MLLR 화자적응 기법을 이용한 적은 학습자료 환경의 화자식별

김세현, 오영환
한국과학기술원 전자전산학과 전산학전공 음성인터페이스연구실

Speaker Identification in Small Training Data Environment using MLLR Adaptation Method

Se hyun Kim, Yung-Hwan Oh
Voice Interface Lab. Div of Computer Science, Dept. of Electrical Engineering and Computer Science, KAIST

shkim@speech.kaist.ac.kr, yhoh@cs.kaist.ac.kr

Abstract

Speaker Identification is the process of automatically identify who is speaking on the basis of information obtained from speech waves. In training phase, each speaker models are trained using each speaker's speech data. GMMs (Gaussian Mixture Models), which have been successfully applied to speaker modeling in text-independent speaker identification, are not efficient in insufficient training data environment. This paper proposes speaker modeling method using MLLR (Maximum Likelihood Linear Regression) method which is used for speaker adaptation in speech recognition. We make SD-like model using MLLR adaptation method instead of speaker dependent model (SD). Proposed system outperforms the GMMs in small training data environment.

Ⅰ. 서론

물리적인 특징들의 차이에 기인하는 화자간 음성의 변이는 음성인식의 경우 이를 극복하도록 모델링 하는 것이 중요하지만, 화자인식에서는 이를 잘 반영할 수 있도록 모델링하는 것이 중요하다. 이처럼 음성인식의 경우에는 화자의 특성보다는 음운정보에 초점을 맞추게 되므로, 학습모델을 훈련시키기 위해 특정화자의 음성뿐 아니라, 임의의 화자들이 발성한 다량의 음성자료를 학습에 이용할 수 있다. 하지만 화자인식의 경우에는 각각의 화자모델을 학습시키기 위해서는 특정화자가 발성한 음성자료만을 이용하여 학습모델을 생성해야 함으로, 각각의 모델을 훈련시키기 위해 사용할 수 있는 자료의 양이 제한적일 수 밖에 없다.

화자인식의 한 분야인 화자식별 시스템은 입력 음성이 등록된 화자들 중 누구의 것인지를 판별하여 결과로 출력하는 시스템이다. 화자식별을 위해서는 학습단계에서 비교대상이 되는 화자들의 모델이 각 화자들의 음성을 이용하여 모두 학습되어 있어야 한다. 학습 방법으로는 일반적으로 GMMs, HMM 등의 통계적 모델링 기법을 사용하지만, 이들 방법은 학습 자료가 충분하지 않은 경우 좋은 성능을 보이지 못하고 있다. 본 논문에서는 화자적응에서 적은 양의 학습자료로 유사화자모델을 생성할 수 있는 MLLR 적응 기법을 이용하여 화자독립 모델로부터 유사 화자모델을 생성하고, 이를 화자식별에 이용하는 방법을 제안하고 비교 실험한다[1]. 논문의 구성은 다음과 같다. 2장에서 기존 화자식별 시스템의 구성 및 기존의 화자모델 생성 방법에 대해 설명한 후, 3장에서 MLLR 적응 기법에 대한 소개와, 제안하는

MLLR 기법을 적용한 화자모델링 방법에 대해서 설명한다. 4장에서는 비교 실험을 통해 제안하는 방법을 검증하고, 5장에서 결론을 맺는다.

Ⅱ. 화자식별 시스템과 화자 모델링

1. 화자식별 시스템

일반적인 화자식별의 과정은 그림 1.과 같다. 먼저, 입력음성에서 음성신호에 포함되어 있는 개인성 정보를 나타내는 특징 파라미터를 벡터열의 형태로 추출한다. 학습과정에서는 추출된 특징 벡터열을 이용하여 각각의 화자모델을 학습시킨다. 모델 학습 방법으로는 DTW (Dynamic Time Warping), 신경회로망 (neural network), 벡터양자화(vector quantization), GMMs, HMM 등을 사용할 수 있다. 최근에는 GMMs이나 HMM을 이용하는 화자식별 시스템이 대부분을 차지하고 있다. 화자 모델을 생성하고 나면 학습과정은 종료된다. 인식단계에서는 입력 음성이 들어오면, 이로부터 입력벡터열과 학습된 모델과의 유사도를 측정하여 인식점수를 구한다. 유사도 측정은 모든 화자 모델들과의 유사도를 모두 계산한 후, 그 중 최대값에 해당하는 화자 아이디를 결과로 출력하게 된다.

즉, 화자식별은 식 1.과 같이 $\lambda_1, \lambda_2, \cdots, \lambda_S$로 표현된 S명의 화자그룹에서 주어진 음성열 X의 사후확률 (A Posterior Probability)를 최대화시키는 화자모델 $\hat{S}$를 찾는 것이다..

$$\hat{S} = \underset{1 \le k \le S}{agrmax}\ P(X|\lambda_k) \quad (1)$$

2. 화자 특성의 표현

화자모델은 화자의 특징을 잘 나타내고 다른 화자들과의 차이를 효과적으로 표현할 수 있도록 모델링되어야 한다. 일반적으로 GMMs, HMM 등의 통계적인 모델링 방법들을 주로 사용하는데, 등록화자별로 각 화자의 음성을 이용하여 각각의 화자모델을 생성해야 함으로 화자별로 필요한 자료의 양이 어느 정도 이상이 되어야 모델을 효과적으로 생성할 수 있다.

화자모델은 식 1.의 λ_k를 생성하는 과정이다. 일반적으로 화자식별 시스템의 화자모델링 방법으로 GMMs이 가장 효과적인 것으로 알려져 있다[2]. GMMs에서는 각 화자의 음성파라미터의 분포를 가우시안 밀도의 가중 합으로 다음과 같이 표현된다.

$$P(\vec{x}|\lambda_i) = \sum_{k=1}^{N} w_k p_k(\vec{x}) \quad (2)$$

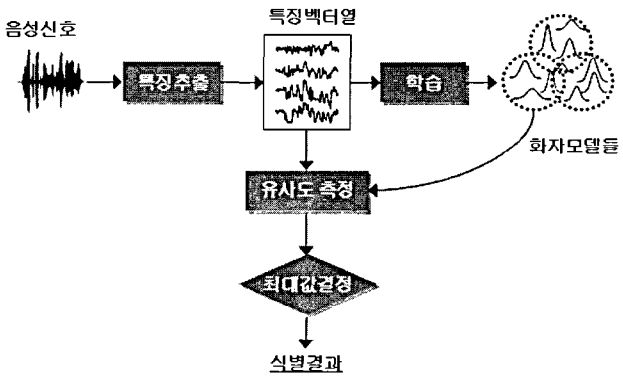

그림 1 화자식별 시스템의 구성도

$$p_k(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp - \frac{1}{2}(\vec{x} - \overline{\mu_k})^t \Sigma_k^{-1} (\vec{x} - \overline{\mu_i}) \quad (3)$$

위 식에서 각각의 화자모델 $\lambda_i = (w_i, \overline{\mu_i}, \Sigma_i)$는 각 가우시안들의 가중치, 평균, 분산으로 표현하게 된다. 표현의 정확성을 위해 화자별로 여러 개의 가우시안을 사용하게 는데, 이에 따라 정확한 모델링을 위해 학습과정에서 예측해야 되는 파라미터의 수도 많아지게 된다. 파라미터의 값들은 신뢰할 수 있는 수준으로 예측하기 위해서는 각 화자별로 충분한 양의 학습자료가 필요하게 된다. 그러나 충분한 양의 학습자료를 확보하는 것은 항상 가능한 것은 아니며, 상황에 따라 학습자료의 양이 제한될 수도 있으므로, 제한된 환경에 적합한 화자모델링 방법이 필요하게 된다. 학습자료의 양이 적은 경우의 접근방법은 크게 2가지로 나누어 볼 수 있다.

① 학습자료를 이용해 예측해야 되는 모델의 파라미터의 수를 줄인다.
② 화자독립모델을 이용하여 각 화자의 특성을 포함하고 있는 유사화자모델을 생성한다.

첫 번째 방법인 파라미터의 수를 줄이는 경우, 정해져 있는 특정 문장이나 단어를 발성하게 하여 화자식별을 행하는 문장종속형 화자식별 시스템에서는 어느 정도 성능을 기대할 수 있으나, 화자의 자유발성을 이용해 화자식별을 행해야하는 문장독립형 화자식별 시스템의 경우에는 그 성능이 현저히 떨어지게 된다.
따라서 본 논문에서는 화자독립 모델에 화자적응 방법을 적용하여 유사 화자모델을 생성하는 두 번째 방법을 사용하여 학습자료의 수가 적은 화자식별에 효과적인 화자 모델링을 제안한다.

Ⅲ. MLLR 기법을 적용한 화자 식별

화자간의 변이는 화자들 간의 서로 다른 특성을 나타내는 것으로 동일한 발성에 대해서도 서로 다른 특징을 보인다. 일반적으로 음성인식에서 특정 화자에 특화되어 학습된 화자종속모델의 경우, 훈련된 화자에 대해서는 더 좋은 인식 성능을 보이게 된다. 그러나 화자종속모델을 훈련시키기 위해서는 목적 화자의 발성으로 구성된 많은 양의 학습자료가 필요하게 된다. 이러한 문제점을 극복하기 위해서 많이 사용되는 방법이 화자적응기법이다. 화자적응은 화자독립모델로부터 목적화자의 적은 양의 적응 자료를 이용하여 유사화자종속 모델을 만들어, 목적화자에 대한 인식 시스템의 성능을 향상시키는 방법이다. 음성인식에서 사용하는 인식기법에는 MAP (Maximum A Posterior)에 기반하는 방법, MLLR과 같은 파라미터 변환에 기반한 방법 그리고 화자군집화에 기반한 방법들로 나누어 볼 수 있다.

1. MLLR 화자적응기법

본 논문에서는 화자모델의 생성을 위한 충분한 자료가 없는 환경을 가정하고 있다. 따라서 화자모델의 생성을 위해 기존의 GMMs, HMM등의 기법들을 적용할 수 없으므로, 음성인식에서 사용하는 화자적응 기법을 이용하여 유사화자모델을 생성한 후, 이를 화자식별에 이용하게 된다. 위의 3가지 화자적응 기법 중 MAP에 기반하는 방법의 경우 화자적응을 위한 적응 자료의 양이 상대적으로 많이 필요하며, 화자군집화 방법의 경우는 화자 적응에는 적은 양의 자료가 필요하지만, 학습단계에서 각 화자별로 모델을 생성할 수 있는 충분한 자료가 필요한 문제점이 있다.. 따라서 본 논문에서는 학습단계에서 화자 모델을 생성할 필요가 없고, 적은 양의 적응자료만으로 유사화자모델을 생성할 수 있는 MLLR 화자적응 기법을 이용하여 식별에 필요한 화자모델을 생성한다. MLLR 화자적응은 주어진 모델과 적응자료 사이의 불일치를 줄여줄 수 있는 변환들의 집합을 찾아내는 것이다.

이때 불일치는 화자간의 불일치나 환경의 불일치 등이 될 수 있다. 좀 더 구체적으로 GMMs, HMM의 경우 가우시안 혼합(Gaussian Mixture)들의 평균과 분산 파라미터를 적응 자료의 화자나 환경에 맞도록 선형변환(linear transformation)하는 화자적응 기법이다.

이를 식으로 표현하면 다음과 같다.

$$\hat{\mu} = A\mu + b \tag{4}$$

μ는 적응되기 전 원 모델의 파라미터이며 이를 변환행렬 A와 편향 항 b를 이용하여 적응 모델 파라미터를 얻게 된다. 변환행렬 A는 주어진 적응 자료의 양에 따라 전역 적응변환행렬 1개만을 사용하거나 유사한 특성을 가지는 가우시안들로 나누어 회귀 클래스(regression class)를 생성해 각 클래스 별로 서로 다른 변환행렬을 사용할 수도 있다. 변환행렬은 주어진 적응 자료를 이용해 구할 수 있다.

2. MLLR 기법을 이용한 화자모델 생성

화자식별을 위해 각 화자모델을 생성할 수 있는 충분한 양의 학습자료가 없으므로, 우선 주어진 모든 학습자료를 이용해 화자독립 모델 λ_{SI}를 생성한다.

S개의 화자 모델 $\lambda_i (i = 1,...,S)$는 모든 화자의 훈련자료를 이용해 생성된 화자독립모델인 λ_{SI}에 i번째 화자의 음성을 화자적응 자료로 이용하여 다음과 같이 MLLR 적응기법을 적용하여 얻는다.

$$\hat{\lambda}_i = A\lambda_{SI} + b = W\lambda_{SI} \tag{5}$$

이는 화자 i의 음성으로만 학습된 화자종속모델 λ_i와 유사한 특성을 지니게 되는 유사화자모델이 된다.

식별 단계에서는 식 6.과 같이 각 화자의 화자모델대신 λ_{SI}에 각 화자의 음성과 MLLR 적응기법을 이용해 구한 유사화자모델 $\hat{\lambda}_i$를 대상으로 유사도를 측정한 후, 확률값이 가장 큰 모델을 결과 값으로 출력해 주게 된다.

$$\hat{S} = \underset{1 \le k \le S}{\arg max}\, P(X|\hat{\lambda}_k) \tag{6}$$

Ⅳ. 실험 및 결과

1. 실험 환경

제안한 방법의 유효성을 검증하기 위해 문장독립형 화자식별 시스템에 대해 성능을 비교하였다. 20명의 남성 화자를 대상으로 실험하였으며, 학습에는 화자별로 30초~5분정도의 자유발화 음성을 이용하였다. 특징벡터는 에너지와 12차 MFCC 특징벡터들과 그 차분값과 2차 차분값을 사용하였고, 인식기는 HTK (HMM Tool KIT) 을 이용해 구현하였다[3].

GMMs의 경우, 학습자료의 양에 따라 12~64개 가우시안들을 사용하였고, HMM은 에르고딕 HMM으로 구현하고 각 상태별로 4~16개의 가우시안을 사용하여 학습모델을 얻었다. MLLR 행렬의 수도 학습자료의 양에 따라 1~8개까지 나누어 학습하였다.

식별실험에는 등록화자 20명이 발성한 2~3s 정도의 음

성을 화자별로 8번씩 실험하였다.

2. 실험 결과

표 1.은 학습자료의 양을 변화시킴에 따라 GMMs, HMM, MLLR Matrix 모델링 방법의 식별 성능을 보인 것이다. 표 1.에서 보이는 바와 같이 학습자료의 양이 적은 경우에는 제안한 MLLR Matrix에 기반한 방법의 성능이 더 우수함을 알 수 있다. 이는 학습자료 양의 부족으로 인해 GMMs, HMM의 경우 통계적으로 충분히 모델링 되지 않아서 나타나는 결과로, 학습자료의 양이 늘어남에 따라 인식성능의 차이의 폭이 줄어들고 충분한(5분 이상) 학습자료가 제공되는 경우 GMMs, HMM의 성능이 MLLR 적응 방법보다 더 좋은 결과를 보이게 됨을 볼 수 있다.

표 1 GMM, HMM 모델링 방법과 비교 실험결과

학습자료 양	GMMs	HMM	MLLR Matrix
30s	75.63%	71.25%	82.50%
60s	80.00%	75.00%	86.25%
120s	83.75%	80.63%	87.50%
300s	92.50%	90.63%	89.38%

학습자료의 양이 충분한 경우에는 기존 화자식별 시스템에 좋은 성능을 보이는 것으로 알려진 GMMs을 사용해 모델링하는 것이 효과적이나, 학습자료가 적은 경우에는 MLLR 화자적응 기법을 사용하는 것이 효과적이다. 따라서 학습자료의 양에 따라 화자모델링 방법을 선택적으로 사용하는 것이 성능향상에 도움이 된다.

V. 결론

본 논문에서는 제한된 양의 학습자료를 이용하는 화자식별 시스템의 성능향상을 위해 MLLR 화자적응 기법을 이용한 화자모델링 방법을 제안하였다. 화자독립 모델로부터 MLLR 화자적응 기법을 이용해 유사화자종속 모델을 생성하는 방법을 통해, GMMs, HMM 등의 기존 통계적인 모델링 방법들에 비해 학습자료의 양의 적은 경우 더 좋은 성능을 보임을 실험을 통해 알 수 있었다. 현 모델링 방법은 가우시안 모델로 구성된 화자 독립모델을 적응 기법을 통해 유사화자모델로 생성하는 방법이나, 향우 MLLR 행렬을 하나의 화자모델로 이용해서 화자식별을 행하는 방법에 대한 연구를 진행중이다.

참고문헌

[1] C. J. Leggetter, P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous densit hidden Markov models," *Computer speech and Language,* vol.9, pp.171-185, 1995

[2] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Transaction on Speech and Audio Processing, vol.3, No.1, pp.72-83, January 1995

[3] Steve Young, "The HTK Book (for HTK Version 3.1)," Cambridge University Engineering Department 2001.