향상된 JA 방식을 이용한 다 모델 기반의 잡음음성인식에 대한 연구

A Study on the Noisy Speech Recognition Based on Multi-Model Structure Using an Improved Jacobian Adaptation

정 용 주*
Yong-Joo Chung

ABSTRACT

Various methods have been proposed to overcome the problem of speech recognition in the noisy conditions. Among them, the model compensation methods like the parallel model combination (PMC) and Jacobian adaptation (JA) have been found to perform efficiently. The JA is quite effective when we have hidden Markov models (HMMs) already trained in a similar condition as the target environment. In a previous work, we have proposed an improved method for the JA to make it more robust against the changing environments in recognition. In this paper, we further improved its performance by compensating the delta-mean vectors and covariance matrices of the HMM and investigated its feasibility in the multi-model structure for the noisy speech recognition. From the experimental results, we could find that the proposed improved the robustness of the JA and the multi-model approach could be a viable solution in the noisy speech recognition.

Keywords: Noisy speech recognition, JA, multi-model structure

1. 서 론

잡음음성인식을 위한 방법은 크게 3 가지 정도로 나누어진다고 보여 진다. 음질의 향상을 통한 방법, 잡음에 강인한 특징 추출방법 그리고 음성인식모델의 보상내지는 적응을 통한 방법들이다. 본 논문에서는 특히 인식모델의 보상방법에 대해서 논의하고자 한다. 인식모델 보상의 방법 중에는 PMC나 JA 방식 등이 주로 많은 관심이 되고 있는 기법이며 성능이 비교적 다른 방식에 비해서 우수한 것으로 알려져 있다[1][2]. JA 방식은 실제 환경과 비슷한 조건하에서 미리 HMM을 훈련한 경우에 매우 효과적인 것으로 알려져 있다. JA 방식에서는 미리 훈련된 기준HMM(reference HMM)의 파라미터 값을 Jacobian 행렬을 이용하여 실제 인식환경의 잡음음성에 적응시키게 된다. 우리는 JA 방식에서 향상된 기준HMM을 구성하기 위한 새로운 방안(Data-driven Jacobian

* 계명대학교 전자공학과

Adaptation: D-JA)에 대한 연구를 수행하였다. 여기서는 기준HMM을 구성하기 위해서 기존에 주로 사용하는 PMC나 NOVO와 같은 모델결합 방식 대신에 기준HMM을 잡음음성을 이용하여 직접적으로 훈련하는 방안이 제시되었다. 기존의 연구에서는 HMM의 정적 평균벡터에 대해서만 보상하는 방법이 제시되었으나 본 논문에서는 차분 특징벡터의 평균과 공분산 행렬에 대한 보상방식을 제안하여 보다 더 인식성능을 향상시키고자 하였다.

다양한 종류의 잡음이 존재하는 실제 인식환경에서는 잡음의 특성이 달라지는 경우에 대한 인식기의 강인성이 매우 중요하다 하겠다. 이러한 잡음의 다양성에 대비하기 위한 중요한 방법의 하나는 가능한 한 다양한 종류의 잡음음성을 준비하고 이를 이용하여 HMM을 미리 훈련함으로서 (Multi-condition training : MCT) 잡음음성에 대한 대처능력을 향상시키는 것이다[3]. 하지만, 이러한 방식은 인식시의 잡음종류를 미리 파악하기 힘든 어려움이 있고, 여러 가지 종류의 잡음을 이용하여 HMM을 구성할 경우 관측벡터에 대한 확률밀도함수가 지나치게 평탄화 되어 인식성능의 저하를 가져올 수 있다는 단점이 있다. 따라서 MCT방식의 단점을 보완하기 위해서 우리는 D-JA 방식을 이용한 다 모델 기반의 인식시스템을 고려하였다[4]. 먼저, 몇 가지 특정잡음의 음성에 대해서 잡음종류별로 기준HMM을 직접훈련방식으로 구성한다. 이후, 테스트(인식) 잡음음성에 대해서 HMM 파라미터들을 적응하기 위해서 D-JA방식에 근거한 모델 보상을 실시한다. 이러한 모델보상 방식이 성공적으로 진행되기 위해서는 미리 훈련된 기준HMM이 새로운 잡음음성에 충분히 적응 가능 하여야 된다. D-JA 방식은 기존의 JA 방식에 비해서 잡음의 변화에 대한 강인성이 뛰어나므로 우수한 적응 성능을 보일 것으로 생각된다. 이와 같은 다 모델(multi-model)기반의 인식방식은 다수의 훈련된 HMM 집합(set) 을 사용하여야 한다는 단점은 있지만, 단일 HMM만을 이용하는 기존의 모델보상방식에 비해서 우수한 성능을 보여주리라 기대된다.

본 논문의 구성은 2 장에서는 JA 방식을 간략히 소개하고 3 장에서는 향상된 D-JA 방식과 이에 기반 한 다 모델기반 잡음음성 인식에 대해서 소개하며 4 장에서는 제안된 방식에 의한 연구 결과를 소개하며 5 장에서 결론을 맺고자 한다.

2. JA 방식의 개요

켑스트럼(cepstrum) 영역에서 잡음음성 $\boldsymbol{n}$에 의해서 원래의 음성신호 $\boldsymbol{x}$는 다음과 같이 변환된다고 가정된다.

$$\boldsymbol{y} = \boldsymbol{C}\left[\log\{\exp(\boldsymbol{C}^{-1}\boldsymbol{x}) + \exp(\boldsymbol{C}^{-1}\boldsymbol{n})\}\right] \tag{1}$$

여기서 $\boldsymbol{y}$는 잡음음성신호이며 $\boldsymbol{C}$는 discrete cosine transformation(DCT)을 나타낸다[5].
잡음신호 $\boldsymbol{n}$에 대한 잡음음성신호의 변화율을 나타내는 Jacobian 행렬은 다음과 같이 표현된다 [2].

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}} = \boldsymbol{C}\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{C}^{-1} \tag{2}$$

여기서 R_y는 대각행렬이며 k번째 대각원소 $R_{y,k}$는 다음과 같다.

$$R_{y,k}=\frac{(\exp(C^{-1}\mu_n))_k}{(\exp(C^{-1}\mu_x))_k+(\exp(C^{-1}\mu_n))_k} \tag{3}$$

여기서 μ_n은 잡음신호의 평균값이고 μ_x는 연속밀도 HMM의 각 혼합성분별 평균벡터를 나타낸다. 위의 식 (2)와 (3)을 이용하면 주어진 잡음신호에 대해서 기준HMM의 각 혼합성분별로의 Jacobian 행렬을 구할 수 있게 된다. 이와 같이 얻어진 Jacobian 행렬을 이용하여 다음식과 같이 잡음신호가 n에서 $\tilde{n}$으로 변할 경우의 잡음음성신호 y의 변이를 나타낼 수 있다.

$$\tilde{y}=y+\frac{ROUND\,y}{ROUND\,n}(n-\tilde{n}) \tag{4}$$

한편 잡음음성신호에 대한 HMM의 각 혼합성분별 평균값을 얻기 위해서는 위식의 양변에 평균자를 취하여 다음과 같이 구한다.

$$E\{\tilde{y}\}=E\{y\}+\frac{ROUND\,y}{ROUND\,n}(E\{n\}-E\{\tilde{n}\}) \tag{5}$$

위식에서는 식(2)에서 구한 각 혼합성분별의 Jacobian 행렬을 이용하게 된다.
잡음음성의 차분벡터(delta-cepstrum) $\dot{y}$에 대해서도 다음식과 같이 Jacobian 행렬이 얻어진다.

$$\frac{\partial \dot{y}}{\partial n}=CR_{\dot{y}}C^{-1} \tag{6}$$

여기서 $R_{\dot{y}}$는 대각행렬이며 k번째 대각원소 $R_{\dot{y},k}$는 다음과 같이 얻어진다[2].

$$R_{\dot{y},k}=\frac{N_k\dot{X}_k-N_k\bar{X}_k}{(X_k+N_k)^2} \tag{7}$$

여기서 $\bar{X}_k=X_k(C^{-1}\mu_{\dot{x}})_k$ 이고 $X_k=(\exp(C^{-1}\mu_x))_k$ 이며, $\dot{x}$는 x에 해당하는 차분벡터이다. 식(6)을 이용하여 잡음신호에 의한 차분평균벡터의 변이는 다음과 같이 표현된다.

$$E\{\tilde{\dot{y}}\}=E\{\dot{y}\}+\frac{ROUND\,\dot{y}}{ROUND\,n}(E\{n\}-E\{\tilde{n}\}) \tag{8}$$

한편, HMM의 공분산 행렬에 대한 보상은 다소간의 통계적 가정을 기초로 하여 다음식과 같이 표현된다[2].

$$cov(\tilde{y})=cov(y)+\frac{ROUND\,y}{ROUND\,n}(cov(\tilde{n})-cov(n))\frac{ROUND\,y}{ROUND\,n}^T \tag{9}$$

$$cov(\tilde{\dot{y}})=cov(\dot{y})+\frac{ROUND\,\dot{y}}{ROUND\,n}(cov(\tilde{n})-cov(n))\frac{ROUND\,\dot{y}}{ROUND\,n}^T \tag{10}$$

3. D-JA 방식을 이용한 다 모델 기반 인식시스템

3.1 D-JA 방식의 개요

D-JA 방식에서는 기준HMM을 모델결합방식을 이용하여 얻는 대신에 잡음음성을 이용하여 직접 훈련하는 방식이 채택되었다. 이것은 일반적으로 모델결합방식으로 얻어진 HMM이 실제 환경의 잡음음성을 이용하여 직접 훈련된 HMM에 비해서 그 인식성능이 다소 떨어진다는 생각에 기반하고 있다. 비록 JA 방식이 기준HMM을 Jacobian 행렬을 이용하여 잡음음성에 적응시키는 것을 주요장점으로 하고 있지만, 기준HMM의 성능이 우수할 경우 보다 나은 적응 결과를 얻을 수 있으리라 생각된다. D-JA 방식에서는 Jacobian 행렬을 얻기 위해서는 Baum-Welch 알고리듬에 기반한 추정 방식을 사용하였다.

HMM에 기반한 음성인식에서 HMM 파라미터들은 보통 Baum-Welch 알고리듬에 의해서 얻어진다[6]. 연속밀도 HMM의 상태 j의 혼합성분 k에 해당하는 평균벡터는 다음과 같은 수식에 의해서 추정된다.

$$E(\boldsymbol{x}_t)=\frac{\sum_{t=1}^{T}\gamma_t(j,k)\,\boldsymbol{x}_t}{\sum_{t=1}^{T}\gamma_t(j,k)} \tag{11}$$

여기서 $\gamma_t(j,k)$는 켑스트럼 특징벡터 $\boldsymbol{x}_t$가 상태 j의 혼합성분 k에 의해서 발생될 확률을 의미한다. 식(4)와 (11)을 이용하면, 잡음음성신호에 대한 평균벡터는 다음과 같이 추정된다.

$$E(\widetilde{\boldsymbol{y}}_t)=\frac{\sum_{t=1}^{T}\gamma_t(j,k)\left(\boldsymbol{y}_t+\frac{ROUND\boldsymbol{y}_t}{ROUND\boldsymbol{n}_t}(\boldsymbol{n}_t-\widetilde{\boldsymbol{n}}_t)\right)}{\sum_{t=1}^{T}\gamma_t(j,k)} \tag{12}$$

만약, 잡음신호의 차이 $\Delta\boldsymbol{n}$ $(=\boldsymbol{n}_t-\widetilde{\boldsymbol{n}}_t)$의 값이 시간에 대한 평균치로서 대체가 가능하다면 위의 식은 다음과 같이 전개될 수 있다.

$$\begin{aligned}E(\widetilde{\boldsymbol{y}}_t)&=\frac{\sum_{t=1}^{T}\gamma_t(j,k)\,\boldsymbol{y}_t}{\sum_{t=1}^{T}\gamma_t(j,k)}+\frac{\sum_{t=1}^{T}\gamma_t(j,k)\frac{ROUND\boldsymbol{y}_t}{ROUND\boldsymbol{n}_t}}{\sum_{t=1}^{T}\gamma_t(j,k)}\Delta\boldsymbol{n}\\&=E(\boldsymbol{y}_t)+\frac{\sum_{t=1}^{T}\gamma_t(j,k)\left(\boldsymbol{C}\boldsymbol{R}_{y_t}\boldsymbol{C}^{-1}\right)}{\sum_{t=1}^{T}\gamma_t(j,k)}\end{aligned} \tag{13}$$

$$\mu_{\tilde{y}}=E(\boldsymbol{y}_t)+E(\boldsymbol{C}\boldsymbol{R}_{y_t}\boldsymbol{C}^{-1})\Delta\boldsymbol{n}=\mu_y+\mu_J\Delta\boldsymbol{n} \tag{14}$$

여기서 μ_y는 기준HMM의 평균벡터를 의미하고 μ_J는 추정된 Jacobian 행렬값이다. Δn은 훈련시의 기준잡음의 평균값과 인식시의 관측잡음 신호의 평균값의 차이를 나타낸다. 또한, D-JA 방식을 통해서는 정적벡터의 평균값 뿐만 아니라 공분산 행렬도 다음과 같이 추정이 가능함을 알 수 있다.

$$cov(\tilde{y}_t) = \{E(\tilde{y}_t - \mu_{\tilde{y}})(\tilde{y}_t - \mu_{\tilde{y}})^T\} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)((\tilde{y}_t - \mu_{\tilde{y}})(\tilde{y}_t - \mu_{\tilde{y}})}{\sum_{t=1}^{T} \gamma_t(j,k)} \quad (15)$$

위의 식에서 $\tilde{y}_t$는 식(4)을 적용하고 $\mu_{\tilde{y}}$에는 식(14)의 결과를 대입함으로서 공분산의 추정값을 얻을 수 있다. 한편, 본 논문에서는 차분벡터의 평균값에 대해서도 추정식을 유도하였다. 본 연구에서 사용된 잡음음성의 차분특징벡터는 다음과 같이 유도된다.

$$\dot{y}_t = \mu \sum_{k=-K}^{k=K} k \; y_{t+k}$$

위식의 각 항의 y_{t+k}에 식(4)를 적용하고 양변에 평균자를 취하면, 차분벡터 $\tilde{y}_t$의 평균값은 다음과 같다.

$$E\{\tilde{y}_t\} = \mu \sum_{k=-K}^{K} k \; E(y_{t+k}) + \mu \sum_{k=-K}^{k=K} k E\left(\frac{ROUND\; y_{t+k}}{ROUND\; n_t}\right)\Delta n \quad (16)$$

위식의 $E\{y_{t+k}\}$와 $E\{\frac{ROUND\; y_{t+k}}{ROUND\; n_t}\}$ 등은 식(12), (13) 등에서 정적벡터의 평균값을 추정할 때와 유사하게 얻어진다.

3.2 다 모델 기반의 인식시스템

D-JA 방식에서는 기본적으로 식(4)에 의한 선형식에 기반하여 기준HMM의 모델파라미터 값이 충분히 배경잡음에 잘 적응된다는 가정을 한다. 하지만, 기준HMM을 훈련할 때의 잡음과 인식시의 잡음간의 주파수 특성의 차이가 크다면 식(4)의 정확도는 떨어질 것이며 이로 인해 인식성능의 저하가 발생하리라 생각된다. 따라서 단일 기준HMM을 이용한 D-JA 방식에서는 잡음에 대한 적응 정도에 한계가 있으리라 생각된다. 다 모델 기반의 음성인식에서는 인식시에 가정되는 다양한 잡음 환경에 대해서 미리 다수의 기준HMM을 구성하도록 하며, 인식시에는 이러한 여러 가지 기준 HMM중에서 테스트 잡음음성에 가장 잘 맞는 것을 선택하여 인식모델로 삼는다. 이런 과정을 거침으로서 단일 기준HMM을 사용한 경우에 비해서 보다 다양한 잡음에 적응할 수 있는 강인성을 높일 수 있으리라 생각된다.

인식시의 잡음음성에 가장 적합한 인식모델을 찾기 위해서는 잡음음성에 대한 신호 대 잡음비(SNR: Signal to noise ratio)를 추정하고 잡음음성에 포함된 잡음신호의 분류를 하는 것이 필요하다. <그림 1>에는 다 모델 기반의 인식시스템에 대한 흐름도가 나타나 있다.

훈련과정에서는 먼저 다양한 잡음종류별로 별도의 기준HMM을 구성하게 된다. 이때, 같은 종류

의 잡음에 대해서도 몇 가지 SNR 레벨에 따라서 각각 기준HMM을 구성하게 된다. 기존의 연구결과에 따르면 필요한 SNR 레벨은 잡음종류별로 2~3 개가 적당하리라 생각된다. 인식과정에서는 잡음신호의 분류를 하고 잡음음성의 SNR 값을 추정한 후 그에 가장 적합한 기준HMM을 선택하게 된다. 선택된 기준HMM은 D-JA방식을 통해서 HMM 파라미터값이 보상되어 최종인식결과를 얻게 된다.

4. 인식실험 결과

4.1 기반 인식시스템의 개요

잡음환경에서 화자독립 단어 인식실험을 통해서 제안된 방식의 성능을 평가하였다. 인식대상 어휘는 음소분포가 비교적 고르게 되어 있는 한국어 75 단어이며 음향모델을 위한 기본단위는 32 개의 유사음소를 사용하였다. 각각의 유사음소단위는 연속밀도 HMM에 의해서 모델링된다. 화자의 수는 80 명이며 이들은 각각 75 단어를 한번 씩 발성하였다. 인식실험을 위해서 잭-나이프(Jack-knife) 방식을 이용하였다. 전체 화자를 20 명씩 4 개의 그룹으로 나눈 후, 그 중 하나의 그룹은 인식용으로 나머지 3 그룹은 훈련용으로 활용하였다. 이와 같은 과정을 4 회 반복하여 인식실험을 수행하여 인식화자의 수를 4 배로 증가시키는 효과를 거두도록 하였다. 잡음음성을 얻기 위해서는 원래의 깨끗한 음성에 차량(car)잡음, 배블(babble)잡음, 전시회(exhibition)잡음 그리고 지하철(subway)잡음을 다양한 신호대잡음비에 맞추어 더해 주었다. 잡음신호는 AURORA 2 데이터에 있는 잡음파일로부터 얻었다[7]. 인식특징벡터로는 13 차의 멜주파수 (mel-frequency) 켑스트럼 계수(MFCC)와 그의 차분계수(delta-MFCC)를 사용하였다[5].

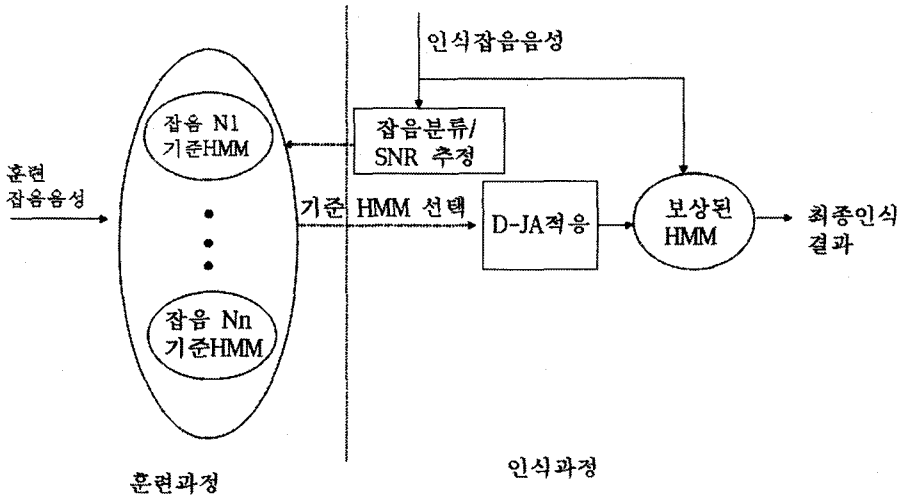

그림 1. D-JA 방식을 이용한 다 모델 기반 음성인식기의 구조

4.2 인식실험 결과

<표 1>과 <표 2>에는 차량잡음과 배블잡음 환경에서의 D-JA 방식과 기존의 JA 방식 및 PMC 방식의 인식성능비교가 나타나 있다. 아무런 모델보상을 취하지 않은 베이스라인 인식기의 경우에는 약 10 dB부터 인식성능의 저하가 심하게 나타남을 알 수 있다. 모델보상방식에 따른 효과를 자세히 살펴보기 위해서 HMM 파라미터들을 순차적으로 보상한 결과를 표에 나타내었다. 먼저, 정적 MFCC의 평균만을 보상해준 경우를 나타내고, 그 다음으로 차분벡터의 평균을 함께 보상한 경우와 마지막으로 공분산행렬까지 보상해준 경우들을 각각 나타내었다. 전반적으로 보상해주는 파라미터 수를 증가시킬수록 인식율은 향상되는 것을 알 수 있었으나, SNR값이 높은 경우에는 그 차이가 별로 크지 않은 것으로 보이며, 따라서 계산량이 중요한 실시간 인식을 위해서는 정적특징벡터 평균의 보상만으로도 어는 정도 원하는 결과를 얻을 수 있다는 것을 알 수 있다.

<표 1>에서 자동차잡음의 경우에는 JA 방식이 PMC 에 비해서 다소 나은 인식결과를 보임을 알 수 있다. 이것은 JA 방식이 Jacobian 행렬에 의한 보상을 통해서 PMC 방식에서 발생 가능한 HMM 파라미터 보상 오차를 다소 나마 보완해 주는 역할을 해 주기 때문이라 생각된다[8]. 하지만. 배블(웅성거림)잡음의 경우에서는 두 가지 방식사이에는 큰 차이가 없는 것으로 나타나는데, 아마도 배블잡음의 불규칙(non-stationary) 특성으로 인하여 Jacobian 적응을 위한 잡음값 추정이 다소 신뢰성이 떨어진 것이 그 원인인 것으로 보인다.

표 1 차량잡음과 배블잡음환경 음성인식에서 D-JA 방식과 기존의 모델보상간의 단어인식율(%)의 비교

		차량잡음				배블잡음			
		0dB	10dB	20dB	clean	0dB	10dB	20dB	clean
베이스라인 인식기		12.6	60.7	92.5	98.6	13.6	61.7	92.5	98.6
재훈련 방식		82.1	95.0	97.5	98.6	79.7	93.9	96.8	98.6
PMC	정적평균	59.8	87.8	95.3	98.6	54.9	86.6	95.1	98.6
	+차분평균	62.7	88.3	95.4	98.3	60.5	86.8	94.8	98.3
	+공분산	66.6	88.6	95.4	98.3	62.8	87.6	95.0	98.3
JA	정적평균	59.9	87.8	95.4	98.4	54.2	86.7	95.3	98.6
	+차분평균	62.4	88.2	95.5	98.4	59.3	86.8	94.8	98.4
	+공분산	68.6	89.5	95.8	98.4	61.5	87.1	95.3	98.4
D-JA	정적평균	82.4	95.0	97.4	98.6	79.7	94.0	96.9	98.6
	+차분평균	82.0	95.0	97.4	98.6	79.8	94.0	96.9	98.6
	+공분산	82.0	94.8	97.5	98.6	79.6	93.9	96.9	98.6

위의 결과에서 우리는 D-JA 방식이 JA와 PMC에 비해서 월등히 나은 인식성능을 보임을 알 수 있다. 이는 기준HMM을 잡음음성을 이용하여 직접 훈련하여 얻은 결과라고 생각되어진다. 상대적 인식에러감소율(%)은 기존 방식에 비해서 약 40~50(%) 정도이다. 위와 같이 우수한 성능을 보이기 위해서, D-JA 방식에서는 각 잡음의 SNR별로 독립적인 기준HMM을 구성하여야 한다. 하지만, 실제 인식시에 많은 수의 기준HMM을 가지고 있는 것이 쉽지 않으므로 우리는 D-JA 방식의 SNR의 변화에 따른 강인성을 조사하였다.

표 2. 자동차잡음환경에서 인식시의 SNR값이 변하는 경우의 D-JA 방식과 기존방식간의 단어인식율의 비교

	인식 / 훈련	0dB	5dB	15dB	25dB
D-JA	10dB	80.4	91.2	95.7	93.8
	20dB	71.4	87.4	96.7	97.8
JA	10dB	66.2	81.3	92.8	91.3
	20dB	67.7	81.9	93.3	96.7
MCT		71.1	86.3	94.3	96.0
PMC		62.8	78.3	92.2	97.0

표 3. 배블잡음환경에서 인식시의 SNR값이 변하는 경우의 D-JA 방식과 기존방식의 단어인식율의 비교

	인식 / 훈련	0dB	5dB	15dB	25dB
D-JA	10dB	77.5	89.9	95.5	94.5
	20dB	66.3	84.9	95.2	97.6
JA	10dB	59.8	77.2	90.8	89.9
	20dB	62.4	79.2	92.4	96.5
MCT		68.2	85.1	94.2	96.2
PMC		62.8	78.3	92.2	97.0

<표 2>와 <표 3>에는 기준HMM이 10 dB 와 20 dB에서 훈련된 경우, 자동차잡음환경과 배블잡음환경에서 테스트음성의 SNR이 변하는 경우에 D-JA 방식과 기존의 모델보상방식들의 인식성능의 변화를 나타내었다. MCT(Multi-condition training) 경우의 인식율도 함께 나타내었는데 이 훈련방식은 AURORA DB에 관한 인식실험에서 소개된 훈련방식으로 여러 가지의 잡음종류와 다양한 SNR을 고려한 잡음음성 데이터 set을 훈련시에 이용하여 HMM 모델을 구하는 방식이다. 따라서, MCT 훈련을 통해서 다양한 음향조건이 HMM파라미터에 반영되도록 할 수 있다. 본 연구에서는 MCT방식에 의한 훈련을 위해서 차량잡음과 배블잡음외에도 전시회잡음과 지하철잡음을 고려하였다. 표에 나타난 인식율은 평균벡터와 공분산 행렬을 함께 보상한 경우이다.

위의 표에서 보면 D-JA 방식은 기존의 방식에 비해서 잡음음성의 SNR 변화에 상당히 강인함을 보임을 알 수 있다. 예를 들어, 배블잡음환경의 경우에 기준HMM이 10 dB에서 훈련된 경우에 D-JA 방식은 테스트음성의 SNR값이 0 dB와 25 dB인 경우 각각 77.5(%)와 94.5(%)의 인식율을 보이지만, 기존의 JA 방식은 동일조건하에서 각각 59.8(%)와 89.9(%)의 인식율을 보임을 알 수 있었다. D-JA 방식은 전반적으로 MCT방식과 PMC 방식에 비해서도 우수한 성능을 보임을 알 수 있다.

이러한 결과로부터 우리는 특정 SNR 값에서 훈련된 기준HMM이 상당히 넓은 범위의 테스트음성 SNR 값들에서 우수한 인식성능을 보임을 알 수 있었다. 따라서 테스트음성의 SNR의 추정시 다소간의 오차가 있더라도, 위와 같이 2 개의 SNR 값(10dB, 20dB)에서 훈련된 기준HMM을 인식시에 이용할 수 있다면, 기존의 방식에 비해서 매우 우수한 인식성능을 보일 수 있으리라 생각된다. 이와 같은 다 모델기반의 음성인식시스템은 잡음음성인식환경에서 매우 유용한 방안이라 생각된다.

<표 2>, <표 3>의 경우에서 우리는 미리 잡음의 종류를 안다고 가정하였는데, 그렇지 못한 경우에는 테스트음성에 포함된 잡음의 종류를 미리 파악할 필요가 있을 것이다. 우리는 이를 위해서 GMM(Gaussian mixture model) 에 기반한 잡음모델을 가정하고 4 가지 종류의 잡음신호에 대해서 훈련과정을 통하여 평균과 공분산을 추정하였는데, 이때 혼합성분의 개수는 5로 하였다. 그리고 테스트 음성에 포함된 잡음의 종류를 분류하는 실험을 하였다. 표 4에는 그 결과가 나타나 있다. 잡음신호의 분류를 위해서는 테스트 음성의 묵음구간인 처음 20 프레임을 이용하였으며, 테스트음성의 SNR 값이 유사하다는 가정을 두었다.

표 4. GMM 방식에 근거한 잡음신호의 분류 결과

결과 / 테스트잡음	자동차(%)	웅성거림(%)	전시회(%)	지하철(%)
자동차	99.2	0.8	0	0
웅성거림	0.4	99.6	0	0
전시회	0.1	0	99.9	0
지하철	0	0	0.3	99.7

위의 표에서 보듯이 잡음분류는 상당히 정확한 인식성능을 보임을 알 수 있었다. 따라서 잡음의 종류를 어느 정도 미리 파악할 수 있는 환경에서는 이와 같은 잡음분류기법을 인식 전 단계에 두어서 테스트잡음음성에 가장 적합한 기준HMM을 선택할 수 있으리라 생각된다.

5. 결 론

본 논문에서는 D-JA 방식을 이용한 다 모델기반의 잡음음성인식시스템에 대하여 고찰하였다. D-JA 방식은 기존의 인식모델 보상방식에 비해서 테스트음성의 변이에 대해서 보다 강인한 특성을 가지며, 이를 이용하면 비교적 적은 수의 HMM set을 이용하여 잡음음성인식에서 우수한 성능을 얻을 수 있을 것으로 생각된다. 다 모델 기반의 잡음음성인식시스템이 보다 유용하기 위해서는 미리 가정되지 않은 잡음에 대해서도 기준HMM이 충분히 적응되어야 하는 문제가 있는데, 이를 위해서는 보다 향상된 잡음분류방법이나 모델보상절차에 대한 깊이 있는 연구가 향후 필요하리라 생각된다.

참 고 문 헌

[1] Gales, M. J. F. 1995. Model based techniques for robust-speech recognition, Ph. D. Dissertation, University of Cambridge.

[2] Sagayama, S., Yamaguchi, Y. & Takahashi, S. 1997. "Jacobian adaptation of noisy speech models", *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.

396-403.

[3] Yapanel, U., Hanse, J. H. L., Sarikaya, R. & Pellom, B. 2001. "Robust digit recognition in noise: An evaluation using the AURORA Corpus", *Eurospeech 2001, Aalborg, Denmark.*

[4] Xu, H., Tan, Z., Dalsgaard, P. & Linderg, B. 2005. "Robust speech recognition based on noise and SNR classification- a multiple-model framework", *Interspeech 2005, Lisbon, Portugal.*

[5] Davis S. B. & Mermelstein P. 1980. "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp.357-366.*

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

[7] ETSI draft standard doc. speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, *ETSI ES 202 108 V1.1.2 (2000-04), Apr. 2000.*

[8] Hung, J-W., Shen, J-L. & Lee, L-S. 2001. "New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques", *IEEE Trans. Speech and Audio Processing, vol. 9, no. 8, pp. 842-855.*

[9] 정용주, 2004. "직접데이터 기반의 모델적응방식을 이용한 잡음음성인식에 관한 연구", *음성과학 제11권 제2호*, p.111-119.

접수일자: 2006. 5. 1
게재결정: 2006. 5. 24

▲ 정용주
대구광역시 달서구 신당동 1000 (우: 704-701)
계명대학교 전자공학과
Tel: +82-53-580-5925
E-mail: yjjung@kmu.ac.kr