Adaptation Data의 Quality를 고려한 강인한 화자 적응

표현아°, 김세현, 오영환
한국과학기술원 전자전산학과 전산학전공

Flexible Speaker Adaptation Reflecting the Quality of Adaptation Data

Hyun-A Pyo°, Se-Hyun Kim, Yung-Hwan Oh
Division of Computer Science
Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology
e-mail : netty@bulsai.kaist.ac.kr

요약

최근 음성 인식 시스템의 성능 향상을 위해 화자 적응 (speaker adaptation)에 대한 연구가 활발히 진행되고 있다. HMM 기반 인식 시스템의 모델 파라미터를 수정하는 화자 적응의 경우, MAP 방법과 MLLR 방법에 대한 연구가 주류를 이루고 있다. 두 방법은 adaptation data의 양에 따라서 서로 다른 성능을 보인다. 본 논문에서는 adaptation data의 quality를 정의하고, 이를 기존 두 방법의 가중치로 이용하여 화자 적응을 수행하는 방법을 제안한다. 제안한 방법을 KAIST 통신연구실에서 구축한 한국어 도시이름 500단어 인식 시스템에 적용하여 성능을 개선하였다.

1. 서론

음성 인식 시스템의 성능이 상용화 단계에 이를 정도로 향상되었지만, 화자나 환경의 불일치로 인하여 인식 성능의 저하를 가져오게 된다. 특정 화자에 대한 학습 자료가 충분할 경우에 화자 종속 (speaker-dependent; SD) 시스템이 화자 독립 (speaker-independent; SI) 시스템보다 2-3배 이상 우수한 성능을 보인다. 그러나 실용시스템의 경우, 한 화자에 대한 충분한 자료를 얻을 수 없으므로, 적은 adaptation data를 이용하여 화자 독립 시스템의 파라미터를 재예측하는 화자 적응 방법에 대한 연구가 활발히 진행되고 있다.

본 논문에서 대상으로 하는 HMM 기반 인식 시스템에서 많이 이용되는 화자 적응 방법은 크게 MAP (Maximum a Posteriori) 방법과 MLLR (Maximum Likelihood Linear Regression) 방법의 두 가지로 나누어 볼 수 있다. MAP 방법은 adaptation data에서 관측된 모델들의 파라미터만을 재예측하므로 adaptation data가 증가할수록 화자 종속 시스템에 접근하게 되어 성능이 높아지게 된다. 반면에 MLLR 방법은 비슷한 특성을 지닌 모델들을 클래스 (class)로 묶어서 선형회귀 방법을 적용함으로써, 적은 adaptation data에 대해서 효과적인 특징을 가지고 있다. 그러나 adaptation

data가 증가할수록 클래스 별로 동일한 변환을 하게 되므로 성능은 현저히 떨어진다. 즉, adaptation data의 양에 따라 두 가지 방법 중 적절한 적응 방법을 선택하여 적용하여야 한다.

본 논문에서는 adaptation data의 quality를 정의하고, 화자 적응을 수행할 때, 기존의 화자 적응 방법에서 구한 파라미터에 가중함으로써 quality에 따라 서로 다른 반영도를 보이는 적응 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 화자 적응 방법인 MAP, MLLR 방법에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 adaptation data의 quality를 이용한 화자 적응 방법에 대해 설명한다. 제안한 방법의 유효성 검증을 위한 실험 방법 및 결과를 4장에서 보인 후, 5장에서 결론을 맺는다.

2. 화자 적응

최근 화자 적응에 대한 연구는 MAP 적응 방법과 MLLR 적응 방법이 주류를 이루고 있는데, 이들은 adaptation data의 양에 따라 각각 장,단점을 가지게 된다[1].

2.1 MAP 적응 방법

MAP 적응 방법은 예측하고자 하는 목적 파라미터를 랜덤 변수로 가정하고 목적 파라미터에 대한 선험 정보를 이용하는 적응 방법이다. Adaptation data X를 이용해 재예측된 파라미터 λ'는 식 (1)과 같이 구할 수 있다.

$$\lambda' = \arg\max_{\lambda} p(\lambda \mid X) = \arg\max_{\lambda} p(X \mid \lambda) p_0(\lambda) \quad (1)$$

$p_0(\lambda)$는 선험 확률값으로 일반적으로 화자 독립 시스템의 파라미터를 사용한다. State s의 Gaussian mixture mean값은 식 (2)와 같이 SI mean과 adaptation data mean의 가중합으로 구할 수 있다[2].

$$\hat{\mu}_s = \frac{N_s}{N_s + \tau} \bar{\mu}_s + \frac{\tau}{N_s + \tau} \mu_s \quad (2)$$

where $\bar{\mu}_s$: *adaptation data mean*

μ_s : *SI mean*

N_s : *adaptation data의 관측 확률*

τ : *weight*

식 (2)에서 보는 바와 같이 MAP 방법은 adaptation data의 양이 증가할수록, N_s 값이 커지게 되므로 SD mean에 접근하게 되어 성능이 증가한다. 그러나 adaptation data의 양이 적은 경우에는 관측된 모델만 수정하게 되어 적은 adaptation data에 대해서는 오히려 성능이 저하된다.

2.2 MLLR 적응 방법

MLLR 적응 방법은 adaptation data가 적은 경우 관측되지 않는 모델이 나타나는 것을 고려하여 유사한 모델들을 클래스로 묶어서 식 (3)과 같은 선형 변환을 통해 화자 적응을 수행하는 방법이다.

$$\hat{\mu}_s = A\mu_s + b = W\xi_s \quad (3)$$

where W : *transformation matrix*

ξ_s : *extended mean vector*

변환 행렬 W를 구하기 위해 식 (4)와 같은 목적 함수를 정의한다. 목적 함수는 adaptation data에 의해 관측된 모델의 우도를 최대화 한다. 목적 함수를 구하기 위해서 보조 함수를 식 (5)와 같이 정의하고, 보조 함수를 최대화하는 W를 식 (6)과 같이 구할 수 있다[3].

$$F(X \mid \lambda) = \sum_{\theta \in \Theta} F(X, \theta \mid \lambda) \quad (4)$$

$$Q(\lambda, \lambda') = \sum_{\theta \in \Theta} F(X, \theta \mid \lambda) \log(F(X, \theta \mid \lambda')) \quad (5)$$

$$\sum_{t=1}^{T} \gamma_s(t) C_s^{-1} o_t \mu_s' = \sum_{t=1}^{T} \gamma_s(t) C_s^{-1} W \mu_s \mu_s' \quad (6)$$

where $\gamma_s(t)$: *time* t *에 state* s *에 머무를 확률*

C_s^{-1} : *inverse covariance matrix*

MLLR 방법은 적은 adaptation data에 대해서는 효과적이지만, adaptation data의 양이 증가하는 경우 각 클래스 별로 동일한 변환 행렬 W를 적용하게 되므로 오히려 성능의 저하를 가져오게 된다.

2.3 Implement issues

화자 독립 시스템에 대해서 화자 적응을 수행하는 경우, 일반적으로 adaptation data의 양에 따라서 MAP와 MLLR 방법 중 선택하게 된다. 즉, adaptation data의 양이 많으면 MAP 방법이 효과적이며, 적으면 MLLR 방법이 더 효과적이다. 그러나 adaptation data의 양뿐만 아니라, 주어진 adaptation data의 quality에 따라서 성능이 크게 좌우된다. 전체적으로 adaptation data의 양이 증가하더라도 특정 모델들에 대해 관측되는 adaptation data가 증가한다면, MAP 방법보다는 MLLR 방법이 더 효과적이다. 실용 시스템이 경우 adaptation data의 변이가 크므로, 화자 적응 방법을 미리 결정하는 것은 효과적이지 못하다. 따라서 화자 적응 방법을 미리 결정하지 않고 입력되는 adaptation data에 따라서 SI, MAP, MLLR에서 얻어진 파라미터들의 중요도를 측정한 후, 각 모델 별로 서로 다른 적응을 수행함으로써 강인한 적응 시스템을 구성하는 방법이 효과적이다.

3. Adaptation data의 quality

3.1 Adaptation data의 quality

Adaptation data의 모델에 대한 분포에 따라서 화자 독립 시스템의 적응 정도가 달라진다. 따라서 adaptation data의 모델에 대한 분포를 quality로 정의하고, adaptation data의 양과 함께 화자 적응시에 같이 고려해야 한다. 본 논문에서는 adaptation data X 의 quality를 $q_{s,X}(\nu_s,\varphi_X)$ 로 정의하고 각각의 파라미터는 다음과 같이 정의한다.

$$\nu_s = \sum_{x\in X}\sum_{t=1}^{T}\gamma_s(t) \quad (7)$$

$$\varphi_X = \frac{\sum_{s\in S}\nu_s}{|S|} \quad (8)$$

ν_s 는 모든 adaptation data에 대해서 state s 에 머무를 확률 $\gamma_s(t)$ 의 합으로 나타냄으로써, 한 state s 에 대한 중요도를 나타내며, φ_X 는 모든 state에 대한 평균 ν_s 값으로, adaptation data X 가 모든 state를 관측하면 1이상의 값을, adaptation data가 적으면 1보다 작은 값을 가지며, adaptation data의 분포를 standard gamma distribution $f(x;\varphi_X)$ 로 표현하게 된다. ν_s 는 φ_X 로 정해진 adaptation data에서 각 state에 대한 관측 정도를 나타낸다.

3.2 Weight class

φ_X 값이 작을수록 전체적으로 MLLR 방법에 의한 결과의 신뢰도가 높지만, ν_s 값도 작은 경우에는 모델 적응의 신뢰도가 낮은 경우를 고려해야 하며, φ_X 값에 따라서 ν_s 값의 신뢰도도 다르다. 따라서 SI mean의 가중 정도, MAP와 MLLR mean의 가중 정도와 가중 증가 정도를 고려해 weight class를 실험적으로 정의한다. φ_X 값에 의해 weight class를 선택하고, 선택된 weight class내에서 ν_s 값으로 weight vector ω_s 를 선택한다.

3.3 Adapted mean

화자 독립 시스템의 state s 의 mean vector μ_s 에 대해서 주어진 adaptation data를 이용해서 각각 MAP 방법과 MLLR 방법으로 $\mu_s^{map}, \mu_s^{mllr}$ 을 얻었다. 본 논문에서 제안한 $q_{s,X}(\nu_s,\varphi_X)$ 값을 이용해서 weight vector $\omega_s = (\alpha_s,\beta_s,\gamma_s)$ 를 선택하고, 구한 mean들의 가중합으로 식 (9)와 같은 새로운 adapted mean을 구한다.

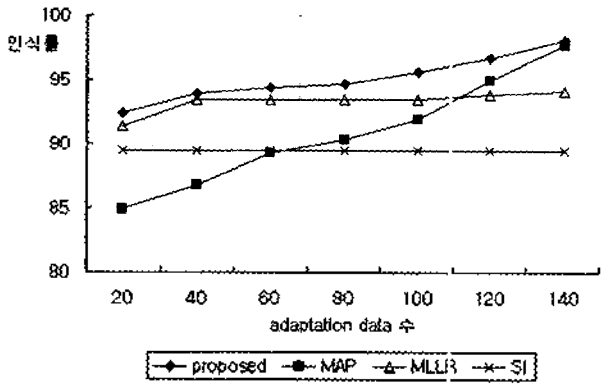
$$\hat{\mu}_s = \alpha_s\mu_s + \beta_s\mu_s^{map} + \gamma_s\mu_s^{mllr} \quad (9)$$

4. 실험 및 결과

본 논문에서 사용된 DB는 KAIST 통신 연구실에서 구축한 한국어 도시이름 500단어 DB이다. 남성화자 34명, 여성화자 14명 중 남녀 각각 26, 10명을 학습단계에서 사용하였고, 인식 및 적응화자로 남녀 각각 8, 4명을 사용하였다. 특징벡터는 MFCC 12차와 에너지를 포함해서 모두 39차를 사용하였고, HMM 모델은

single mixture로 3개의 states로 구성하였다. 화자독립 시스템 및 MAP, MLLR 방법은 HTK를 사용해 실험하였다[4]. Adaptation data set은 각 화자당 20, 40, 60, 80, 100, 120, 140개의 단어로 모두 7개의 set으로 구성하였다.

[그림 1]은 각 adaptation data set에 대해서 12명의 화자에 대한 각각의 인식 성능의 평균을 비교한 그래프이다. 본 논문에서 제안한 방법은 SI 시스템에 대해서 최대 8.63%, MLLR 적응 시스템에 대해서는 3.99%, MAP 적응 시스템에 대해서는 7.4%의 인식률 향상을 보였다.



[그림 1] 인식 성능 비교

Adaptation data의 수가 적은 경우에는 MLLR 방법이 MAP 방법보다 인식률이 높고, 반대로 adaptation data의 수가 많은 경우에는 MAP 방법의 인식률이 높음을 볼 수 있다. 또한 본 논문에서 제안한 방법이 MAP나 MLLR 방법보다 항상 인식률이 높음을 볼 수 있다. 실험 결과를 통해 본 논문에서 제안한 state 별 가중치에 따라 각 적응 방법을 통해 얻은 파라미터 값을 결합하는 방법이 효과적임을 알 수 있다.

5. 결론

본 논문에서는 HMM 기반 화자 독립 시스템에서 특정 화자의 adaptation data를 이용해서 인식 성능을 향상시키는 MAP 방법과 MLLR 방법의 장, 단점에 대해서 살펴보았다. Adaptation data의 quality를 state 별 출현 빈도의 확률모델로 정의하고, 기존의 두 적응 방법으로 얻은 파라미터와 화자 독립 시스템의 파라미터 값의 가중합으로써 파라미터를 재예측하는 방법을 제안하였다. 실험 결과 기존의 적응 방법에 대해서 최고 7.4%의 인식률 향상을 보였다. 화자 적응의 큰 주류를 이루고 있는 두 가지 방법에 대해서 state 별로 반영 정도를 달리함으로써, adaptation data의 quality에 따라 유연하게 파라미터를 재예측하는 것이 효과적임을 알 수 있었다. 현재는 weight class로 분류되어 있는 weight vector의 값을 수식적으로 계산하는 부분에 대한 연구를 수행 중이다.

참고문헌

[1] P.C. Woodland, "Speaker adaptation: techniques and challenges", ASRU, 1999.

[2] J.L. Gauvain and C.H. Lee, "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains.", IEEE Trans. SAP, Vol. 2, pp.291-298, 1994.

[3] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol. 9, pp.171-185, 1995.

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P.C. Woodland, "The HTK Book (for HTK version 3.0)", Microsoft Corporation, 2001.