SUFFICIENT HMM 통계치에 기반한 UNSUPERVISED 화자 적응

고 봉 옥 , 김 종 교
전북대학교 전자정보공학부

Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics

Bong-Ok Ko , Chong-Kyo Kim
Div. of Electronics and Information Engineering, Chonbuk National University
E-mail : moakdosa@hanmail.net

Abstract

This paper describes an efficient method for unsupervised speaker adaptation. This method is based on selecting a subset of speakers who are acoustically close to a test speaker, and calculating adapted model parameters according to the previo usly stored sufficient HMM statistics of the selected speakers' data. In this method, only a few unsupervised test speaker's data are required for the adaptation. Also, by using the sufficient HMM statistics of the selected speakers' data, a quick adaptation can be done. Compared with a pre-clus tering method, the proposed method can obtain a more optimal speaker cluster because the clustering result is determined according to test speaker's data on-line. Experiment results show that the proposed method attains better improvement than MLLR from the speaker independent model. Moreover the propos ed method utilizes only one unsupervised sentence utterance, while MLLR usually utilizes more than ten supervised sentence utterances.

1. 서 론

현재까지 다양한 종류의 화자적응 방법들이 제안되어져 왔다. 화자종속 모델은 특정한 화자의 데이터 또는 특정한 화자와 밀접한 화자 데이터를 이용함으로써 훈련된다. 훈련을 위해 많은 양의 데이터를 이용하게 되면 acoustic model을 만드는데 많은 시간이 소요되는 단점이 있다. 따라서 이런 model adaptation 형식은 온라인 적응 모드에서 사용되어지기가 어렵다. 위의 문제를 해결하기 위해서 프리-클러스터링(pre-cluste ring)방법을 사용하였다[4]. 프리-클러스터링 방법은 적응모드 전에 여러 개의 화자종속 모델들을 준비하여 적응모델을 구한다. 이 방법은 어떤 종류의 화자 화자종속 모델을 준비할 것인가를 결정하는 것이 중요하다. 또, MLLR방법은 매우 인기 있는 방법이고 광범위하게 사용이 된다. MLLR은 화자 독립 모델에서 크게 향상된 인식률을 얻을 수 있다. 본 논문에서 제안하는 방법은 선택된 화자의 충분한 HMM 통계치를 이용하여 오직 몇 개의 unsupervised학습 화자의 데이터만을 가지고 화자적응을 하는 방법으로 빠른 적응을 할 수 있게 된다. 본 논문의 구성은 2장에서 기존의 화자적응 방법인 프리-클러스터링 방법과 MLLR방법에 대해

소개를 하고, 3장에서는 제안한 방법에 대해서 설명을 한다. 4장에서는 MLLR방법과 제안한 방법을 실험을 통하여 에러율을 비교하였다.

2. 화자적응 방법

2.1 프리-클러스터링

화자 프리-클러스터링은 화자들의 의해서 음향적 공간을 partitioning하는 것으로 보여 질 수 있다. 각각의 화자 클러스터에 대해서 하나의 음향시스템은 클러스터에 속한 화자들로부터의 음성데이터를 이용하여 훈련되어진다. 학습 화자의 데이터를 이용할 수 있을 때, 학습 화자와 각각의 클러스터와의 거리에 따라서 클러스터-종속 시스템의 등급을 나눈다. 음향적으로 학습 화자에 가까운 subset이 선택되어지고 각각의 선택된 클러스터들의 모델이 학습 화자의 음향공간에 가깝도록 가져오기 위해 변환되어진다. 이 방법은 훈련화자 모델에 대해서 과도한 저장공간 문제를 해결한다. 왜냐하면 클러스터들의 숫자가 훈련화자의 숫자보다 훨씬 적기 때문이다.

2.2 MLLR 적응방법

MLLR(maximum likelihood linear regression)적응방법은 MAP방법의 대안이 될 수 있는데 모델 파라미터의 선형변환 행렬을 추정하는 것이다[1][2][3]. 적응데이터에 의해 추정된 변환 행렬을 이용해서 관측되지 않는 모델에 대해서도 변환을 공유함으로써 매우 적은 양의 적응 데이터에 대해서도 빠른 적응이 가능하다. MLLR 에서 Gaussian의 평균 추정치는 다음과 같은 식에 의해서 update 된다.

$$\hat{u} = W_s \xi_s \qquad (2.1)$$

여기서, W_s 는 상태 s 에서 적응 데이터의 likelihood를최대로하는 n×(n+1)행렬이고, $\xi_s = [1, u_s, u_{s_1}, u_{s_n}]'$은 상태 s 에서의 확장된 평균 벡터이다. 각각의 상태가 단일 mixture Gaussian 분포를 가지고, 공분산 행렬은 대각 행렬이라고 가정하면, 현재의 모델을 λ, update된 모델을 $\hat{\lambda}$, 모든 가능한 상태열의 집합을 Θ일 때, 전체 likelihood 함수는

$$P(O \mid \lambda) = \sum_{\theta=0} P(O, \theta \mid \lambda) \qquad (2.2)$$

가 된다. 여기서 $P(O, \theta \mid \lambda)$는 주어진 모델 하에서 상태열 θ를 이용했을 때의 적응데이터의 likelihood함수이다. 식(2.2)를 최대로 하는 변환 행렬 W_s 를 찾기 위해서 latent variable가 추가된 compete data의 log-likelihood함수의 기대값으로 표시, 즉 함수는 식(2.3)와 같이 정의될 수 있다. 따라서, W_s는 이 함수를 최대로 하는 방향으로 변화 시켜가며 반복하는 EM(expectation maximization)알고리듬에 의해 구해질 수 있다.

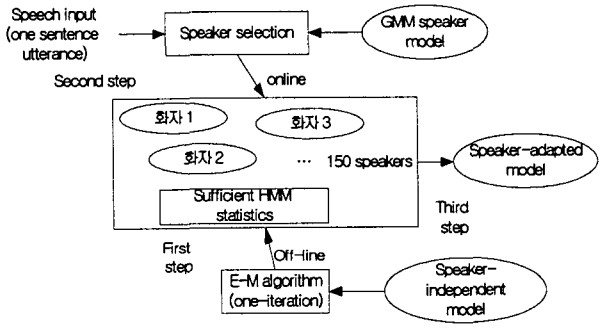
$$Q(\lambda, \bar{\lambda}) \quad \sum_{\theta \in \Theta} P(O,\theta \mid \lambda) \log\{ P(O,\theta \mid \bar{\lambda}) \} \qquad (2.3)$$

3. 제안한 방법

그림 3.1은 제안한 방법을 나타낸다. sufficient HMM 통계치에 기반한 unsupervised화자적응 과정은 세 단계로 구성되어 있다.

그림3.1 제안된 방법

3.1 첫 번째 단계 : sufficient HMM 통계치



sufficient HMM 통계치는 mean, variances 그리고 HMM모델들의 E-M count같은 음향 모델의 통계적 파마미터들이다. 이러한 파라미터들은 각각의 화자 에 대해서 개별적으로 계산되어 저장된다. sufficient HMM 통계치는 각각의 화자들의 데이터와 화자 독립 HMM 모델을 이용하여 E-M알고리즘을 한번 반복 수행함으로써 추정 할 수가 있다.

3.2 두 번째 단계 : 화자의 subset 선택

이 단계에서는 가우시안 mixture 화자모델을[5] 이용한 화자선택을 알아본다. 좋은 적응모델을 구하기 위해서는 테스트 화자와 음향적으로 가까이 있는 화자들의 Subset을 선택하는 것이 중요하다. 본 논문에서는 화자들의 집합을 선택하기 위하서 phone-independent one-state HMM인 32-가우시안 mixture

모델로 이루어진 화자 모델을 사용하였다[5]. 학습 화자의 데이터와 다른 화자의 데이터들 간의 거리에 따라서 적응 데이터에 가우시안 mixture모델인 acoustic likelihood가 사용되었다.. 화자들에 가장 가까운 Top N이 적응된 음향 모델을 계산하기 위한 화자들의 집합 subset으로 선택된다.

3.3 세 번째 단계 : 적응모델

학습 화자의 데이터가 입력이 되어질 때 3.2절 과정을 적용하여 훈련 화자 내에서 음향학적으로 비슷한 학습 화자들의 subset을 결정한다.

sufficient HMM 통계치에 대한 개념을 간단히 소개를 하면, 적응과정에서 음향 모델을 계산하는데 적은 시간이 소요된다. 왜냐하면 이런 한 값들은 오프라인에서 미리 적응되기 전에 계산이 되어진다. 이러한 방법은 데이터베이스를 이용하는 대신에, sufficient HMM 통계치를 적응 과정에 사용하여 적응을 하였다. 화자적응 음향모델은 통계적인 계산방법을 이용해서 선택된 화자들의 sufficient HMM 통계치로부터 계산되어진다. 이 과정은 화자 독립 모델로부터 HMM훈련을 한번 반복하는 것과 같다. 선택된 화자의 sufficient HMM 통계치로부터 화자 적응된 음향 모델은 다음과 같은 통계적 계산 방법으로 계산된다.

$$u_i^{adp} = \frac{\sum_{j=1}^{Nsel} C_{mix}^{j} u_i^{j}}{\sum_{j=1}^{Nsel} C_{mix}^{j}} \quad (i=1,\ldots,N_{mix}) \tag{3.1}$$

$$v_i^{adp} = \frac{\sum_{j=1}^{Nsel} C_{mix}^{j}(v_i^{j}+(u_i^{j})^2)}{\sum_{j=1}^{Nsel} C_{mix}^{j}} - (u_i^{adp})^2 \quad (i=1,\ldots,N_{mix}) \tag{3.2}$$

$$a^{adp}[i][j] = \frac{\sum_{k=1}^{N_{sel}} C_{state}^{k}[i][j]}{\sum_{j=1}^{N_{state}} \sum_{k=1}^{N_{sel}} C_{state}^{k}[i][j]} \tag{3.3}$$

여기서 u_i^{adp} v_i^{adp}는 각각의 적응된 모델과 선택된 화자에 대해서 mean, variance이다.

$a^{adp}[i][j](i,j=1,2,\ldots,N_{state})$는 상태 i,j의 천이확률이다. N_{mix}, N_{state}는 Gaussian과 state의 수를 표시한다. $C_{mix}^{j}(j=1,\ldots,N_{sel})$와 $C_{state}^{k}[i][j]$ $(i,j=1,2,\ldots,N_{sel}, i,j=1,2,\ldots,N_{state})$각각의 가우시안과 상태 천이의 E-M count이다.

4. 실험 방법 및 결과

본 논문의 실험을 위해서 사용한 음성 데이터베이스는 한국과학기술원 통신연구실에서 제작한 무역 상담용 DB로 조용한 환경에서 녹음된 대화형의 문장들을 낭독체 스타일로 발화한 연속음성으로 한 문장은 대략 10~20개의 단어로 구성되어 있고, 남성화자 100명과 여성화자 50명이 평균 98문장씩을 발성한 DB이다. 본 논문에서는 46개의 음소 set을 사용하여 phone-independent 3-state HMM 모델을 구성하여 가우시안 mixture 모델을 사용하여 MLLR과 비교 실험을 하였다.

4.1 MLLR과 비교

제안한 방법과 MLLR방법을 monophone 단위에서 가우시안 화자모델을 구성하여 실험을 하였다. 문장에서 16 가우시안과 32 가우시안을 구성하였다. 제안한 방법에서는 문장 하나로 unsupervised 적응실험을 하였고 MLLR방법에서는 supervised방법으로 10문장, 50문으로 실험을 하였다. 16 가우시안 일 때에 제안한 방법에서는 15.1%의 에러 율을 나타냈으며 MLLR의 방법은 15.9%, 13.1%의 에러 율을 나타내 10문장보다는 에러 율이 적었으나 50문장보다는 그렇지 못했다. 32 가우시안으로 구성하여 실험을 한 결과 비교 실험한 MLLR의 10, 50문장보다도 에러 율의 감소가 있었다. MLLR방법과 제안한 방법의 실험 결과를 표 4.1과 그림 4.1과, 그림 4.2에서 나타냈다.

표 4.1 MLLR방법 와 에러율 비교

		제안한 방법			MLLR		화자 독립 모델
		unsupervised			supervised		
문장 발성		1	2	10	10	50	
word error rate (%)	16 가우시안	15.1	14.7	13.9	15.9	13.1	19.7
	32 가우시안	11.8	11.7	11.4	13.8	12.1	15.8

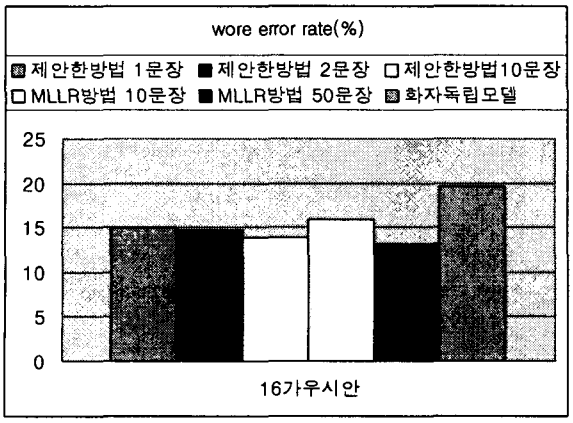


그림 4.1 16 Gaussian 에러율 비교

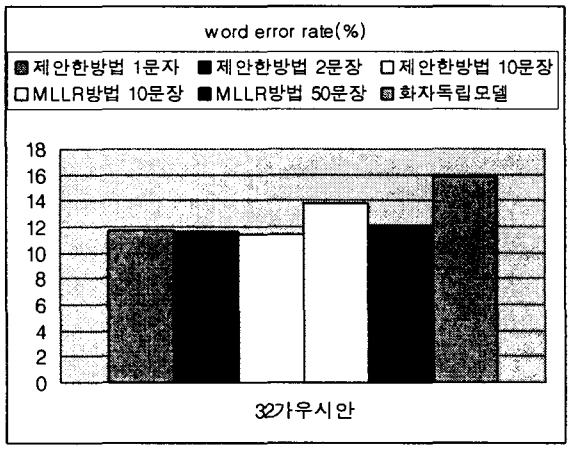


그림 4.2 32 Gaussian 에러율 비교

5. 결론

본 논문은 새로운 적응방법을 제안하였다. 이 방법은 적응하는데 몇 개의 unsupervised 테스트 화자의 데이터만 필요로 한다. 선택된 화자의 sufficient HMM 통계치를 사용하여 빠른 적응을 할 수가 있었다. 실험결과 제안된 방법이 화자적응의 대표적인 방법인 MLLR방법보다 16 가우시안과 32 가우시안에서 에러율이 감소됨을 알 수 있었고, 적은 테스트 화자 데이터로 효과적인 화자적응을 할 수 있었다. 따라서 온라인 상에서 빠르고, 에러율이 적은 화자인식에 적용 할 수 있을 것이다.

참 고 문 헌

[1] C. J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adapta tion of continuous density hidden Markov models ", *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, September 1995

[2] C. J. Leggetter, *Improved Acoustic Modeling for HMMs Using Linear Transformations*, Ph.D. thesis, Cambridge University, 1995

[3] Sam-joo Doh, *Enhancements to Transformation-Based Speaker Adaptation Principal Component and Inter-Class Maximum Likelihood Linear Regression*, Ph.D .thesis, Carnegie Mellon University, 2000

[4] Tuqing Gao, Mukund Padmanabhan and Michael Picheny, "Speaker adaptation based on pre-clustering training speakers" *Proceedings of the Eurospeech*, pp.2091-2094, 1999

[5] D. A Reynolds and R. C. Rose, "Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, jan. 1995.