논문 13-38C-02-10

한국통신학회논문지 '13-02 Vol.38C No.02
http://dx.doi.org/10.7840/kics.2013.38C.2.213

인자화된 최대 공산선형회귀 적응기법을 적용한 해양IT융합기술을 위한 HMM기반 음성합성 시스템

성 준 식◆, 홍 두 화*, 정 민 아**, 이 연 우**, 이 성 로**, 김 남 수°

Factored MLLR Adaptation for HMM-Based Speech Synthesis in Naval-IT Fusion Technology

June Sig Sung◆, Doo Hwa Hong*, Min A Jeong**,
Yeonwoo Lee**, Seong Ro Lee**, Nam Soo Kim°

요 약

은닉 마코프 모델 (hidden Markov Model, HMM) 기반 음성 합성 시스템에서 파라미터 적응을 위해 널리 쓰이는 기법으로 최대 공산 선형 회귀 (maximum likelihood linear regression, MLLR)이 있다. 이전 연구에서 우리는 각 MLLR 파라미터를 인자화된 MLLR (Factored MLLR, FMLLR) 형태로 확장하는 형태를 제안하였다. FMLLR 파라미터를 기존의 EM 알고리즘 형태로 구하는 기법 역시 제안하였고, 이를 통해 보완 정보를 활용하여 적응 학습을 수행할 수 있게 하였다. 본 논문에서는, FMLLR 기법을 스펙트럼 파라미터에 사용하는 것뿐 아니라 피치에도 적용하여 그 성능을 향상시키는 것에 대한 탐구를 수행하였다. 감정 음성을 생성하는 여러 실험을 통해, 우리는 제안하는 기법이 피치 및 스펙트럼에 대해 효과적으로 작용하는 것을 확인하였다.

Key Words : Speech synthesis, adaptation, MLLR, HMM, expressive speech

ABSTRACT

One of the most popular approaches to parameter adaptation in hidden Markov model (HMM) based systems is the maximum likelihood linear regression (MLLR) technique. In our previous study, we proposed factored MLLR (FMLLR) where each MLLR parameter is defined as a function of a control vector. We presented a method to train the FMLLR parameters based on a general framework of the expectation-maximization (EM) algorithm. Using the proposed algorithm, supplementary information which cannot be included in the models is effectively reflected in the adaptation process. In this paper, we apply the FMLLR algorithm to a pitch sequence as well as spectrum parameters. In a series of experiments on artificial generation of expressive speech, we evaluate the performance of the FMLLR technique and also compare with other approaches to parameter adaptation in HMM-based speech synthesis.

Ⅰ. 서 론

해양IT융합기술에서 사용자와의 이상적인 최종

※ 본 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원(No. 2012R1A2A2A01045874) 및 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임(2009-0093828).
◆ 주저자 : 서울대학교 전기컴퓨터공학부 뉴미디어통신공동연구소, jssung@hi.snu.ac.kr, 학생회원
° 교신저자 : 서울대학교 전기컴퓨터공학부 , nkim@snu.ac.kr, 종신회원
* 서울대학교 전기컴퓨터공학부 뉴미디어통신공동연구소, dhhong@hi.snu.ac.kr, 학생회원
** 목포대학교, majung@mokpo.ac.kr, 정회원, ylee@mokpo.ac.kr, 종신회원, srlee@mokpo.ac.kr, 정회원
논문번호 : KICS2013-01-037, 접수일자 : 2013년 1월 14일, 최종논문접수일자 : 2013년 2월 19일

통신 방법은 육성을 통해 지시하고 지시한 결과를 음성으로 보고받는 것이다. 이는 선박을 조정하는 사용자를 조타 혹은 선박 운항장치의 조정에 집중할 수 있게 하여 운항 과정에서 발생할 수 있는 사고를 완화하는데 큰 역할을 하게 된다. 사용자의 취향에 맞는 음성을 생성하기 위해서는 해당 스타일의 음성을 모델로 가지는 합성음 시스템을 필요로 하는데, 요구되는 스타일이 다양할 경우 각 경우마다 합성음을 위한 원본 음성을 채집하기에는 많은 노력을 요구한다. 이러한 간극을 줄이기 위해서는 파라미터 적응 학습의 사용이 필수적이다. 파라미터 적응 기법은 화자, 채널 및 발성 환경의 차이에 따른 학습 데이터 및 테스트 데이터 간의 차이를 줄이기 위해 사용된다. 최대 공산 선형 회귀 (maximum likelihood linear regression, MLLR) 기법은 가장 널리 쓰이는 적응 기법 중 하나로, 은닉 마코프 모델 (hidden Markov Model, HMM) 기반 시스템의 원래 파라미터를 의사 변환 매트릭스 (affine transformations)를 통해 작은 양의 적응 파라미터로 사상하여 그 차이가 최소화될 수 있도록 하는 것이다[1]. 의사 변환 매트릭스로 나타나는 MLLR 파라미터는 일반적으로 최대 공산 (maximum likelihood, ML) 방식에 따라 추정된다. 보통 적응 파라미터의 양이 작기 때문에 MLLR 파라미터들은 비슷한 발성을 보이는 음성들 간 공유하게 되며 각 파라미터로의 접근은 현재 발성을 위한 음성의 종류에 따라 구분된다. MLLR 기법은 음성 인식 분야뿐 아니라 음성 합성 분야에서도 널리 쓰이고 있다[2-4].

그러나 감정 및 가창 음성 모델링의 경우 음성의 종류만으로 파라미터를 구분하는 것은 불충분하다. [5]에서 보이는 바와 같이, 같은 음성이라도 발성음의 높이 및 감정 정도에 따라 스펙트럼은 크게 다른 양상을 보이기 때문이다. 따라서 기존의 MLLR 방식 만으로의 접근은 충분한 성능을 보장할 수 없다. 이를 위해 우리는 기존의 MLLR을 인자화된 MLLR (Factored MLLR, FMLLR)로 확장하는 기법을 제안했었다[6-8]. FMLLR은 각 MLLR 파라미터는 제어 벡터의 함수로 정의되며, 이는 회귀 함수와 제어 벡터를 변환한 파라미터와의 내적에 따라 구하는 형태로 정의했다. 우리는 제안한 FMLLR 기법을 reading 스타일 음성의 스펙트럼 파라미터를 가창 스타일 음성으로 변환하는 것에 적용해 보았다. 가창 스타일 음성의 스펙트럼은 가사에 따른 음성뿐 아니라 악보 정보에 따라 다양하게 나타나게 되는데 [5], 이를 표현하기 위해 악보 정보를 제어 벡터 파라미터로 하는 FMLLR을 통해 가창 음성 합성을 수행하였다.

본 논문에서는 FMLLR을 스펙트럼에만 적용하였던 [7]을 확장하여 피치 정보 역시 FMLLR로 변환하는 과정을 수행하였다. 피치가 제어 벡터의 파라미터로 사용되는 가창 음성과 달리 감정 음성에서는 감정의 정도를 제어 벡터로 표현하기 때문에 피치 역시 FMLLR framework에서 변환이 가능하게 된다. 기존의 기법들과의 비교를 수행하여 제안하는 기법이 나은 성능을 보임을 실험을 통해 입증하였다.

Ⅱ. 본 론

2.1. MLLR

기존의 MLLR 적응 기법은 mean 벡터 μ_s를 아래의 수식에 따라 변환된 $\hat{\mu_s}$로 나타난다.

$$\hat{\mu_s} = M\mu_s + b \tag{1}$$

M은 p by p 회귀 매트릭스이며 b는 bias 벡터이다. 분포 s는 평균 μ_s와 공분산 Σ_s의 Gaussian으로 가정된다. 추가적으로 우리는 M과 Σ_s를 diagonal로 가정하였다. M, b의 각 파라미터는 ML 기법에 따라 추정하며, 이를 위해 우리는 EM 알고리즘을 적용하였다. $X = (x_1, x_2, ..., x_T)$ 를 주어진 적응 데이터로 정의하고, 그에 따라 다음과 같은 수식을 구할 수 있다[1].

$$\begin{bmatrix} \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}^2}{\sigma_{s,i}^2}\right) & \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}}{\sigma_{s,i}^2}\right) \\ \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}}{\sigma_{s,i}^2}\right) & \left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}\right) \end{bmatrix} \begin{bmatrix} \hat{M_i} \\ \hat{b_i} \end{bmatrix} = \begin{bmatrix} \left(\sum_{t=1}^{T}\gamma_t(s)\frac{x_{t,i}\mu_{t,i}}{\sigma_{s,i}^2}\right) \\ \left(\sum_{t=1}^{T}\gamma_t(s)\frac{x_{t,i}}{\sigma_{s,i}^2}\right) \end{bmatrix} \tag{2}$$

여기서 $\gamma_t(s)$는 E step에서 구한 사후 확률, $x_{t,i}$, $\mu_{s,i}$ 및 $\sigma_{s,i}^2$는 각각 x_t, μ_s, Σ_s의 i번째 원소를 뜻한다.

2.2. MRHSMM

MRHSMM은 여러 감정에 대해 적응 학습을 각

각 수행한 결정 트리가 존재할 때, 이를 하나의 결정 트리에서 감정 정도를 인자로 하여 조절하는 형태의 표현을 하는 방법이다[8]. MLLR은 만들어진 커다란 크기의 결정 트리에 적은 양의 데이터를 사용하여 각 트리의 node 값을 변환해주는 과정이기 때문에, 같은 결정 트리로 적응 학습을 수행한 결과는 같은 형태의 결정 트리를 갖게 된다. 따라서 감정 혹은 화자의 수가 증가할 때마다 그 크기가 선형으로 증가하는데, MRHSMM은 각 node에서 다양한 감정 혹은 화자를 아우르는 multiple regression matrix를 새로 생성하고, 각 감정 혹은 화자를 표현하는 파라미터를 통해 원하는 결과를 만드는 것을 목적으로 한다.

MRHSMM에서, μ_s는 다음과 같은 형태로 정의한다.

$$\mu_s = H_s\xi \tag{3}$$

여기서 ξ는 감정 혹은 화자 스타일의 정도를 표현하는 L차의 제어 파라미터를 의미하며, H_s는 $p \times L$ 차의 multiple regression matrix를 표현한다. 앞에서 주어진 적응 데이터 $X = (x_1, x_2, ..., x_T)$를 통해 표현되는 H_s의 i번째 원소는 다음과 같이 표현된다.

$$\widehat{H}_s(i) = \left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}x_{t,i}\xi_t'\right)\left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}\xi_t\xi_t'\right)^{-1} \tag{4}$$

2.3. FMLLR

M과 b가 제어 파라미터 η에 의해 표현되는 것을 생각해보자. (1)은

$$\widehat{\mu_s} = M(\eta)\mu_s + b(\eta) \tag{5}$$

와 같이 각 변환 매트릭스 및 벡터는 η를 인자로 갖는 형태로 변형이 가능하다. 여기서, $M(\eta)$이 diagonal한 성분만을 갖고 있는 것을 가정하면, 각 변환 수식은 다음과 같이 표현할 수 있다.

$$\begin{aligned} M(\eta) &= diag(w_1'\xi, w_2'\xi, \ldots, w_p'\xi) \\ b(\eta) &= (v_1'\xi, v_2'\xi, \ldots, v_p'\xi) \end{aligned} \tag{6}$$

이에 따라, 추정 수식은 다음과 같은 모양으로 구성할 수 있다.

$$\begin{aligned} &\begin{bmatrix} \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}^2}{\sigma_{s,i}^2}\xi_t\xi_t'\right)\left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}}{\sigma_{s,i}^2}\xi_t\xi_t'\right) \\ \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}}{\sigma_{s,i}^2}\xi_t\xi_t'\right)\left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}\xi_t\xi_t'\right) \end{bmatrix}\begin{bmatrix} \widehat{w_i} \\ \widehat{v_i} \end{bmatrix} \\ &= \begin{bmatrix} \left(\sum_{t=1}^{T}\gamma_t(s)\frac{x_{t,i}\mu_{t,i}}{\sigma_{s,i}^2}\xi_t\right) \\ \left(\sum_{t=1}^{T}\gamma_t(s)\frac{x_{t,i}}{\sigma_{s,i}^2}\xi_t\right) \end{bmatrix} \end{aligned} \tag{7}$$

여기서 ξ_t는 시간 t에서 η_t에 따라 변형된 제어 벡터이다. 자세한 유도 과정은 [7]을 참고하여 확인할 수 있다.

Ⅲ. 실 험

성능 증명을 위한 실험으로, 본 논문에서는 낭독형 음성을 감정 음성으로 적응하여 그 성능을 비교하였다. 실험을 위해, 525분길이 (4,000문장)의 낭독형 남성 음성을 HMM 기반 음성 모델 생성에 사용하였다. 각 음성 파일은 16 kHz로 샘플링 되었으며 5 ms 간격의 frame shift, 20 ms Hamming window를 적용하였다. 피치를 위한 feature로 1차의 log-scaled f0를 사용하였고 ($\log(F0)$), 여기에 delta 및 delta-delta 값을 추가하여 3차의 feature를 사용하였다. 모델링 과정은 [9]과 같다. 합성을 위한 기본 유닛으로 triphone을 사용하였으며, 각 triphone은 5 state left-to-right with no skip HMM으로 이루어졌다. 각 state는 single Gaussion PDF로 구성되었다. 적응 학습을 위해 평균 21분 길이의 (254문장) 남성 감정 음성을 확보하였고, 위와 같은 방식으로 파라미터를 추출하였다. 실험을 위해 사용한 제어 벡터는 (1, 1, 0, 0, 0), (1, 0, 1, 0, 0), (1, 0, 0, 1, 0), (1, 0, 0, 0, 1)이며, 이는 각각 화남, 슬픔, 즐거움, 두려움 감정에 해당한다.

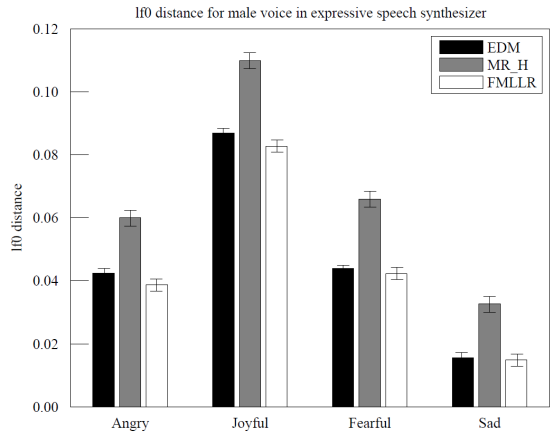


그림 1. 남성 음성 평균 lf0 distance. EDM은 MLLR을, MR_H는 MRHSMM기법을 뜻함.
Fig. 1. Average lf0 distance for male voice. EDM denotes MLLR, MR_H indicates MRHSMM.

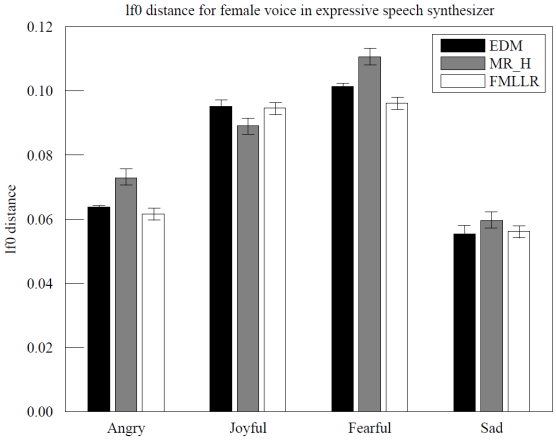


그림 2. 여성 음성 평균 lf0 distance.
Fig. 2. Average lf0 distance for female voice.

비교를 위해, 학습한 낭독형 음성으로부터 각 감정으로의 적응 학습을 기존 기법 및 제안하는 기법에 대해 피치 에 대하여 적용하였다. 스펙트럼에 대해 적용한 결과는 [7]을 참고할 수 있다. Objective test를 위해 실제 lf0와 생성한 lf0간의 euclidean distance를 측정하였고 그 결과를 그림 1, 2에서 나타내었다. Euclidean distance의 수식은 다음과 같이 나타나며

$$d(x,y) = \|x - y\|^2 \tag{8}$$

이들의 평균을 통해 값을 나타내었다. 제안하는 기법이 대부분의 경우에서 기존에 더 가까운 모델을 생성함을 확인할 수 있다. 듣기 평가를 위해 20명의 청자에게 각 감정 별로 여섯 문장을 들려주었으며, 평균 의견 점수 (mean opinion score, MOS) 기법으로 점수를 측정하여 그 결과를 그림 3, 4에 나타내었다.

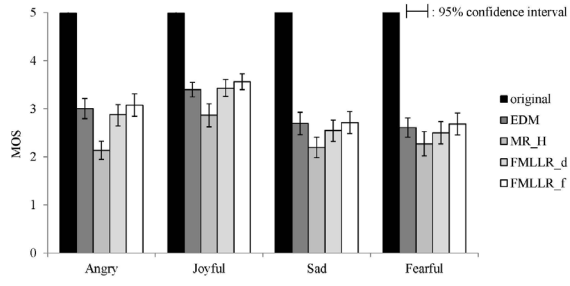


그림 3. 남성 음성의 주관적 듣기 평가 결과. _d 및 _f는 FMLLR을 diagonal 및 full matrix 형태로 가져간 것을 의미.
Fig. 3. Result of subjective listening test for male voice. _d and _f indicate diagonal and full matrix FMLLR, respectively.

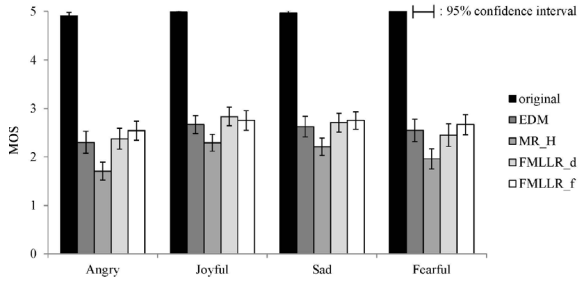


그림 4. 여성 음성의 주관적 듣기 평가 결과.
Fig. 4. Result of subjective listening test for female voice.

각 문장은 피치 뿐 아니라 스펙트럼 역시 각 기법들을 통해 적응 학습을 수행하였다. 실험 결과에서 확인할 수 있듯, 제안하는 기법이 피치에 적용했을 때도 효과적으로 동작하며, 기존 기법들보다 나은 결과를 보여주는 것을 알 수 있다.

Ⅳ. 결 론

본 논문에서는 MLLR에 제어 벡터를 통해 변환 매트릭스를 조정하는 FMLLR 기법을 스펙트럼 뿐 아니라 피치에도 적용하고 그 성능을 평가해 보았다. MLLR이 확장된 형태인 FMLLR을 통해 각 감정을 위한 제어 벡터를 효과적으로 사용하여 실제 음성에 더 가까운 모델을 생성할 수 있도록 하였다. 제안하는 기법을 통해 생성한 합성음은 기존의 기법들에 비해 향상된 품질을 보임을 distance 측정 및 듣기 평가를 통해 증명하였다.

References

[1] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[2] Y. Sung, C. Boulis, and D. Jurafsky, “Maximum conditional likelihood linear regression and maximum a posteriori for hidden conditional random fields speaker adaptation,” *in Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4293–4296

[3] J. Yamagishi et al., “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[4] T. Nose, Y. Kato, and T. Kobayashi, “A speaker adaptation technique for MRHSMM-based style control of synthetic speech,” *in Proc. ICASSP, Honolulu, HI, 2007*, pp. 833–836.

[5] J. Sundberg, “The acoustics of the singing voice,” Sci. Amer., pp. 82–91, Mar. 1977

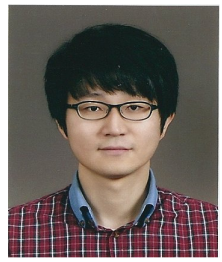
[6] N. S. Kim, J. S. Sung and D. H. Hong, “Factored MLLR adaptation,” *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 99-102, Feb. 2011.

[7] J. S. Sung, D. H. Hong, H. W. Koo and N. S. Kim, “Factored MLLR Adaptation Algorithm for HMM-based Expressive TTS,” *in Interspeech2012*, Portland, Sep. 2012.

[8] J. S. Sung, S. J. Kang, J.-H. Chang, and N. S. Kim, “Factored MLLR Adaptation for HMM-based Singing Voice Synthesis“, *in Proc. KICS Int. Conf. Commun. 2011*, Jeju Island, Korea, June, 2011.

[8] H. Zen et al., “The HMM-based speech synthesis system version 2.0,” *in Proc. of ISCA SSW6*, Bonn, Germany, Aug. 2007.

성 준 식 (June Sig Sung)



2006년 2월 서울대학교 전기컴퓨터공학부 학사

2008년 2월 서울대학교 전기컴퓨터공학부 석사

2008년 3월~현재 서울대학교 전기컴퓨터공학부 박사과정

<관심분야> 신호처리, 음성인식, 음성합성

홍 두 화 (Doo Hwa Hong)

2007년 8월 KAIST 전기전자공학부 학사

2007년 9월~현재 서울대학교 전기컴퓨터공학부 석박통합과정

<관심분야> 신호처리, 음성인식, 음성합성

정 민 아 (Min A Jeong)

1994년 2월 전남대학교 전산통계학과 석사

2002년 2월 전남대학교 전산통계학과 박사

2005년 3월~현재 목포대학교 컴퓨터공학과 부교수

<관심분야> 데이터베이스/데이터마이닝, 생체인식시스템, 무선통신응용분야(RFID, USN, 텔레매틱스), 임베디드시스템

이 연 우 (Yeonwoo Lee)

1994년 2월 고려대학교 전자공학과 석사

2000년 2월 고려대학교 전자공학과 박사

2000년 10월~2003년 12월 영국 Edinburgh 대학교 Research Fellow

2004년 1월~2005년 8월 삼성종합기술원

2005년 9월~현재 국립목포대학교 공과대학 정보통신공학과, 부교수

<관심분야> 해상무선통신, e-Navigation, Cognitive Radio, 4G 이동통신

이 성 로 (Seong Ro Lee)

1987년 2월 고려대학교 전자공학과 공학사
1990년 2월 한국과학기술원 전기및전자공학과 공학석사
1996년 8월 한국과학기술원 전기및전자공학과 공학박사
1997년 9월~현재 목포대학교 공과대학 정보전자공학과 교수
<관심분야> 디지털통신시스템, 이동 및 위성통신시스템, USN/텔레매틱스응용분야, 임베디드시스템

김 남 수 (Nam Soo Kim)

1988년 2월 서울대학교 전자공학과 학사
1990년 2월 한국과학기술원 전기 및 전자공학과 석사
1994년 8월 한국과학기술원 전기 및 전자공학과 박사
1998년 3월~현재 서울대학교 전기컴퓨터공학부 교수
<관심분야> 음성신호처리, 음성인식, 음성합성, 음성향상, 통계적 신호처리, 패턴 인식