화자독립 음성인식을 위한 GMM 기반 화자 정규화

신 옥 근†

요 약

화자독립 음성인식기의 화자 정규화를 위해 GMM(Gaussian mixture model)분포를 이용하는 방법에 대해 실험한다. 이 방법은 벡터 양자화기를 이용한 선행 연구를 개선한 것으로, 정규화된 학습용 특징벡터들의 확률분포를 최적의 클러스터의 수를 갖는 GMM분포로 모델링한 다음, 이 분포를 이용하여 시험용화자의 워핑계수를 추정한다. 이 연구의 목적은 기존의 ML을 이용한 방법의 단점을 개선하는 동시에 벡터 양자화기를 이용한 선행연구와'soft decision'이라 불리는 확률 분포를 이용한 방법의 성능을 비교하는데 있다. TIMIT 코퍼스를 대상으로 한 음소 인식 실험에서 클러스터의 수를 적절한 크기로 설정한 GMM분포를 이용함으로써 벡터 양자화기를 이용한 방법에 비해 약간 나은 인식률을 얻을 수 있었다.

키워드 : 음성인식, 화자정규화, VTLN, GMM

Speaker Normalization using Gaussian Mixture Model for Speaker Independent Speech Recognition

Ok Keun Shin†

ABSTRACT

For the purpose of speaker normalization in speaker independent speech recognition systems, experiments are conducted on a method based on Gaussian mixture model(GMM). The method, which is an improvement of the previous study based on vector quantizer, consists of modeling the probability distribution of canonical feature vectors by a GMM with an appropriate number of clusters, and of estimating the warp factor of a test speaker by making use of the obtained probabilistic model. The purpose of this study is twofold: improving the existing ML based methods, and comparing the performance of what is called 'soft decision' method with that of the previous study based on vector quantizer. The effectiveness of the proposed method is investigated by recognition experiments on the TIMIT corpus. The experimental results showed that a little improvement could be obtained by adjusting the number of clusters in GMM appropriately.

Key Words : Speech Recognition, Speaker Normalization, VTLN, GMM

1. 서 론

화자 사이의 성도의 길이가 달라서 생기는 음향학적 특성의 변이 때문에 화자독립 음성인식기의 성능은 화자종속 인식기의 성능에 미치지 못한다. 이 성능차이를 극복하기 위한 연구가 꾸준히 진행되고 있으며, 그 중 한 가지가 성도 정규화(vocal tract normalization; VTN), 또는 화자 정규화라 불리는 방법[1, 2, 3]이다. 이들 중 대표적인 것으로 Lee 등[3]과 Welling등[4]이 제안한 ML(maximum likelihood) 방법이 있는데 이들의 공통점은 정규화특징벡터로 만들어진 HMM 음성인식모델 $\hat{H}$을 만든 다음, (식 1)과 같이 표현되는 유사도(likelihood)를 최대화시키는 워핑계수 $\hat{\alpha}_i$를 화자 i의 계수로 추정한다는 것이다.

$$\hat{\alpha}_i = \arg\max_{\alpha} p(x_i^{\alpha} \mid \hat{H}, W_i) \quad (1)$$

여기서 x_i^{α}는 α로 워핑된 화자 i의 특징벡터, W_i는 x_i의 전사(transcription)이다. 이 식을 이용하여 워핑계수를 추정하기 위해서는 $\hat{H}$뿐 아니라, x_i의 전사 W_i에 대한 정보도 필요하다. 먼저 $\hat{H}$을 구하기 위해 이들은 워핑되지 않은 특징벡터로 인식 모델 H를 만들고, 이 모델에서 화자별로 최대의 유사도를 갖는 계수를 구한 다음 워핑된 특징벡터로 $\hat{H}$을 만들거나[4], 워핑되지 않은 벡터로 모델 H를 만든 다음, 만들어진 모델을 이용하여 화자별 계수를 추정하고, 다시 새로운 모델을 만드는 방법을 반복함으로써 구하였다[3]. 시험용 발화가 주어졌을 때 W_i를 구하기 위해 이들은 여분의 인식과정을 한번 더 거치는 방법을 이용한다. 즉, 첫 인식과정에서는 정규화되지 않은 인식기에서 발화의 전사를 추정한 다음, 이를 이용하여 워핑계수를 구하여 워핑하고, 두 번

† 정 회 원 : 한국해양대학교 IT공학부 부교수
논문접수 : 2004년 11월 11일, 심사완료 : 2005년 6월 24일

째 인식과정에서 최종적인 인식을 수행하였다. 이 ML에 기초한 방법은 인식기를 직접적으로 이용하여 워핑계수를 추정한다는 장점이 있으나, 인식과정에 계수추정을 위한 또 다른 인식과정이 필요하다는 단점이 있다. 한편 Lee 등[3]은 인식단계에서의 계산량을 줄이기 위해, 학습용 발화의 워핑되지 않은 벡터들을 계수별로 모아 각각의 GMM(Gaussian mixture model)분포를 구하였다. 시험용 발화가 주어지면 이 발화의 특징벡터를 구하여, 이미 구해 놓은 GMM들 중에서 최대의 유사도를 갖게 하는 모델을 찾음으로 워핑 계수를 추정하였다. 이 방법은 결국 학습용화자의 계수는 인식모델을 이용하여 구하는 한편, 시험용화자의 계수는 GMM분포를 이용하여 구하는 것이 되어 대칭성의 문제가 있을 수 있다.

한편, 본 연구의 선행 연구[5]에서는 이러한 방법들의 단점을 개선하기 위한 한 가지 방편으로 워핑계수 추정에 인식모델을 직접 이용하는 대신, 반복적인 벡터 양자화기의 학습을 통해 정규화된 양자화기를 구하고 이것을 이용하여 시험용 화자의 워핑계수를 추정하는 방법을 제안한 바 있다. 이 방법은 같은 코퍼스를 이용한 다른 연구와 비교하여 비슷한 성능을 가짐을 확인하였으나, 양자화기의 크기를 2의 지수 승으로 밖에 설정할 수 없는 k-means 알고리즘의 제약으로 인해 클러스터의 수를 최적으로 설정할 수 없는 문제점이 있었다.

본 연구에서는 클러스터의 수를 자유롭게 정할 수 있는 GMM과 최적의 클러스터의 수를 정하기 위한 MDL(minimum description length) 클러스터링을 이용하여 '정규화된 학습용 벡터의 GMM분포'(정규화 분포)를 만든 다음, 시험용화자의 계수는 이 분포에서 최대의 유사도를 가지게 하는 계수로 추정하는 방법을 제안한다. 이 방법의 효용성을 알아보기 위해 TIMIT 코퍼스와 HTK 툴킷을 이용하여 연속음성의 음소인식 실험을 수행하였으며, 실험결과, 제안한 방법이 벡터 양자화기를 이용한 방법에 비해 약간 나은 결과를 얻을 수 있었다.

본 논문의 구성은 다음과 같다. 제 2장에서 정규화 분포가 주어졌을 때 워핑계수를 추정하는 방법에 대해 설명한 다음, 제 3장에서 정규화 분포를 구하는 방법에 대해 기술한다. 제 4장에서는 제안한 방법을 이용한 실험방법과 결과를 기술하고 고찰한 다음, 마지막 제 5장의 결론으로 끝맺는다.

2. 주파수 워핑

이 장에서는 본 연구에서 사용한 워핑함수와 워핑함수 추정에 사용되는 특징벡터에 대해 먼저 설명한 다음, 주어진 정규화 분포를 이용하여 워핑계수를 추정하는 방법에 대해 기술한다.

2.1 워핑함수와 특징벡터

본 연구에서 사용한 워핑함수와 특징벡터는 선행연구[5]에서 이용한 것과 같다. 워핑 함수는 구간선형 함수로 샘플링 주파수가 16KHz인 스펙트럼에서 0~6KHz 사이는 선형계수를 적용하고 나머지 6KHz에서 Nyquist 주파수인 8KHz까지는 끝점이 Nyquist 주파수에서 만나도록 하였다. 0~6KHz 구간에 적용하는 선형 계수는 0.02의 간격을 갖는 0.77~1.22 사이의 값이다. 워핑계수의 적용은 Umesh 등[6]이 사용한 방법과 같이 주파수 영역의 스펙트럼을 직접 워핑하는 방법을 이용하였다.

워핑계수 추정을 위한 특징벡터는 인식모델의 학습과 시험을 위해 사용하는 특징벡터와는 달리 모음의 주요 포만트만 포함하기 위해 모음의 평균적인 F4(3.6KHz)까지의 파워 스펙트럼 성분을 취하여 구하였다. 따라서 이 특징벡터에 적용되는 워핑함수는 위에서 설명한 워핑함수의 선형 구간의 일부분인 상수 계수이다. 워핑계수 추정을 위한 특징벡터는 12차의 MFCC이며 발화의 파워 스펙트럼을 주어진 선형계수에 따라 워핑한 다음, 워핑된 주파수 영역에서 F4까지의 성분을 이용해 구한다.

2.2 워핑계수의 추정

화자별로 정규화된 특징벡터 $\hat{x}_i$ (i=1,..N)가 주어졌다고 가정하고 이들의 확률분포 $\hat{\lambda}$를 GMM으로 모델링하면 화자 i의 정규화된 특징벡터 $\hat{x}_i$의 유사도는 (식 2)와 같이 표현될 수 있다.

$$p(\hat{x}_i \mid \hat{\lambda}) = \sum_{k=1}^{K} \pi_k b_k(\hat{x}_i) \tag{2}$$

(식 2)에서 b_k는 다음 (식 3)과 같이 표현되는 컴포넌트 밀도함수(component density)이며 π_k는 GMM을 구성하는 컴포넌트 k의 가중치이다.

$$b_k(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{D/2}} \exp\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\} \tag{3}$$

이 식에서 D는 벡터 x의 차원, μ_k와 Σ_k는 각각 x의 평균벡터와 covariance 행렬이다. (식 2)의 모델 $\hat{\lambda}$는 정규화된 특징벡터 x_i들의 분포모델이므로 화자 i의 정규화된 특징벡터 $\hat{x}_i$는 정규화되지 않은 특징벡터 x_i와 비교하여 다음의 관계가 성립한다.

$$p(\hat{x}_i \mid \hat{\lambda}) > p(x_i \mid \hat{\lambda}), \ \ if\ x_i \neq \hat{x}_i \tag{4}$$

따라서 유한한 워핑계수의 집합 A={ $\alpha_l : l$ = 1,..,L}를 고려할 때 화자 i의 워핑계수 $\hat{\alpha}_i$는 (식 5)와 같이 추정할 수 있다.

$$\hat{\alpha}_i = \arg\max_{\alpha} p(x_i^{\alpha} \mid \hat{\lambda}) \tag{5}$$

여기서 x_i^{α}는 특징벡터 x_i를 계수 α로 워핑한 것이다.

3. 정규화된 특징벡터들의 GMM분포

이 절에서는 정규화된 특징벡터들의 GMM분포 $\hat{\lambda}$를 구하는 방법에 대하여 기술한다. 먼저, 임의의 샘플 데이터가 주어졌을 때 이들의 GMM분포는 EM (Expectation-Maximization) 알고리즘을 이용하여 구할 수 있다[7]. EM 알고리즘은 안정적 (stable)이며 비교적 간단하게 구현할 수 있다는 장점이 있으나, 이 알고리즘으로 구한 최적치는 국부 최적치(local optimum)이며, 알고리즘 구동시에 설정되는 초기 조건에 민감하게 의존한다는 단점이 있다[8]. 본 연구에서는 정규화된 특징벡터의 GMM분포를 구하기 위해 EM 알고리즘을 이용하되, 이러한 EM의 단점을 극복하기 위해 적절한 초기조건으로 설정한 EM알고리즘을 반복적으로 구동한 다음, 최적의 컴포넌트의 수를 결정하는 방법을 택한다.

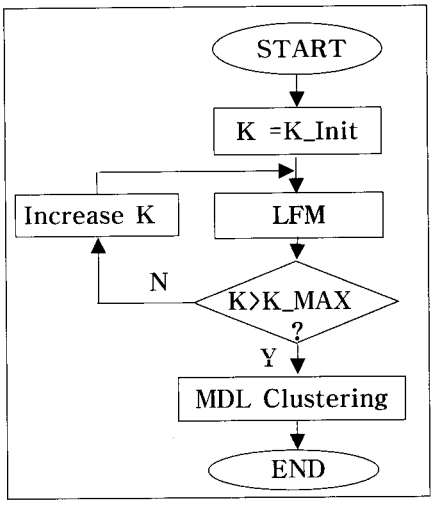
일반적으로 샘플 데이터의 GMM분포를 구한다는 것은 이 데이터의 GMM 파라미터 $(\pi_k, \mu_k, \Sigma_k; k=1,..,K)$를 구하는 것이다. 이들 중 컴포넌트의 수 K는 적절한 값으로 미리 주어질 수도 있지만, 경우에 따라서는 최적화 되어야 하는 대상이 될 수도 있다. 본 연구에서 파라미터 K는 궁극적으로는 최적화되어야 할 대상이지만, EM에 적절한 초기 조건을 제공하기 위한 제어 파라미터로도 사용된다.

또, 벡터의 분포 모델에서 컴포넌트의 수가 크면 샘플의 분포를 세밀하게 표현할 수 있으나, 너무 커지면 overfitting의 문제가 발생한다. 반면에 충분히 크지 않으면 분포의 대체적인 윤곽을 표현하는데 그친다.

따라서 화자 정규화의 관점에서 볼 때, 정규화되지 않은 벡터들의 분포에 대해서는 대체적인 윤곽을 얻고, 정규화가 진행될수록 컴포넌트의 수를 크게 하여 세밀한 분포를 얻는 것이 의미가 있을 것이다. 이러한 가정에 기초하여 본 연구에서는 다음과 같이 두가지 모듈을 이용하여 정규화 벡터의 GMM분포를 구한다.

먼저 Local Fitting Module (LFM)은 초기분포와 함께 컴포넌트의 수가 정해져서 구동되는데, 주어진 초기 분포를 다음과 같은 방법으로 반복적으로 학습시켜 국부적으로 최적의 분포를 구하는 역할을 한다. 즉, 각 화자의 발화를 워핑계수의 집합의 모든 원소로 워핑한 다음, 주어진 초기 분포모델과 비교하여 가장 높은 유사도를 갖는 계수를 찾는다. 이렇게 구한 계수들로 특징벡터들을 워핑하고 이들의 GMM분포를 구한 다음, 다시 이 분포에서 가장 높은 유사도를 갖는 계수를 찾는 과정을 반복한다. 이 과정을 반복하면 모든 화자들의 계수가 변하지 않는 수렴 상태에 도달하게 되는데 다음의 3.1절에서 LFM은 국부적인 최적치에 수렴한다는 것을 보인다.

또 다른 하나의 모듈은 Global Fitting Module(GFM)이라 부르는 것으로 (그림 1)의 순서도에 보인 것과 같이 컴포넌트의 수 K를 증가시켜 가면서 반복적으로 LFM를 구동한 다음, 마지막으로 MDL clustering을 수행하는 과정이다. GFM은 LFM을 구동할 때 초기 GMM 분포와 컴포넌트의 수 K를 제공함으로써 LFM을 제어하는데, K는 미리 정해진



(그림 1) GFM 순서도

방법대로 증가시키는 한편, LFM에 제공하는 초기 분포는 직전의 LFM 구동에서 구한 계수로 워핑한 특징벡터들의 분포이다. 따라서 본 연구에서 제안하는 방법은 컴포넌트의 수를 점진적으로 늘려 가면서 각 화자의 워핑계수를 더 정확한 값으로 추정해 나가는 방법으로 설명할 수 있다. 충분한 수의 컴포넌트를 갖는 분포 모델이 구해진 다음, 마지막으로 최적의 컴포넌트의 수 $\hat{K}$와 각 컴포넌트의 파라미터를 구하는 MDL clustering을 수행한다. 3.2절에서 이 모듈에 대해 자세히 설명한다.

3.1 Local Fitting Module

이 모듈은 주어진 수의 컴포넌트를 갖는 GMM 모델을 반복적으로 생성함으로써 국부적으로 최적인 화자별 워핑계수를 추정는 모듈이다. 이 절에서는 LFM이 국부적으로 최적인 상태에 수렴함을 보인다.

먼저, GFM에 의해 주어진 초기분포를 λ_0, 컴포넌트의 수를 K라 하고, 이 모듈의 j번째 iteration에서 생성된 GMM 모델을 λ_j^K, j번째 iteration에서 구한 화자 i의 워핑계수를 $\alpha(i, j)$라고 할 때, $\alpha(i, j+1)$은 다음 (식 6)과 같이 표현되며,

$$\alpha(i, j+1) = \arg\max_{\alpha \in A} p(x_i^{\alpha} \mid \lambda_j^K) \tag{6}$$

분포 λ_{j+1}^K는 (식 6)에서 구한 계수로 워핑한 특징 벡터들로 만들어진다. 초기에 주어진 분포모델에서 출발하여 위의 과정을 반복하면 j번째와 $(j+1)$번째에 구해진 특징 벡터들의 λ_j^K에 대한 유사도는 다음 (식 7)의 관계를 만족한다.

$$p(x_i^{\alpha(i,j+1)} \mid \lambda_j^K) \geq p(x_i^{\alpha(i,j)} \mid \lambda_j^K) \tag{7}$$

이 식에서 λ_j^K는 특징벡터 $x_i^{\alpha(i,j)}$들의 분포 모델이며, 좌측 항은 (식 6)에 의해 구해지는 계수로 워핑한 벡터 $x_i^{\alpha(i,j+1)}$

의 유사도 이다. 또 유사도 $p(x_i^{\alpha(i,j+1)} \mid \lambda_{j+1}^K)$ 와 $p(x_i^{\alpha(i,j+1)} \mid \lambda_j^K)$ 를 비교할 때 전자는 모델 λ_{j+1}^K 을 만드는데 사용된 벡터 $x_i^{\alpha(i,j+1)}$ 의 유사도 이므로 다음 (식 8)이 성립한다.

$$p(x_i^{\alpha(i,j+1)} \mid \lambda_{j+1}^K) \geq p(x_i^{\alpha(i,j+1)} \mid \lambda_j^K) \tag{8}$$

(식 7)과 (식 8)로부터 다음 (식 9)를 얻을 수 있다.

$$p(x_i^{\alpha(i,j+1)} \mid \lambda_{j+1}^K) \geq p(x_i^{\alpha(i,j)} \mid \lambda_j^K) \tag{9}$$

따라서 이 과정이 반복될수록 유사도는 단조증가하고, 결국에는 일정한 값에 수렴한다. 수렴상태에 도달했을 때의 워핑계수와 분포모델은 초기에 주어진 분포와 컴포넌트의 수에서 시작하여 구할 수 있는 국부적인 최적치이다.

3.2 Global Fitting Module

이 절에서는 컴포넌트의 수를 늘려가며 LFM을 구동하는 GFM의 역할에 대해 먼저 설명한 다음 MDL Clustering에 대해 설명한다. LFM에 최초로 제공되는 분포는 워핑되지 않은 특징벡터들의 분포이다. 만약 이 단계에서부터 컴포넌트의 수가 너무 크다면 어떤 음소들은 필요이상으로 많은 컴포넌트들로 모델링될 것이다. 이렇게 세분화된 분포 모델과 여러가지 계수로 워핑된 특징벡터들을 비교하여 최대의 유사도를 가지는 특징벡터를 찾는다면 특징벡터를 중 일부는 처음과는 다른 음소의 컴포넌트에서 최대의 유사도를 가지게 될 가능성이 높아진다. 반면에 초기의 컴포넌트의 수가 적당히 작다면 각 화자의 정확한 워핑계수 대신 대략적인 워핑의 방향밖에 구할 수 없겠지만 다른 음소의 컴포넌트에서 높은 유사도를 가지는 빈도는 출어 들 것이다.

두번째 이후의 LFM 구동에서 GFM은 직전의 LFM에서 구한 계수를 이용하여 초기화하기 때문에 컴포넌트의 수를 어느 정도 증가시켜도 잘못된 컴포넌트에서 최대의 유사도를 갖는 빈도는 크게 증가하지 않는 반면, 특징 벡터의 분포는 더 정교해진다. 따라서 컴포넌트의 수를 늘려가며 반복적으로 LFM을 구동함으로써 EM에 적절한 초기조건들을 제공할 수 있게 된다.

지금까지의 과정에서 컴포넌트의 수 K는 제어파라미터로 이용되었다. K가 커질수록 유사도는 높아지지만 필요 이상으로 커지면 overfitting현상이 발생하므로 최종적인 컴포넌트의 수 $\hat{K}$ 을 구할 필요가 있다. 이를 위해 컴포넌트의 수 K를 충분히 크게 한 다음 MDL clustering을 수행한다. MDL Clustering은 샘플 데이타로 부터 최적의 컴포넌트의 수 $\hat{K}$ 와 함께 $\hat{\lambda}$ 의 파라미터를 결정하는 방법이며, Rissanen[9]이 제안한 (식 10)의 criterion을 최소화시키는 K를 최적의 컴포넌트의 수 $\hat{K}$ 으로 간주한다. $\hat{K}$ 을 구하기 위해 먼저 주어진 컴포넌트의 수 K에서 시작하여 컴포넌트의 수를 하나씩 감소시켜가며 EM 알고리즘을 적용함으로써 GMM 파라미터와 MDL criterion을 구한다.

$$MDL(K,\lambda) = -\log(p_x(x \mid K,\lambda)) + \frac{1}{2} R \log(MD) \tag{10}$$

다음에 최소의 MDL criterion을 갖게 하는 컴포넌트의 수 $\hat{K}$ 와 그 때의 GMM 파라미터를 선택하고 이것을 샘플 데이타의 GMM분포로 삼는다. 이 식에서 M과 D는 각각 샘플의 수와 특징벡터의 차원이며 R은 $\hat{\lambda}$ 의 파라미터들 $(\pi_k, \mu_k, \Sigma_k; k = 1,..K)$ 을 표현하는데 필요한 데이타의 길이이며 다음 (식 11)과 같이 표현할 수 있다.

$$R = K(1 + D + \frac{(D+1)D}{2} - 1 \tag{11}$$

4. 실험 및 고찰

이 장에서는 제안한 방법의 성능을 알아보기 위해 수행한 인식실험과 결과를 기술한다. 본 연구에서는 벡터 양자화기를 이용한 화자 정규화의 결과와 비교하기 위해 선행연구와 같은 조건에서 실험하였다. 다음에 실험환경에 대해 요약한 다음, 실험결과를 보이고 고찰한다.

4.1 실험환경

HMM toolkit인 HTK[10]를 이용하여 음소 단위의 인식실험을 수행하였다. 이 toolkit을 음소별로 3개의 state를 갖는 CDHMM으로 설정하였다. 코퍼스는 TIMIT[11]을 이용하였으며 음소들의 빈도균형을 유지하기 위해 SA1과 SA2를 제외한 학습용 발화로 인식기를 학습시켰고 인식시험에는 시험용의 모든 발화를 이용하였다. 인식기의 학습과 시험에 사용할 특징 벡터는 각화자의 발화를 0.97의 계수를 갖는 프리엠퍼시스 필터로 필터링한 다음, 25ms 길이의 해밍 윈도우를 적용한 프레임을 초당 100개의 비율로 추출하였다. 이어서 각 프레임의 파워 스펙트럼을 구한 후, 화자별로 주어진 워핑함수로 주파수 영역에서 워핑하고 이들로부터 12차의 MFCC와 에너지를 먼저 구하였다. 마지막으로 이 MFCC와 에너지를 합친 13차원의 벡터, 그리고 이들의 1차 및 2차 미분을 포함시킨 39차원의 특징벡터를 만들어 인식기에 사용하였다.

한편 워핑계수를 추정하기 위해 사용한 음소는 13개의 모음 (iy, ih, eh, ey, ae, aa, ah, ao, ow, uh, uw, ux, er)이다. 성비 균형을 위해 TIMIT의 학습용 발화중 남여 각각 136명씩의 화자를 임의로 선정하여 SA1와 SA2를 제외한 모든 발화로부터 위의 13개의 모음음소를 추출하고, 인식기를 위한 특징벡터 추출시와 같은 방법으로 처리하여 파워 스펙트럼을 구하였다. 마지막으로 이들을 계수에 따라 워핑한 다음, F4에 해당하는 3.6KHz까지의 성분으로부터 12차원의 MFCC를 추출하여 GMM분포를 만들기 위한 특징벡터로 사

용하였다. GMM 분포가 만들어진 다음에는 화자별 특징벡터를 이 분포와 비교하여 최대의 유사도를 갖게 하는 계수를 찾아 화자별 워핑계수로 추정하였다.

4.2 인식 실험 및 고찰

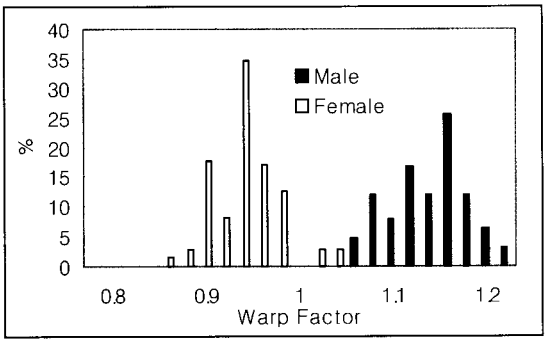
GMM의 클러스터의 수를 15에서 시작하여, 30, 45, 60, 90, 150, 220, 300, 400, 512의 순으로 증가 시켜가며 LFM를 구동하였다. GMM 클러스터의 수가 512가 되었을 때 MDL Clustering을 수행하였으며 그 결과 최적의 클러스의 수, $\hat{K}$ 은 355였다.

다음 표 1에 본고에서 제안한 방법을 이용하여 얻은 실험결과를 Baseline 인식률, 그리고 선행 연구의 벡터 양자화기를 이용하여 워핑하였을 때의 인식률[5]과 비교하여 보인다. Baseline 인식률은 본 연구와 같은 인식환경에서 화자 정규화를 거치지 않았을 때의 인식률이다. 또 선행연구에서 이용한 벡터 양자화기는 k-means 알고리즘을 이용한 것이며, 따라서 양자화기의 크기는 2의 지수 승으로 제한되었다.

〈표 1〉 인식실험 결과

	Cluster Size	Accuracy (%)	Improve-ment
Baseline	-	52.67	-
VQ	512	54.91	2.24
GMM	512	54.92	2.25
GMM	355	55.13	2.46

이 표에서 보는 것처럼 컴포넌트의 수, 혹은 클러스터의 수가 같을 때는 GMM 분포를 이용한 경우나 벡터 양자화기를 이용한 경우의 인식률 (Accuracy) 차이는 없었다. MDL Clustering을 통하여 적절한 컴포넌트의 수를 정한 GMM 분포를 이용해서 워핑한 경우, Baseline 보다는 2.46%, 벡터 양자화기와 비교해서는 0.22%의 절대 인식률 증가를 보였다. 이 결과는 벡터 양자화기를 이용한 클러스터링과 GMM 분포를 이용한 클러스터링의 차이는 없으나, 벡터 양자화기를 이용한 경우 클러스터의 수가 너무 커서 overfitting현상이 일어난 반면, GMM을 이용한 경우 클러스터의 수를 적절하게 조정할 수 있어 조금 더 나은 인식률을 얻을 수 있는 것으로 해석된다. 다음 (그림 2)에 GMM분포를 이용하여 워핑했을 때의 워핑계수의 분포를 보인다. 이 그림에서 남성화자의 발화는 주파수 영역에서 확장이, 그리고 여성화자의 경우에는 압축이 필요함을 알 수 있다.



(그림 2) 남여화자의 워핑계수 분포

5. 결 론

본 연구에서는 화자 정규화의 한 가지 방법으로 정규화된 학습용 특징벡터의 분포를 GMM으로 모델링한 다음, 이를 이용하여 시험용 화자의 워핑계수를 추정하는 방법을 제안하였다. 먼저 정규화 특징벡터의 GMM분포를 구하기 위해 워핑하지 않은 벡터들로 작은 크기의 GMM분포를 만든 다음, 점차 클러스터의 수를 늘려가면서 반복해서 GMM을 모델링하였으며, 이 반복과정들은 국부적인 최적치에 수렴함을 보였다. 충분한 크기의 GMM분포를 얻고 난 다음에는 최적의 클러스터의 수를 구하기 위해 MDL criterion을 최소화시키는 클러스터의 수를 구하고, 이를 이용하여 최종적인 GMM을 만들었다. 시험용화자의 워핑계수는 주어진 화자의 발화를 모든 가능한 계수를 적용하여 워핑한 다음 정규화 GMM에서 최대의 유사도를 갖는 계수로 추정하였다.

기존의 인식모델을 이용한 ML 방법들은 인식모델을 직접 이용한다는 장점이 있으나, 시험용화자의 워핑계수를 추정하기 위해 두 번의 인식 과정을 거치거나, 학습용 데이터의 계수추정은 인식기를 이용하는 반면 시험용 데이터의 계수추정에는 확률모델을 이용하는 비대칭성등의 단점이 있었다. 이에 비해 제안한 방법은 인식모델을 직접 이용하지는 않지만, 국부적인 최적 분포를 연속적으로 구하였으므로 양질의 워핑계수를 추정할 수 있으리라 기대된다. 제안하는 방법을 이용하여 Baseline 인식기보다 2.46%의 절대 인식률을 향상시킬 수 있었으며 선행연구에서 이용한 벡터양자화기보다는 약 0.2%의 인식률 향상을 얻을 수 있었다. 벡터양자화기에 비해 인식률이 약간 나은 것은 적절한 클러스터의 수를 구할 수 있었던 것에 기인하는 것으로 보인다. 전체적으로 볼 때, 대용량 어휘의 연속음성을 위한 화자정규화는 그간의 노력에 비해 아직 큰 효과를 얻지는 못하고 있으며 향후 비선형 워핑 함수, 화자 적응 등의 방법과 연계한 더 많은 연구가 필요하다.

참 고 문 헌

[1] P. Zhan, M. Westphal, "Speaker Normalization Based on Frequency Warping", Proc. ICASSP '97. pp.1039-1042, 1997.

[2] S. Molau, S. Kanthak, H. Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition," Proc. ESSV, pp.209-216, 2000.

[3] L. Lee and R. C. Rose, "A Frequency Warping Approach to Speaker Normalization", IEEE Trans. on Speech and Audio Processing, Vol.6, N0.1, pp.49-60. Jan., 1998.

[4] L. Welling, S. Kanthak, H. Ney, "Improved Methods for Vocal Tract Normalization", Proc. of ICASSP, pp.797-800, Mar., 1999.

[5] 신옥근, "연속음성 인식기를 위한 벡터양자화기 기반의 화자 정규화", 한국음향학회지, 제23권 제8호, pp.583-589, 2004.

[6] S. Umesh, L. Cohen and D. Nelson, "Frequency Warping and the Mel Scale", IEEE Signal Processing Letters, pp.104-107, Vol.9, No.3, March, 2001.

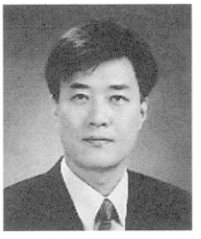
[7] E. Redner & H. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithms", SIAM Review, Vol.26, No.2, pp.195-239, Apr., 1984.

[8] G. J. McLachlan, T. Krishnan, "The EM Algorithm and Extentions", New York, Wiley, 1997.

[9] J. Rissanen, "A universal Prior for Integers and Estimation by Minimum Description Length", Annals of Statistics, Vol.11 No.2, pp.417-431, 1983.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, The HTK Book. ver.3., Microsoft Corp., 2000.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet and N. L. Dahlgren, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus: CDROM, NIST., 1993.



신 옥 근

e-mail : okshin@bada.hhu.ac.kr

1981년 서강대학교 전자공학과(학사)

1983년 부산대학교 전자공학과(공학석사)

1989년 프랑스 Université de Franche-Comté(공학박사)

1983년~1995년 한국전자통신연구소 선임연구원

1995년~현재 한국해양대학교 IT공학부 부교수

관심분야 : 신호처리, 음성신호처리, 음성인식