2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

MLLR을 이용한 한국어 음성의 화자 적응

김태형, 이건웅, 이상호, 홍재근
경북대학교 전자공학과
e-mail : soma@speech.knu.ac.kr

A Speaker Adaptation of Korean Speech Using MLLR

Tae-Hyeong Kim, Keon-Ung Lee, Sang-Ho Lee, Jae-Keun Hong
Dept. of Electronic Engineering, Kyungpook National University

요약

화자 독립 인식은 훈련 화자와 시험 화자의 차이로 인해 화자 종속의 경우보다 인식률이 떨어진다. 따라서, 인식률을 향상시키기 위해 화자 독립 모델을 화자에 적응시킬 필요가 있다. 본 논문에서는 효과적인 적응 방법인 MLLR(Maximum Likelihood Linear Regression) 적응 방법을 한국어 음성에 적용하여 적응 성능을 향상시켰고, 온라인 상에서 적용 가능하도록 증가 적응 방법을 이용하였다. PBW 445 음성 데이타베이스에 대한 실험 결과, 400개의 적응 데이터를 사용하였을 때, 제안한 방법이 기존의 화자 독립 시스템보다 7.02% 향상된 성능을 보였다.

1. 서론

화자 독립 인식 시스템은 수십 년간 발전하여 왔으나, 화자 독립 인식 성능은 여전히 화자 종속 시스템보다 인식률이 낮다. 이것은 훈련 화자와 실험 화자의 음성 특징이 다르기 때문이다. 따라서, 기존의 모델은 다른 새로운 화자에 대해서 좋은 인식 성능을 보여 주지 못한다. 따라서, 기존의 화자 독립 시스템을 새로운 화자가 발음한 데이터로 적응시켜 기존의 모델을 화자에 맞게 재조정하여야 인식률을 향상시킬 수 있다.

화자 적응 방법에는 크게 새로운 화자의 음성을 기존의 화자 독립 인식 시스템에 잘 맞게 적응시키는 정규화 방법[5]과 새로운 화자에 맞게 기존의 화자 독립 시스템의 모델을 적응시키는 모델 적응 방법[2] 등이 있다. 모델 적응 방법은 새로운 화자의 적응 데이터의 양에 따라 그 성능이 크게 달라진다. 만약 적응시켜야 하는 모델의 수가 많고, 적응 데이터의 수는 제한되어 있다면, 좋은 인식 성능을 보여줄 수 없다.

본 논문에서는 여러 가지 화자 적응 방법[2][6][7] 중 최근에 화자 적응에 많이 쓰이는 MLLR 적응 방법을 이용하여, 한국어 음성에 잘 적용시킬 수 있는 방법을 연구하였다. 한국어 음성은 영어와는 달리 초성, 중성과 종성으로 이루어져 있고 중성과 종성의 전이 부분의 변화가 영어보다 크다. 따라서, 본 논문에서는 한국어 음성에 맞게 모델을 재조정하고, 적응시켜 인식률을 향상시키는 화자 적응 방법을 연구하였다. 또한, 적응 데이터의 수가 항상 충분한 것은 아니기 때문에 적응 데이터 수를 순차적으로 증가시켜가며 기존의 모델을 갱신하여 온라인 적응을 할 수 있도록 하였다.

2. MLLR 화자 적응

2.1 MLLR 화자 적응

MLLR 화자 적응 방법은 초기의 화자 독립 CD-HMM 시스템의 모델 파라미터를 갱신하여 새로운 화자의 적응 데이터와의 화자 독립 시스템과의 유사도가 최대가 되게 하는 것이다.

가우시안 분포를 가지는 기존의 화자 독립 CD-HMM 시스템의 어떤 특정한 분포 s는 평균 벡터 μ_s와 분산 행렬 C_s로 나타낼 수 있다. 주어진 관측 벡터 o에 대한 확률 밀도 함수 $b_s(o)$를 다음 식에 나타내었다.

$$b_s(o) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{-1/2(o-\mu_s)' C_s^{-1}(o-\mu_s)} \quad (1)$$

여기서 n은 관측 벡터의 차원이다.

만약 평균 벡터만을 적응한다고 하면, 적응된 평균 $\hat{\mu}_s$는 다음과 같이 된다.

$$\hat{\mu}_s = W_s \xi_s \quad (2)$$

여기서 W_s는 적응 데이터의 유사도를 최대화하는 $n \times (n+1)$ 행렬이고, ξ_s는 기존의 화자 독립 모델의 평균이다. 따라서, 특정한 밀도 s에 대한 화자 적응 시스템의 관측 확률은 다음과 같다.

$$b_s(o) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{-1/2(o-\hat{\mu}_s)' C_s^{-1}(o-\hat{\mu}_s)} \quad (3)$$

2.2 MLLR 변환 행렬의 추정

적응 데이터 O가 T개의 프레임으로 구성되어 있다면, 관측 데이터의 총 유사도는 기존의 화자 독립 모델 λ에 대해 다음과 같이 나타낼 수 있다.

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

여기서 $F(O, \theta|\lambda)$는 주어진 모델에 대한 관측 음성의 유사도이다. 화자 독립 모델 λ와 재추정된 모델 $\overline{\lambda}$의 유사도를 최대화하기 위해 보조 함수 $Q(\lambda, \overline{\lambda})$를 쓰는 것이 편리하다. 보조 함수는 식 (5)와 같이 유사도 함수에 재추정된 유사도 함수의 로그값의 곱으로 나타낼 수 있다.

$$Q(\lambda, \overline{\lambda}) = \sum_{\theta \in \Theta} F(O, \theta|\lambda) \log(F(O, \theta|\overline{\lambda})) \quad (5)$$

원하는 파라미터인 변환 행렬 W_s가 측정되면, 확률 밀도 함수 b_s가 변하게 되므로, 유사도에 관한 식 (5)는 변하는 성분 $b_s(o)$에 대해서 나타낼 수 있다. 변하는 성분 $b_s(o)$에 의해 식 (5)를 수정하여 다음 식에 나타내었다.

$$Q(\lambda, \overline{\lambda}) = \text{constant} + \sum_{\theta' \in \Theta} \sum_{t=1}^{T} F(O, \theta'|\lambda) \log b_{\theta_t}(o_t) \quad (6)$$

시간 t에서 상태 s가 있을때, 후위 확률 밀도 함수 $\gamma_s(t)$는 다음 식과 같다.

$$\gamma_s(t) = \frac{1}{F(O|\lambda)} \sum_{\theta \in \Theta} F(O, \theta_t = s|\lambda) \quad (7)$$

식 (7)로 식 (6)을 다시 쓰면,

$$Q(\lambda, \overline{\lambda}) = \text{constant} + F(O|\lambda) \sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_j(t) \log b_j(o_t) \quad (8)$$

보조 함수를 최대화하기 위해서 식 (8)을 변환 행렬 W_s에 대해 미분하여 최적 변환 행렬을 구하면 변환 행렬의 일반식을 다음과 같이 구할 수 있다.

$$\sum_{t=1}^{T} \gamma_s(t) C_s^{-1} o_t \xi_s' = \sum_{t=1}^{T} \gamma_s(t) C_s^{-1} \overline{W_s} \xi_s \xi_s \quad (9)$$

2.3 변환 행렬 공유

적은 수의 적응 데이터로 모든 모델의 변환 행렬을 구하기 어려우므로, 비슷한 특징을 가지는 모델의 경우에 대해 같은 변환 행렬을 공유하는 방법을 사용한다. 각각의 모델의 유클리드 거리를 구한 후, 2진 분리 알고리듬을 사용하여 회귀 트리를 만든다 [8]. 회귀 트리는 변환 행렬의 기본이 되고, 데이터가 충분하지 못하면, 독립적인 변환 행렬이 되지 못한다. 2진 회귀 트리 구조를 그림 1에 나타내었다. 회귀 트리의 수가 4개라고 가정하면, 적응 데이터가 충분할 경우는 실선으로 나타내고, 적응 데이터가 충분하지 못할 경우에는 점선으로 나타내었다. 그림 1에서 보면, 적응 데이터가 충분한 경우는 2, 3, 4 노드인 경우이다. 따라서, 변환 행렬은 2, 3, 4 노드에 해당하는 행렬이 된다. 적응 데이터가 충분할 경우에는 그 모델에 해당하는 변환 행렬을 가지게 된다. 각각의 모델들은 이 세 가지 중에 속하게되어

변환 행렬을 공유하게 된다.

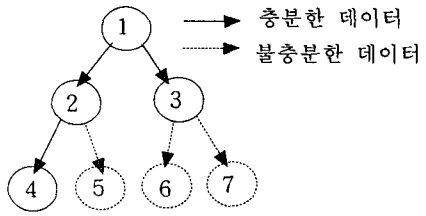

그림1. 2진 회귀 트리

2.4 한국어 음성 모델

한국어 음성의 구조는 초성, 중성, 종성으로 되어 있다. 종성은 나타나는 경우도 있고, 나타나지 않는 경우도 있다. 그리고 초성에 나오는 자음과 종성의 자음은 음소적으로는 같으나 음운학적인 특징이 다르다. 본 논문에서는 영어와는 다르게 음성의 모델을 초성은 영어와 같이 모델링하였고, 종성의 경우에는 앞의 중성에 더하여 모델링하였다. 예를 들면, '받침대'는 'ㅂ'+ 'ㅏ' + 'ㅊ'+'ㅁ' + 'ㄷ'+'ㅐ'로 모델링된다. 또한, 전이 부분에서의 변화를 잘 나타내기 위해 문맥 종속적으로 나타내었다. 위의 '받침대'의 경우는 'ㅂ-ㅏ+ㅊ', 'ㅏ-ㅊ+ㅁ', 'ㅊ-ㅁ+ㄷ', 'ㅁ-ㄷ+ㅐ'로 된다.

2.5 증가 적응

화자 독립의 경우에는 온라인 적응이 필요한 경우가 많다. 그러한 경우에 모든 적응 데이터를 처리한 후에 적응하는 것 보다 적응 데이터의 수를 점차 증가시키면서 적응시키는 방법이 온라인 상에서 사용하기에 더 편리하다. 본 논문에서는 적응 데이터를 순차적으로 증가 시켜 온라인 적응을 하였다.

3. 실험 결과

3.1 실험 환경

한국어 PBW445 음성으로 인식 실험을 하였다. 21명의 남성 화자와 19명의 여성 화자가 발음한 데이터에서 18명의 남성과 17명의 여성 발성을 훈련 데이터로 하였고, 나머지 화자 음성은 시험 데이터로 사용하였다. 적응 데이터는 시험 화자의 발음 중에서 임의로 추출하여 사용하였다. 445 음성 데이터는 16㎑ 표본화되어 있고, 16bit로 코딩되어 있다. 본 연구에서는 음성의 프레임 크기를 15㎳로, 5㎳씩 이동시켜 사용하였다.

프레임마다의 특징 벡터는 12 MFCC, 1개의 가지로 나타내었다. 인식률을 높이기 위해 87개의 모델을 문맥 의존적인 1541개의 triphone 모델로 만들어 실험을 하였다. 모든 훈련, 적응, 인식 실험은 HTK Ver. 2.2 (HMM Toolkit)[4] 시스템을 사용하였다.

3.2 실험 결과

화자 종속, 화자 독립, 화자 적응 인식률을 표 1에 나타내었다. 화자 종속 인식 실험 결과 93.22%의 인식률을 보였으나, 화자 독립의 경우에 85.03%를 보였다. 여전히 화자 독립 인식이 화자 종속의 경우보다 오인식률이 2배 정도 높았다. 적응 데이터의 수가 400개 일 때 제안한 방법으로 인식 실험한 결과를 표 1에 나타내었다. 화자 종속의 경우보다 낮은 단어 인식률을 보였으나, 화자 독립의 경우보다 7.02% 향상된 단어 인식률을 보였다.

표 1. 화자 종속, 화자 독립과 적응 후 단어 인식률

	화자 종속	화자 독립	적응 후
단어 인식률(%)	93.22	85.03	91.02

적응 데이터의 수에 따른 변환 행렬 수와 각각의 단어 인식률을 표 2에 나타내었다. 적응 데이터가 적을 때는 변환 행렬 수가 적어서 적응 후의 인식률이 화자 독립 인식률보다 낮은 경우도 있었으나, 적응 데이터의 수가 점차 증가함에 따라 인식 성능이 향상됨을 알 수 있다.

표 2. 적응 데이터 양에 따른 변환 행렬 수와 단어 인식률 비교

	적응 데이터 수				
	30	100	200	300	400
변환 행렬 수	3	7	15	25	41
단어 인식률(%)	84.13	86.15	88.67	90.15	91.02

적응 데이터의 수에 따른 단어 인식률의 변화를 그림 2에 나타내었다. 적응 데이터 수가 30개일 때의 인식률은 화자 독립 인식률 보다 1.1% 낮았으나, 적응 데이터 수가 점차 증가함에 따라 인식률이 향상된다는 것을 볼 수 있다.

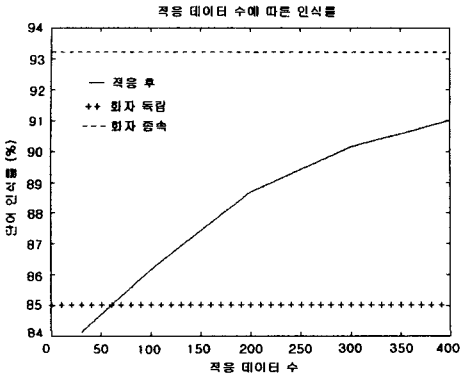

그림 2. 적응 데이터 수에 따른 인식률

4. 결론

본 논문에서는 한국어 화자 인식 시스템의 성능 향상을 위해 MLLR 적응 방법을 사용하였다. 기존의 화자 적응 방법으로 널리 쓰이는 MLLR 방법이 한국어 음성 인식에서도 잘 적용되는지 실험하였고, 문맥 의존 모델을 사용하여 인식 성능을 향상시키고, 증가 적응 방법을 사용하여 온라인 적응이 가능하게 하였다.

인식 실험 결과 기존의 85.03%의 화자 독립 인식률 보다 향상된 91.02%의 향상된 인식률을 얻었다. 그리고, 적응 데이터 수에 비례하여 인식 성능은 점차 향상되었다.

참고문헌

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* vol. 77, pp. 257-285, Feb. 1989.

[2] 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.

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.,* vol 41, pp. 164-171, 1970.

[4] S. J. Young, P. C. Woodland, and W. J. Byrne, *HTK-Hidden Markov Model Toolkit, Ver. 2.2,* Cambridge University Engineering Department and Entropic Research Lab. Inc, 1999.

[5] 이상호, 이건웅, 김광태, 홍재근, "왜곡계수 추정 함수를 이용한 화자 정규화," *제12회 신호처리 합동학술대회 논문집,* vol. 12, no. 1, pp. 839-842, 1999. 10.

[6] J. T. Chien and H. C. Wang, "Telephone speech recognition based on Bayesian adaptation of hidden Markov models," *Speech Communication 22,* pp. 369-384, 1997.

[7] I. Illina, M. Afify and Y. Gong, "Environment adaptation using mixture stochastic trajectory," *Speech Communication 26,* pp. 245-258, 1998.

[8] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proceedings of the ARPA Human Language Technology Workshop,* M. Kaufman, Princeton NJ, pp. 405-410, 1994.

[9] Y. Gotoh, M. M. Hochberg, and H. F. Silverman, "Efficient Training Algorithms for HMM's Using Incremental Estimation," *IEEE Trans. Speech Audio Processing,* vol. 5, pp. 539-548.