화자 적응 기술을 이용한 한국어 화자 확인

최 동 진, 오 영 환
한국과학기술원 전자전산학과 전산학전공

Korean Speaker Verification Using Speaker Adaptation Methods

Dong-jin Choi and Yung-Hwan Oh
Voice Interface Laboratory, CS Division, KAIST
E-mail : {cdjin,yhoh}@speech.kaist.ac.kr

Abstract

Speaker verification systems can be implemented using speaker adaptation methods if the amount of speech available for each target speaker is too small to train the speaker model. This paper shows experimental results using well-known adaptation methods, namely Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR). Experimental results using Korean speech show that MLLR is more effective than MAP for short enrollment utterances.

I. 서론

화자 확인은 발성된 음성이 원하는 화자인지 아닌지를 구분해 내는 기술이다. 화자 확인 기술은 지정된 문장을 발성하게 하느냐에 따라 문장 종속형과 문장 독립형으로 나눠질 수 있다. 최근 문장 독립형 화자 확인 시스템은 GMM(Gaussian Mixture Model)에 기반하여 만들어지는 경우가 많다. 그리고 목적 화자와 사칭자간의 변별력을 높이기 위해서 UBM (Universal Background Model)이 사용되기도 한다[1]. 화자 확인 시에는 사용자가 목적 화자일 때의 우도와 사칭자일 때의 우도간의 차이를 임계치와 비교하여 결과를 낸다.

화자 모델을 훈련시키기 위해서는 많은 양의 목적 화자의 음성이 필요하다. 하지만, 현실적으로 목적 화자의 음성을 많이 요구할 수 없는 문제가 나타날 수 있다. 예를 들어 20초미만의 음성만으로 목적 화자의 모델을 훈련시켜야 할 경우 일반적인 방법으로는 목적 화자의 특성을 잘 나타내는 모델을 만들 수 없다.

이 논문에서는 화자 적응 시스템에 일반적으로 많이 사용되는 MAP[2]와 MLLR[3] 방법을 이용하여 UBM을 목적 화자에 적응 시켜 목적 화자의 모델을 구성하는 방법을 제시한다. UBM은 여러 화자가 발성한 음성으로 훈련될 수 있으므로 음성 자료의 양에 제약을 받지 않는다. 화자 적응 기술을 이용하면 목적 화자의 모델을 직접 훈련시킬 때 보다 적은 양의 자료만으로 목적 화자의 모델과 비슷한 성능을 나타내는 모델을 구성할 수 있게 된다.

논문의 구성은 다음과 같다. 2장에서 GMM을 사용하는 일반적인 화자 확인 방법을 설명하고, 3장에서 화자 적응 방법들에 대해 기술한다. 4장에서 화자 적응 기술은 화자 확인에 적용시켰을 때의 실험 결과를 보이고, 5장에서 결론을 맺는다.

II. GMM을 이용한 화자 확인

2.1 GMM

최근 대부분의 문장 독립형 화자 확인 시스템에서는 GMM을 사용한다. GMM을 사용하기 위해서는 음성의 각 프레임들이 서로 독립적이며, 각 프레임간의 특징벡

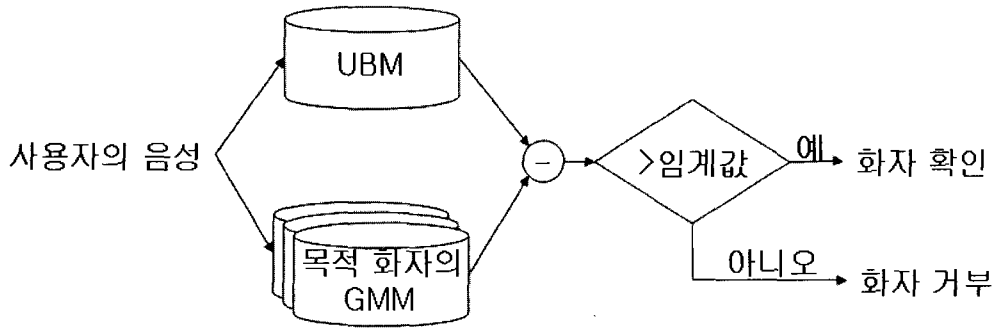

그림 1. 화자 확인 시스템의 구성도

터가 서로 상관되어 있지 않다고 가정한다. N개의 가우시안을 가진 GMM이 주어졌을 때, 입력 음성 $X=\{x_1, x_2, \ldots, x_T\}$에 대한 확률은 다음 식과 같이 계산될 수 있다.

$$p(X)=\prod_{t-1}^{T} p(x_t)=\prod_{t-1}^{T}\sum_{n=1}^{N} w_n \cdot N(x_t;\mu_n,\sigma_n) \quad (1)$$

이 때, w_n은 평균 μ_n과 표준편차 σ_n을 가지는 가우시안 $N(x_t;\mu_n,\sigma_n)$의 가중치이다.

GMM은 일반적으로 EM 알고리즘을 잉용하여 훈련된다. 이 알고리즘은 다음 식과 같이 주어진 데이터 X의 확률분포를 최대화하는 특징 Θ를 추정한다.

$$\hat{\theta}=\arg max_{\theta}\, p(X|\theta) \quad (2)$$

2.2 화자 확인 시스템의 구성

일반적인 화자 확인 시스템에서는 GMM-UBM 시스템을 사용한다. 이 시스템에서는 각 목적 화자 음성으로 훈련되는 GMM 이외에 UBM이라는 GMM을 하나 더 필요로 한다. UBM은 음성 특징의 화자 독립 분포를 표현하기 위해 훈련되는 하나의 큰 GMM이다. UBM은 일반적으로 여러 화자의 음성을 이용하여 훈련시키지만, 경우에 따라서 화자의 성별이나 사용 환경에 따라 여러개의 UBM을 구성할 수도 있다.

목적 화자의 GMM과 UBM이 만들어지면, 그림 1과 같은 화자 확인 시스템을 구성할 수 있다. 화자 확인을 하려는 음성이 입력되면, 이것을 UBM과 목적 화자의 GMM에 통과시키고 여기서 나오는 우도간의 차를 임계값과 비교하여 화자 확인 또는 거부를 하는 방법이다.

III. 화자 적응 기술

화자 적응 기술의 대표적인 방법으로는 MAP와 MLLR이 있다.

3.1 MAP를 이용한 화자 적응 방법

MAP 방법은 ML (Maximum Likelihood) 방법과는 달리 $p(X|\Theta)$의 Θ를 선험 분포 $p(\Theta)$를 가지는 랜덤 변수로 가정한다. MAP 방법은 다음과 같은 식으로 정리될 수 있다.

$$\begin{aligned}\hat{\theta} &= \arg max_{\theta}\, p(\theta|X) \\ &= \arg max_{\theta}\, p(X|\theta)\cdot p(\theta)\end{aligned} \quad (3)$$

GMM의 평균값만 다시 계산한다면, 적응 자료를 이용한 새로운 평균은 다음 식과 같다.

$$\hat{\mu}=\frac{\tau\mu+\sum_{t=1}^{T}\gamma(t)o(t)}{\tau+\sum_{t=1}^{T}\gamma(t)} \quad (4)$$

이 때, τ는 적응 자료와 이전 평균간의 편향을 나타내는 상수이고, $\gamma(t)$와 $o(t)$는 t번째 프레임에서의 확률과 적응 자료의 특징벡터를 각각 나타낸다.

3.2 MLLR을 이용한 화자 적응 방법

MLLR 방법은 기존의 평균들의 가중합을 이용하여 새로운 평균을 계산하는 방법이다. 이를 식으로 나타내면 다음과 같다.

$$\hat{\mu}=A\mu+b \quad (5)$$

이 때, 행렬 A와 벡터 b는 적응 자료의 우도를 최대화 하도록 하는 값이 된다.

적응 자료가 많을 때에는 회귀 트리를 이용하여 여러개의 행렬 A와 벡터 b를 만들어 사용할 수도 있다.

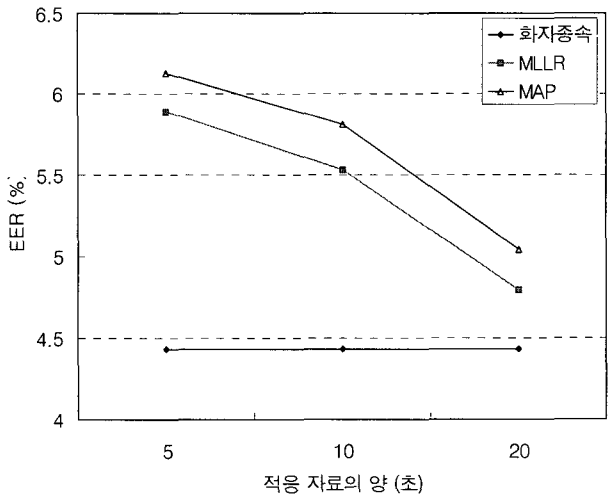

그림 2. 적응 자료의 양에 따른 EER의 변화

Ⅳ. 실험 및 결과

4.1 실험 환경

실험에 사용한 DB는 국립국어연구원에서 배포한 "서울말 낭독체 발화 말뭉치"이다. 여기에는 20대부터 60대까지 다양한 연령의 남성 화자 60명, 여성 화자 60명이 문장을 발성한 음성이 들어 있다. 80명의 화자는 930개의 문장을 발화하였고, 40명의 화자는 404개 문장을 발화하였다. 각 음성은 16bit, 16kHz로 샘플링 되었다.

이 중 60명 화자의 음성중 약 2시간 분량을 UBM을 훈련시키는데 사용하였고, 다른 60명 화자의 음성중 일부는 화자 적응에, 나머지 음성은 테스트에 사용하였다.

특징벡터는 에너지와 12차 MFCC 특징 벡터와 그 차분값을 사용하였고, 각 GMM의 가우시안의 수는 512개로 하였다. 인식시스템은 HTK[4]를 이용하여 구현하여 실험하였다.

4.2 실험 결과

그림 1의 시스템 구성도에서 임계치의 변화에 따라 목적 화자가 발성한 음성이 거부되는 비율인 false rejection rate와 사칭자가 목적화자로 인식되는 비율인 false alarm rate이 trade-off 관계에 있게 된다. 따라서 이러한 실험에서는 두 오류율이 같아졌을 때의 오류율을 나타내는 EER(Equal Error Rate)를 성능 평가 기준으로 사용할 수 있다.

그림 2는 학습 자료의 양의 변화에 따른 EER(Equal Error Rate)을 나타내는 그래프이다. 화자 등록시 사용되는 적응 자료의 양이 5초에서 20초로 다양하게 변할 때, MLLR을 이용하여 목적 화자 모델을 구성하는 것이 MAP를 이용하는 것보다 높은 성능을 나타냄을 알 수 있다.

이와 같은 실험 결과는 적응 자료의 양이 너무 작아 MAP를 이용한 경우 적응이 제대로 이루어 지지 않아 발생한 결과로 생각할 수 있다.

1시간 이상의 목적 화자 음성으로 훈련된 화자 종속 모델을 사용한 경우보다는 화자 적응을 이용했을 때의 성능이 조금 낮은 것을 확인할 수 있다. 하지만, 실제 환경에서는 목적 화자 음성을 많이 요구할 수 없는 경우가 있으므로 이러한 화자 종속 모델은 구성할 수 없을 수도 있는 것을 생각해야 한다.

Ⅴ. 결론

본 논문에서는 목적 화자를 훈련시킬 음성 자료의 양이 제한되어 있을 때, 화자 확인 시스템을 구성하기 위하여 화자 적응 기술을 이용하였다. UBM으로부터 MAP와 MLLR을 각각 이용하여 목적 화자 모델을 구성하였고, 한국어 연속 음성 자료를 이용하여 실험하였다. 실험 결과 충분한 양의 음성을 이용하여 훈련된 화자 종속 모델에는 못 미치지만, 비교적 좋은 성능의 화자 확인 시스템을 구성할 수 있었다.

앞으로, 한국어 특성에 맞는 특징을 이용하여 화자 확인 성능을 높이는 방법에 대한 연구를 계속할 것이다.

감사의 글

본 연구는 과학기술부의 지원을 받아 2006년도 국가지정연구실을 통해 수행되었음.

참고문헌

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 10-41, 2000

[2] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," IEEE

Trans. Signal Processing, vol. 39, pp. 806-814, 1991
[3] C. Leggetter and P.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] Steve Young, "The HTK Book," Cambridge University Engineering Department, 2001