가 1일 때는 SAT 방식에 의한 평균 음성모델을 의미한다. 결과를 볼 때, 적응 데이터가 많아질수록 최적의 α 값이 줄어들며 적응 데이터가 적을수록 최적의 α 값이 커지는 경향이 있음을 알 수 있다.



(a) (b)

그림 6. CSMAPLR+MAP+DR+DI에서 α 값에 따른 성능비교 (적응데이터: 5분)
(a) 지속시간 RMSE[frame] (b) 지속시간 outlier[%]

Figure 6. Performance comparison of CSMAPLR+MAP+DR+DI according to α value. (Adaptation data: 5 min.)
(a) RMSE of duration[frame] (b) outlier of duration[%]

4.2 주관적 성능평가

객관적 평가에서 지속시간 교정과 보간(DR+DI)을 함께 적용한 방식이 가장 좋은 성능을 나타내었기 때문에 기존의 화자적응 방식 중 가장 좋은 성능을 나타낸 CSMAPLR+MAP와 CSMAPLR+MAP에 지속시간 교정과 보간을 함께 적용한 방식에 대해 주관적 성능평가를 수행하였다. 평가방식은 2.2.1절과 마찬가지로 ABX test를 이용하였으며, 객관적 성능평가에 사용한 30문장 중 10문장을 평가에 사용하였고 청취 평가자의 수도 동일하게 10명이다.

<그림 7>은 주관적 성능 평가 결과를 나타낸 것이다. 객관적 성능평가 결과와 마찬가지로 CSMAPLR+MAP에 지속시간 교정과 보간을 함께 적용한 방식이 63%를 얻어 37%인 기존의 CSMAPLR+MAP 방식 보다 높은 성능을 나타내었다.

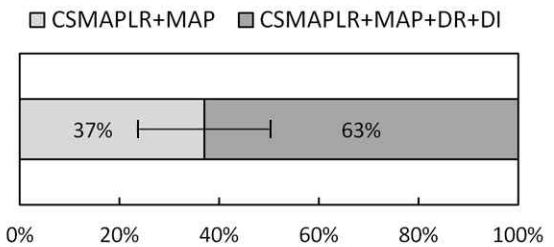


그림 7. CSMAPLR+MAP와 CSMAPLR+MAP+DR+DI의 주관적 성능비교 (적응 데이터 길이: 5분)

Figure 7. Subjective evaluation of CSMAPLR+MAP and CSMAPLR+MAP+DR+DI (Adaptation data: 5 min.)

5. 결 론

본 논문에서는 HMM기반의 한국어 음성합성에서 기존의 대표적인 화자적응 방식들의 성능을 비교하였다. 그 결과 5분 정도의 적은 적응 데이터로 화자적응을 할 경우 CSMAPLR+MAP 방식이 객관적 및 주관적 성능평가에서 가장 좋은 성능을 나타내었다. 그런데 객관적 성능평가 과정에서 켑스트럼 및 F0와 달리 지속시간의 경우 화자적응을 수행해도 평균음성 모델보다도 성능이 개선되지 않는 문제가 관찰되었다. 이러한 문제를 개선하기 위해 본 논문에서는 지속시간 교정 방식과 지속시간 보간 방식을 제안하고 객관적 및 주관적 성능평가로 검증하였다. 그 결과 지속시간 교정과 보간 방식을 함께 적용한 방식이 기존의 화자적응 방식만을 이용하는 것보다 더 좋은 성능을 나타냄을 확인하였다.

참고문헌

[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*, 2347-2350.

[2] http://www.synsig.org/index.php/Blizzard_Challenge_2012_Workshop.

[3] Yamagishi, J. & Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Inf. Syst*. E90-D(2), 533-543.

[4] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. & Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Audio, Speech, Language Process.*, 17(1), 66-83.

[5] Yamagishi, J., Ogata, K., Nakano, Y., Isogai, J. & Kobayashi, T. (2006). HSMM-based model adaptation algorithms for average-voice-based speech synthesis. *Proc. ICASSP*, 77-80.

[6] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis. *Proc. of ICSLP*, 1397-1400.

[7] Shinoda K. & Lee, C.-H. (2001). A structural Bayes approach to speaker adaptation. *IEEE Trans. Speech, Audio Process.*, 9(3), 276-287.

[8] Yamagishi, J. & Kobayashi, T. (2005). Adaptive training for Hidden semi-Markov model. *Proc. ICASSP*, 365-368.

[9] http://hts.sp.nitech.ac.jp/archives/2.2/HTS-demo_CMU-ARCTIC-SLT_STRAIGHT.tar.bz2.

[10] Lee, H. & Kim, H. S. (2012). Performance comparison of speaker adaptation methods for HMM-based Korean speech

synthesis system. *Proc. of Spring Conference of Korean Society of Speech Sciences,* 241-242.
(이혜민, 김형순 (2012). HMM 기반의 한국어 음성합성에서의 화자적응 방식 성능 비교. 한국음성학회 봄 학술대회, 241-242.)

• **이혜민 (Lee, Heamin)**
부산대학교 전자전기공학과
부산시 금정구 부산대학로 63번길 2
Tel: 051-510-1704 Fax: 051-510-1704
Email: oasistony@pusan.ac.kr
관심분야: 음성인식, 음성합성
현재 전자전기공학과 대학원 석사과정 재학 중

• **김형순 (Kim, Hyung Soon)**, 교신저자
부산대학교 전자공학과
부산시 금정구 부산대학로 63번길 2
Tel: 051-510-2452 Fax: 051-515-5190
Email: kimhs@pusan.ac.kr
관심분야: 음성인식, 음성합성, 음성신호처리
현재 전자공학과 교수