다 모델 방식과 모델보상을 통한 잡음환경 음성인식*

정용주(계명대), 곽성우(계명대)

<차 례>

<Abstract>

A Multi-Model Based Noisy Speech Recognition Using the Model Compensation Method

Young-Joo Chung, Seung-Woo Kwak

The speech recognizer in general operates in noisy acoustical environments. Many research works have been done to cope with the acoustical variations. Among them, the multiple-HMM model approach seems to be quite effective compared with the conventional methods. In this paper, we consider a multiple-model approach combined with the model compensation method and investigate the necessary number of the HMM model sets through noisy speech recognition experiments. By using the data-driven Jacobian adaptation for the model compensation, the multiple-model approach with only a few model sets for each noise type could achieve comparable results with the re-training method.

* Keywords: Multiple HMM model sets, Jacobian adaptation, Noisy speech recognition.

* 본 연구는 산업자원부 지방기술혁신사업(RT-104-01-01)지원으로 수행되었음.

1. 서 론

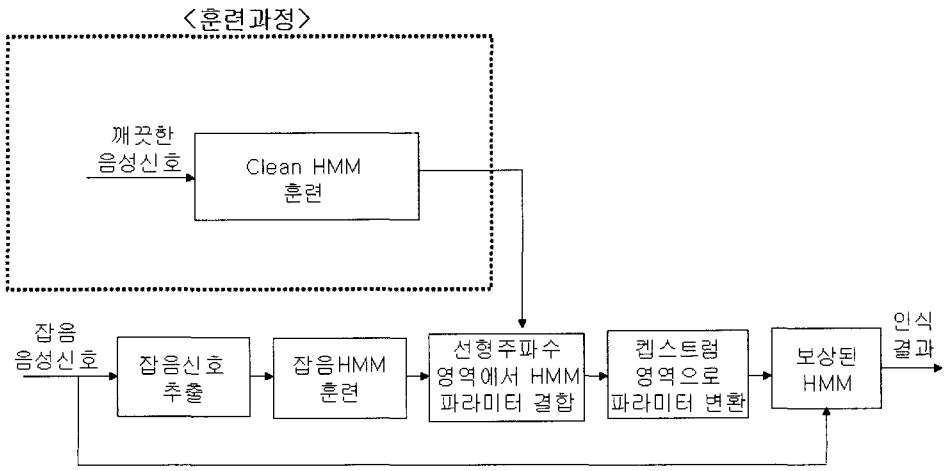
음성인식시스템이 작동하는 환경은 일반적으로 다양한 주변의 잡음신호의 영향을 받게 된다. 훈련환경과 이러한 실제 환경의 차이로 인하여 음성인식시스템은 때때로 원하는 만큼의 성능을 보여주지 못하는 경우가 많다. 이러한 문제점을 극복하기 위해서 음성인식시스템의 강인성을 위한 다양한 연구들이 이루어져 왔다[1]-[3]. 특히, 잡음음성의 음질을 개선한다든가 잡음에 강인한 특징을 인식 시에 사용하는 방법 등이 많이 사용되어 왔으며 최근에는 인식 모델의 보상을 통한 방법이 매우 효과적임이 알려져 있다.

위와 같은 여러 가지 음성인식시스템의 강인성을 위한 방법에도 불구하고 다양한 잡음조건과 환경에 대한 충분한 대처가 되지 못하고 있는 것이 현실이며 실제 환경에서의 음성인식시스템의 성능은 기대에 많이 못 미치고 있다. 최근에는 기존의 방식과는 조금 다른 관점에서 음성인식시스템의 성능을 향상시키고자 하는 노력이 있었는데, 여기에서는 기존의 인식시스템과는 달리 다수의 hidden Markov model (HMM) 모델이 인식 시에 사용된다[4]. 잡음종류별 그리고 신호대잡음비별로 다수의 HMM 모델들을 미리 훈련과정을 통하여 구성하여 둠으로써 인식시스템의 주변 환경이 변하더라도 그에 가장 알맞은 HMM 모델을 찾아서 인식성능의 최대화를 꾀할 수 있게 된다.

이와 같은 다 모델(multi-model) 기반의 음성인식시스템에서는 미리 구성되는 HMM 모델들의 개수를 정하는 것과 입력 잡음음성신호에 가장 적합한 HMM 모델을 찾는 것이 중요한 문제가 된다. 또한, 무조건 많은 수의 HMM 모델을 사용하는 것은 바람직하지는 않으므로 입력음성과 선택된 HMM 모델 사이에는 음향적인 차이가 존재할 수밖에 없으므로 이를 극복하는 방법이 관건이 된다.

본 연구에서는 다 모델 기반의 음성인식시스템에서 필요로 하는 최소한의 모델 개수를 정하기 위해서 다양한 인식실험을 수행하여 이를 실험적으로 추정하였다. 이를 위해서, 입력 잡음음성신호로부터 가장 적합한 HMM 모델을 선정하기 위한 Gaussian mixture model(GMM) 기반의 잡음신호 분류 방식을 사용하였고 voice activity detection(VAD) 기반의 신호대잡음비(SNR: Signal-to-noise ratio) 추정 방식을 이용하였다. 또한, 입력음성과 선택된 HMM 모델간의 차이를 줄이기 위해서는 데이터 기반의 Jacobian adaptation(JA) 방식을 이용하였다. 이러한 다양한 실험을 통해서 우리는 제안된 구조의 다 모델기반의 음성인식시스템이 기존에 사용되던 multi-condition training(MCT) 방식이나 재훈련(re-training) 방식 그리고 최근에 제안된 다양한 모델보상 방식에 비해서도 효과적임을 보일 것이다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 다 모델 기반의 음성인식시스템에서 핵심이 되는 모델보상방식에 대해서 간략히 소개하고자 한다. 3장에서는 모델보상방식과 결합된 다 모델 기반의 잡음음성인식시스템에 대해서 소개하며 4



<그림 1> PMC 방식의 개요

장에서는 인식실험결과를 나타내고 비교 검토하며 마지막으로 5장에서 결론을 맺는다.

2. 잡음음성인식을 위한 모델보상방식

2.1 Parallel Model Combination

잡음음성인식을 위한 모델보상 방식에서 가장 대표적인 parallel model combination(PMC)에 대한 개요가 <그림 1>에 나타나 있다. 여기에서는 깨끗한 음성신호와 잡음신호에 대해서 각각 켑스트럼(cepstrum) 영역의 HMM 파라미터들을 추정한다. 이를 위해서는 깨끗한 음성신호 외에도 잡음음성신호로부터 추출한 잡음신호를 필요로 하며, 일반적으로 잡음신호는 잡음음성신호로부터 VAD 알고리듬을 통해서 잡음구간을 추출하여 얻어지거나 입력 잡음음성의 맨 앞부분의 몇 개의 프레임을 잡음구간으로 간주하여 얻게 된다. 추정된 잡음신호와 원래의 깨끗한 음성신호의 HMM 파라미터 값들은 켑스트럼 영역에서부터 선형주파수 영역으로 변환되며 이를 위해서는 역 discrete cosine transformation(DCT) 변환이 사용된다. 선형 주파수영역에서 잡음신호와 깨끗한 음성신호의 HMM 파라미터 값들은 결합되어 잡음음성신호를 나타내게 된다. 이와 같이 결합된 잡음음성신호의 선형주파수 영역 HMM 파라미터 값들은 실제 음성인식시에 사용되기 위해서 다시 켑스트럼 영역으로 변환된다.

2.2 Jacobian Adaptation

PMC 방식에서는 원래의 깨끗한 음성의 HMM 파라미터값을 추출된 잡음신호의 정보를 이용하여 주어진 인식환경에 적합하게 적응시킨다. 그러나 이 방법은 실시간에서 동작하기 위해서는 다소 많은 계산량이 소요되는 문제점이 있으며 만족할 만한 인식성능을 보이기 위해서는 비교적 많은 양의 잡음신호 샘플이 있어야 한다. JA에서는 인식환경과 훈련환경의 차이가 그리 많지 않다는 가정 하에서, 비교적 구현이 간단하며 적은 수의 잡음샘플만을 이용하여 훈련환경의 HMM을 인식환경에 적합하도록 적응시키는 방법이 제시되고 있다.

켑스트럼 영역에서 부가잡음 신호 n에 의해서 원래의 깨끗한 음성신호 x는 다음과 같이 변환된다고 가정된다.

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

여기서 y는 잡음음성신호이며 C는 DCT를 나타낸다.

잡음신호 n에 대한 잡음음성신호의 변화율을 나타내는 Jacobian 행렬은 다음과 같이 표현된다.

$$\frac{\partial y}{\partial n} = CR_y C^{-1} \tag{2}$$

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

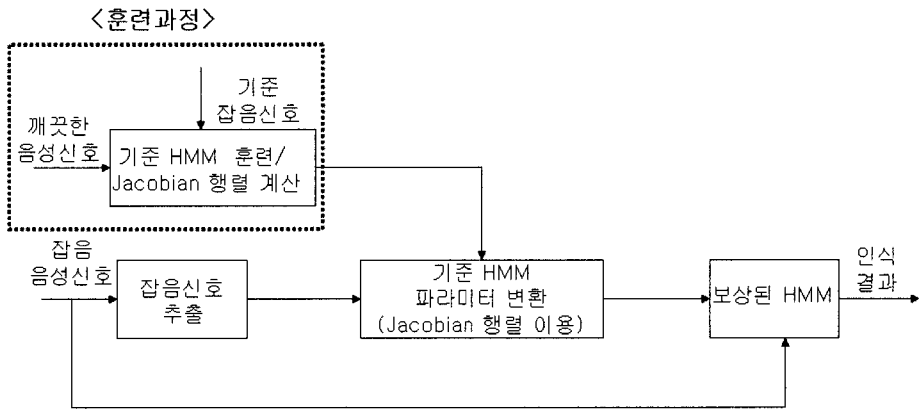
$$R_{y,k} = \frac{(\exp(C^{-1}\mu_n))_k}{(\exp(C^{-1}\mu_x))_k + (\exp(C^{-1}\mu_n))_k} \tag{3}$$

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

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

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

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



<그림 2> JA 방식의 개요

<그림 2>에는 JA 방식에 대한 전반적인 흐름도가 나타나 있다. 먼저, 주어진 잡음환경에서 잡음음성에 대한 HMM 파라미터를 훈련과정에서 추정하게 된다. 이 때 추정된 HMM을 기준 HMM으로 부른다. 기준 HMM을 추정하기 위해서는 PMC 등의 모델보상 방식을 사용하는 것이 일반적이다. 인식시에는 잡음음성신호로부터 잡음신호를 추출한 후, 훈련시에 추정된 기준 HMM의 파라미터값을 Jacobian 행렬을 이용하여 보상하게 된다.

2.3 Data-driven Jacobian Adaptation

Data-driven JA (D-JA) 방식에서는 JA에서 기준 HMM을 모델결합방식을 이용하여 얻는 대신에 잡음음성을 이용하여 직접 훈련하는 방식이 채택되었다[5]. 이것은 일반적으로 모델결합방식으로 얻어진 HMM이 실제 환경의 잡음음성을 이용하여 직접 훈련된 HMM에 비해서 그 인식성능이 다소 떨어진다는 생각에 기반하고 있다. 비록 JA 방식이 기준 HMM을 Jacobian 행렬을 이용하여 잡음음성에 적응시키는 것을 주요장점으로 하고 있지만, 기준 HMM의 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 파라미터값들은 보통 Baum-Welch 알고리듬에 의해서 얻어진다[6]. 연속밀도 HMM의 상태 j의 혼합성분 k에 해당하는 평균벡터는 다음과 같은 수식에 의해서 추정된다.

$$E\{\boldsymbol{x}_t\} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{6}$$

여기서 $\gamma_t(j,k)$는 캡스트럼 특징벡터 x_t가 상태 j의 혼합성분 k에 의해서 발생될 확률을 의미하며 T는 특징벡터의 길이를 나타낸다. 식(4)와 (6)을 이용하면, 잡음음성신호에 대한 평균벡터는 다음과 같이 추정된다.

$$E\{\widetilde{y_t}\} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)(y_t + \frac{\partial y_t}{\partial n_t}(n_t - \widetilde{n_t}))}{\sum_{t=1}^{T}\gamma_t(j,k)} \tag{7}$$

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

$$E\{\widetilde{y_t}\} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)y_t}{\sum_{t=1}^{T}\gamma_t(j,k)} + \frac{\sum_{t=1}^{T}\gamma_t(j,k)\frac{\partial y_t}{\partial n_t}}{\sum_{t=1}^{T}\gamma_t(j,k)}\Delta n \equiv E\{y_t\} + E\{\frac{\partial y_t}{\partial n_t}\}\Delta n \tag{8}$$

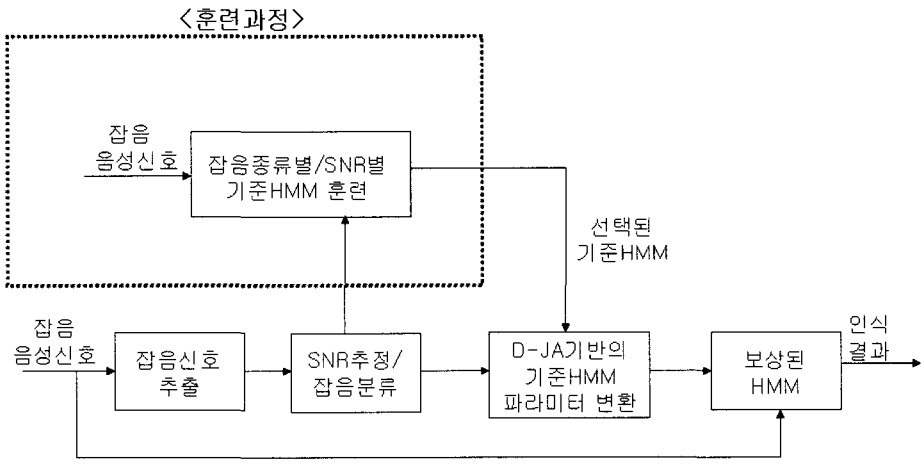
또한, D-JA 방식을 통해서는 정적벡터의 평균값 뿐 만 아니라 공분산 행렬도 비슷한 방식으로 추정이 가능함을 알 수 있다.

3. 다 모델 기반의 잡음음성 인식시스템의 개요

다 모델 기반의 음성인식기에서는 인식시에 가정되는 다양한 잡음환경에 대해서 미리 다수의 기준 HMM을 구성하도록 하며, 인식시에는 이러한 여러 가지 기준 HMM 중에서 테스트 잡음음성에 가장 잘 맞는 것을 선택하여 인식모델로 삼는다. 이런 과정을 거침으로써 단일 기준 HMM을 사용한 경우에 비해서 보다 다양한 잡음에 적응할 수 있는 강인성을 높일 수 있으리라 생각된다.

인식시의 잡음음성에 가장 적합한 인식모델을 찾기 위해서는 잡음음성에 대한 SNR을 추정하고 잡음음성에 포함된 잡음신호의 분류를 하는 것이 필요하다. <그림 3>에는 다 모델 기반의 인식시스템에 대한 흐름도가 나타나 있다.

훈련과정에서는 먼저 다양한 잡음종류별로 별도의 기준 HMM을 구성하게 된다. 이때, 같은 종류의 잡음에 대해서도 몇 가지 SNR 레벨에 따라서 각각 기준 HMM을 구성하게 된다. 연구결과에 따르면 필요한 SNR 레벨은 잡음종류별로 2~3개가 적당하리라 생각된다. 인식과정에서는 잡음신호종류를 분류하고 잡음음성의 SNR 값을 추정한 후 그에 가장 근접한 기준 HMM을 선택하게 된다. 선택된 기준



<그림 3> 다 모델기반의 음성인식시스템의 흐름도

HMM은 D-JA 방식 등의 모델보상 방식을 활용함으로써 더욱 인식성능을 개선시킬 수 있게 된다.

3.1 SNR 추정

SNR을 정의하는 기준은 여러 가지 형태로 표현된다. 일반적으로 많이 알려진 음성신호의 SNR 값은 다음과 같이 정의된다.

$$SNR = 10\log\frac{\sum_{d=0}^{D-1} x^2(d)}{\sum_{d=0}^{D-1} n^2(d)} \tag{9}$$

식 (9)에서의 $x(d)$는 음성샘플을 나타내고 $n(d)$는 잡음샘플을 나타내며 D는 샘플의 개수를 의미한다. 그러나 잡음과 음성신호가 미리 분리되어 있지 않고 잡음음성신호에 서로 섞여있는 경우에는 식 (9)에서 사용되는 잡음샘플과 음성샘플에 대한 전력을 추정하는 과정이 필요하게 된다. 추정 SNR 값은 다음과 같이 정의된다.

$$\widehat{SNR} = 10\ \log\frac{\hat{\sigma}_x^2}{\hat{\sigma}_n^2} \tag{10}$$

식 (10)에서는 잡음신호의 전력 $\hat{\sigma}_n^2$과 음성신호의 전력 $\hat{\sigma}_x^2$이 추정되어야 하며, 이를 위해서는 잡음음성신호 $y(n)$에 대해서 음성구간과 잡음구간을 분리하는 과

정이 필요한데, 일반적으로 VAD 알고리듬을 사용한다. 구간의 길이가 L_n인 분리된 잡음구간에 대해서는 다음과 같이 잡음신호의 전력이 추정된다.

$$\widehat{\sigma_n^2} = \frac{1}{L_n} \sum_{l=0}^{L_n-1} n^2(l) \tag{11}$$

식 (11)에서 구해진 잡음신호의 전력을 이용하여 음성신호의 전력 $\widehat{\sigma_x^2}$은 다음과 같이 잡음음성신호 $y(n)$의 전력과 잡음신호의 전력과의 차이로서 구해진다.

$$\widehat{\sigma_x^2} = \widehat{\sigma_y^2} - \widehat{\sigma_n^2} \tag{12}$$

식 (11)과 (12)를 식 (10)에 대입함으로써 원하는 SNR 값을 얻을 수 있게 된다. 본 연구에서는 위와 같은 방식으로 입력음성신호의 음성구간들에 대해서 식 (10)에 근거한 SNR 값들을 구한 후 그 평균값을 구함으로써 주어진 입력음성에 대한 SNR 값을 추정하였다.

3.2 잡음분류

인식환경에서는 다양한 종류의 잡음이 존재하고 이를 미리 예측하기는 쉽지 않을 것이다. 그러나 다 모델 기반의 음성인식시스템에서는 예상 가능한 여러 가지 종류의 잡음음성신호에 해당하는 각각의 기준 HMM들을 미리 훈련하여 저장한 후, 인식시에는 입력 잡음음성신호에 가장 근접한 기준 HMM을 선택함으로써 최고의 인식성능을 얻을 수 있을 것이다. 따라서 이와 같은 다 모델 기반의 음성인식시스템에서는 입력 잡음음성에 존재하는 잡음이 미리 가정한 잡음신호들 중에서 어느 것에 가장 가까운 것인지를 분류하는 과정이 필요하다. 이를 위해서 우리는 미리 가정된 잡음신호의 특징벡터를 이용하여 GMM을 추정하였다. 각각의 가정된 잡음신호 종류별로 1개씩의 GMM 모델이 생성되었다.

D 차원의 잡음신호 특징벡터 n에 대해서 확률밀도함수는 다음과 같이 정의된다.

$$p(n|\lambda) = \sum_{i=1}^{M} w_i p_i(n) \tag{13}$$

확률밀도함수는 M개의 단일모드 가우시안 밀도함수의 가중 선형결합의 형태를 띠고 있으며, 각각의 가우시한 밀도 함수는 D 차원의 평균벡터 μ_i와 DxD 차원의 공분산 행렬 Σ_i를 갖는다.

$$p_i(n) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\{-\frac{1}{2}(n-\mu_i)'(\Sigma_i)^{-1}(n-\mu_i)\} \quad (14)$$

각 성분의 가중치 w_i는 $\sum_{i=1}^{M} w_i = 1$의 조건을 만족한다. 전체적으로 확률밀도함수의 파라미터값은 $\lambda = \{w_i,\ \mu_i,\ \Sigma_i\}, i = 1,\cdots,M$로 표현된다.

각각의 잡음종류별로 잡음신호의 특징벡터가 구해지면 각 확률밀도함수의 파라미터값들은 expectation-maximization (EM) 방식에 의해서 주어진 잡음신호에 대해서 최대우도를 갖도록 구해지게 된다.

일반적으로 특징벡터들은 서로 독립적이라 가정되며 따라서 잡음특징벡터 열 $n = \{n_1,\cdots,n_T\}$에 대한 로그 유사도(log-likelihood) 값은 다음과 같이 계산된다.

$$\log p(n|\lambda) = \sum_{t=1}^{T} \log p(n_t|\lambda) \quad (15)$$

따라서 주어진 잡음신호에 대해서 각각의 잡음종류별로 위의 로그 유사도를 계산하여 최대값을 나타내는 잡음종류로 잡음신호는 분류되게 된다.

4. 성능비교 실험

4.1 기반인식시스템의 개요

본 연구에서는 잡음환경에서 화자독립 단어 인식실험을 통해서 제안된 방식의 성능을 평가하였다. 인식대상 어휘는 음소분포가 비교적 고르게 되어 있는 한국어 75단어이며 음향모델을 위한 기본단위는 32개의 유사음소를 사용하였다. 각각의 유사음소단위는 3개의 상태(state)를 가진 left-to-right 형태의 연속밀도 HMM에 의해서 모델링되며 각각의 상태는 6개의 혼합성분을 가진다. 화자의 수는 80명이며 이들은 각각 75단어를 한번 씩 발성하였다. 인식실험을 위해서 잭-나이프(Jack-knife) 방식을 이용하였다. 전체 화자를 20명씩 4개의 그룹으로 나눈 후, 그 중 하나의 그룹은 인식용으로 나머지 3그룹은 훈련용으로 활용하였다. 이와 같은 과정을 4회 반복하여 인식실험을 수행하여 인식화자의 수를 4배로 증가시키는 효과를 거두도록 하였다. 잡음음성을 얻기 위해서는 원래의 깨끗한 음성에 차량(car)잡음, 배블(babble)잡음, 전시회(exhibition)잡음, 지하철(subway)잡음 그리고 거리(street)잡음을 다양한 신호대잡음비에 맞추어 더해 주었다. 잡음신호는 AURORA 2 음성데이터에 있는 잡음파일로부터 얻었으며 샘플링율을 16 kHz로 변환한 후 이를 75단어의 음성데이터에 첨가하였다[7]. 인식특징벡터로는 13차의 멜주파수

(Mel-Frequency) 켑스트럼 계수(MFCC)와 그의 차분계수(delta-MFCC)를 사용하였다.

4.2 인식실험 결과

먼저, <표 1>에서는 본 연구에서 사용된 베이스라인(baseline)인식기의 성능 및 재훈련과 MCT 결과 그리고 여러 가지 모델 보상 방식에 대한 인식 성능 결과가 나타나 있다. 인식결과는 차량잡음, 배블잡음, 전시회잡음 그리고 지하철잡음음성의 각각에 대한 인식률을 평균하여 나타내었다.

<표 1> 베이스라인 인식기 및 재훈련과 MCT 그리고 모델보상 방식들에 의한 단어 인식율(%)의 비교

SNR (dB)	0	5	10	15	20	25	30	clean
베이스라인 인식기	10.8	28.1	54.2	77.2	90.1	95.5	96.3	98.6
재훈련 방식	78.9	89.4	93.8	95.9	97.0	97.7	98.0	98.6
MCT	65.0	83.9	91.3	94.1	95.3	95.9	96.3	97.4
PMC	61.3	77.2	86.9	92.0	94.9	96.6	97.4	98.4
JA	63.2	78.3	87.1	92.1	95.1	96.5	97.3	98.4
D-JA	78.7	89.3	93.9	96.0	97.0	97.8	98.0	98.6

베이스라인 인식기는 잡음에 오염되지 않은 원래의 깨끗한 음성데이터를 이용하여 HMM을 Baum-Welch 알고리듬에 근거하여 훈련한 결과를 의미하는데 예상한 대로 SNR 값이 낮은 영역에서는 매우 심한 성능 저하를 보임을 알 수 있다.

재훈련방식은 인식과 훈련시에 동일한 SNR 값을 사용하므로 일반적으로 매우 좋은 인식성능을 보이며 잡음음성인식에 사용되는 여타의 모델보상 방식에 비해서도 성능이 좋은 것으로 나타난다. 본 연구에서도 재훈련방식은 기존의 모델보상 방식인 PMC나 JA에 비해서는 월등히 성능이 우수한 것으로 나타났다.

MCT 방식은 AURORA DB에 관한 인식실험에서 소개된 훈련방식으로 여러 가지의 잡음종류와 다양한 SNR(clean, 5~20 dB)을 고려한 잡음음성 데이터 set을 훈련시에 이용하여 HMM 모델을 구하는 방식이다. 따라서 MCT 훈련을 통해서 다양한 음향조건이 HMM 파라미터에 반영되도록 할 수 있다. 본 논문에서는 MCT 방식에 의한 훈련을 위해서 차량잡음, 배블잡음, 전시회잡음 그리고 지하철잡음을 고려하였다. MCT 방식은 기본적으로 여러 가지 잡음조건을 하나의 HMM 모델에 함축함으로써 다양한 잡음특성을 포함한 HMM이 생성될 수 있으나, 이로 인하여 HMM 모델 파라미터값의 평탄화(smoothing)가 지나쳐서 인식성능이 재훈련 방식에 비해서 우수하지 못한 것이 일반적이며 본 연구에서도 <표 1>의 결과로부터 확인할 수 있었다.

<표 1>의 결과로부터 우리는 기존의 모델보상 방식 중 에서는 D-JA 방식이 상대적으로 우수한 성능을 보임을 알 수 있었는데, 이것은 D-JA 방식은 적응을 위한 기준 HMM으로 재훈련 HMM을 그대로 사용하기 때문이다. 따라서 훈련과 인식시에 잡음특성이 별 차이가 없다면 재훈련 방식과 D-JA 방식은 비슷한 결과를 나타내리라 생각된다. <표 1>의 결과에서는 D-JA 방식과 재훈련 방식에서 테스트 잡음음성의 SNR 값을 미리 안다고 가정하므로 D-JA 방식과 재훈련 방식간의 인식성능의 차이가 그리 크지 않음을 알 수 있다.

<표 1>에서 우리는 베이스라인 인식기의 성능과 재훈련 및 MCT와 같은 HMM의 훈련 방식에 따른 인식기 성능의 차이 및 모델보상방식에 따른 인식성능을 확인할 수 있었다. 전체적으로 보아서 재훈련방식과 D-JA 방식이 PMC나 MCT 등의 방식에 비해서 우수한 성능을 보임을 알 수 있었다. <표 1>의 결과를 바탕으로 우리는 D-JA 방식을 다 모델 기반의 음성인식시스템의 모델보상 방식으로 사용하였으며, 사용되는 HMM 모델의 개수에 따른 인식성능의 변화를 살펴보았다.

<표 2>는 다 모델기반의 음성인식기에서 잡음의 종류가 미리 알려진 경우에 훈련잡음음성의 SNR 값에 따른 기준 HMM 모델의 개수에 따라서 인식성능이 변화하는 것을 나타낸다. 각각의 기준 HMM 모델을 훈련하기 위해서 재 훈련 방식에서 사용한 잡음음성데이터를 그대로 이용하였고 모델 보상방식으로는 D-JA 방식을 사용하였다. <표 1>의 결과와는 달리 <표 2>를 포함하여 향후 언급되는 인식결과들을 얻기 위해서는 테스트 음성의 SNR 값이 3.1절에 언급된 방식으로 추정되어진다.

<표 2> 다 모델기반의 음성인식기에서 잡음의 종류가 미리 알려진 경우에 기준 HMM 모델의 개수에 따른 단어 인식율(%)의 비교 (D-JA 방식 적용시)

SNR (dB) / HMM 모델 개수	0	5	10	15	20	25	30
1 (15 dB)	77.6	89.5	94.1	96.0	97.0	97.5	97.2
2 (10, 20 dB)	79.5	89.9	93.9	96.0	97.1	97.7	97.8
3 (0, 10, 20 dB)	78.9	88.6	93.8	96.0	97.1	97.6	97.8
5 (0, 5, 10, 20, 25 dB)	79.2	89.2	93.7	96.0	97.1	97.8	98.0
7 (0, 5, 10, 15, 20, 25, 30 dB)	79.2	89.2	93.7	96.0	97.1	97.7	98.0

우리는 4가지(자동차, 배블, 전시회, 지하철) 종류의 잡음음성 각각에 대해서 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB, 30 dB 등의 SNR 값에 각각 해당하는 7개의 기준 HMM 모델을 재 훈련과정을 통해서 구성하였다. <표 2>의 다 모델기반의 음성인식결과에서는 이중에서 일부분의 HMM 모델만을 사용할 경우의 인식성능의 변화를 살펴보았다. <표 2>에서 HMM 모델 개수를 나타내는 숫자의 괄호 안에

는 사용된 모델에 해당하는 SNR 값들이 나열되어 있다. <표 2>의 결과를 보면 7개의 전체 모델을 사용한 경우와 단지 2개의 모델(10 dB와 20 dB)만을 사용한 경우의 인식율에 큰 차이가 나지 않음을 알 수 있다. 이 결과를 보면 우리가 생각하는 것보다는 인식결과가 SNR 값에 따른 훈련모델의 개수에 그리 민감하게 반응하지 않음을 알 수 있다. 본 연구에서는 최적의 성능을 나타내는 HMM의 모델 개수는 다소 경험적으로 설정되어진다. 위의 <표 2>에서 보듯이 우리는 10 dB와 20 dB의 HMM 모델을 사용하는 것이 매우 효과적임을 알 수 있는데, HMM 모델의 SNR 값이 다소 변화더라도 인식성능에는 그리 큰 변화가 있지는 않을 것으로 생각된다. 그러나, 다양한 실험결과 제안된 SNR 값들이 가장 무난할 것으로 생각된다.

<표 2>에서 인식모델의 개수에 따른 인식성능이 그리 민감하지 않은 원인이 HMM의 파라미터값이 원래부터 훈련음성의 SNR 값에 그리 민감하지 않기 때문인지 아니면 다 모델기반의 음성인식에서 사용되는 모델보상 방식이 이러한 SNR 값의 변화를 보상시켜주기 때문인지를 알기 위해서 우리는 <표 3>에서 다 모델 기반의 음성인식기에서 D-JA 방식의 모델보상을 사용한 경우와 그렇지 않은 경우에 대해서 비교하였다. 인식실험을 위해서는 <표 2>에서 언급한 4가지 종류의 잡음을 사용하였다.

<표 3> 다 모델기반의 음성인식기에서 잡음의 종류가 미리 알려진 경우에 D-JA방식의 모델보상을 적용한 경우와 그렇지 않은 경우의 단어 인식율(%)의 비교

SNR (dB) / HMM 모델 개수		0	5	10	15	20	25	30
D-JA 보상 적용	1	77.6	89.5	94.1	96.0	97.0	97.5	97.2
	2	79.5	89.9	93.9	96.0	97.1	97.7	97.8
	3	78.9	88.6	93.8	96.0	97.1	97.6	97.8
	5	79.2	89.2	93.7	96.0	97.1	97.8	98.0
	7	79.2	89.2	93.7	96.0	97.1	97.7	98.0
보상 적용 않음	1	55.1	84.6	93.6	95.9	96.6	96.3	95.4
	2	72.0	88.9	93.8	95.5	96.9	97.4	97.3
	3	79.0	86.7	93.4	95.5	96.9	97.3	97.3
	5	79.4	88.8	93.4	95.4	96.8	97.6	97.8
	7	79.4	88.8	93.5	95.7	96.9	97.4	97.9

<표 3>의 결과에서 보면, 보상을 적용하지 않고 재훈련 모델을 그대로 사용한 경우에는 D-JA 보상을 적용한 경우에 비해서 인식성능이 다소 저조한 것을 알 수 있다. 특히 기준 HMM의 모델개수가 적은 경우에는 그 영향이 심하게 나타나는 것을 알 수 있다. 그러나 보상을 사용하지 않은 경우에도 모델의 개수가 3 이상이

면 7개를 전부 사용한 경우와 비교해서 성능의 차이가 그리 크지 않은 것을 알 수 있었다. 이러한 결과로부터 기본적으로 HMM의 모델 파라미터값은 SNR 값의 변화에 그리 민감하지 않음을 알 수 있었다.

<표 2>의 다 모델 기반의 음성인식에서 우리는 잡음의 종류를 안다고 가정하였다. 그러나 실제인식환경에서는 일반적으로 인식 음성에 포함된 잡음의 종류를 미리 알지 못하는 경우가 대부분인데 이러한 경우에는 자동적으로 이를 판단하는 잡음분류기가 필요하다. GMM 기반의 잡음모델을 이용하여 4가지 종류의 잡음신호에 대하여 훈련을 통하여 GMM의 평균값과 공분산 행렬들을 추정하였다. GMM의 혼합성분의 개수는 5로 하고 테스트 음성에 포함된 잡음의 종류를 분류하는 실험을 수행하였다. <표 4>에 그 결과가 나타나 있다. 잡음신호의 분류를 위해서는 테스트 음성의 묵음구간인 처음 20프레임을 이용하였다. 아래의 표에서 보듯이 높은 잡음분류성능을 나타냄을 알 수 있다.

<표 4> GMM 방식에 근거한 잡음신호의 분류 결과(%) 비교

훈련잡음 / 테스트잡음	자동차	배블	전시회	지하철
자동차	99.2	0.8	0.0	0.0
배블	0.4	99.6	0.0	0.0
전시회	0.1	0.0	99.9	0.0
지하철	0.0	0.0	0.3	99.7

<표 5> 다 모델기반의 음성인식기에서 잡음의 종류를 미리 모르는 경우에 기준 HMM 모델의 개수에 따른 단어 인식율(%)의 비교 (D-JA 방식 적용시)

SNR (dB) / 잡음종류별 HMM 모델 개수		0	5	10	15	20	25	30
D-JA 보상 적용	1	77.7	89.6	94.1	96.1	97.0	97.5	97.2
	2	79.5	89.9	93.9	96.0	97.1	97.6	97.8
	3	78.9	88.6	93.8	96.0	97.1	97.6	97.8
	5	79.2	89.2	93.6	96.0	97.1	97.8	98.1
	7	79.2	89.2	93.7	96.0	97.1	97.7	98.0

<표 5>에서는 잡음분류기를 인식의 전단계에 둔 경우의 다 모델 기반 인식기의 성능을 나타내었다. 인식율은 <표 2>의 결과의 거의 유사한 것을 알 수 있었는데, 이것은 GMM에 기반한 잡음분류기의 성능이 4가지 종류의 잡음에 대해서 99% 이상으로 매우 우수하여 잡음분류의 오류에 의한 영향을 최소화하기 때문으로 생각된다.

마지막으로 <표 6>에서 우리는 훈련 중에 사용되지 않은 새로운 잡음에 오염된 잡음음성이 입력될 경우에 다 모델 기반의 음성인식기의 인식율을 나타내었다. 사용된 잡음신호는 AURORA DB set B에 있는 길거리(street)잡음신호이다. D-JA 보상방식을 사용한 경우와 보상을 적용하지 않고 재훈련모델을 그대로 사용한 경우에 대해서 성능비교를 하였는데, 우리는 모델 보상을 사용한 경우가 재 훈련모델을 사용한 경우에 비해서 월등한 성능을 나타냄을 알 수 있었다. 이것은 <표 3>의 경우처럼 테스트 입력 음성에 포함된 잡음이 미리 훈련에 사용된 경우와 대조되는데, 이것은 모델보상 방식을 사용함으로써 훈련에 사용되지 않은 잡음신호가 테스트 잡음음성에 존재하더라도 충분히 보상이 가능함을 보여준다. 이 경우에도 필요로 하는 HMM 모델의 개수는 각 잡음종류별로 2~3개 정도면 충분한 것으로 보인다.

<표 6> 다 모델기반의 음성인식기에서 새로운 잡음이 테스트 잡음 음성에 존재하는 경우에 D-JA방식의 모델보상을 적용한 경우와 그렇지 않은 경우의 단어 인식율(%)의 비교

SNR (dB) / 잡음종류별 HMM 모델 개수		0	5	10	15	20	25	30
D-JA 보상 적용	1	80.6	90.5	94.7	96.3	96.7	96.2	95.4
	2	82.7	91.0	94.2	96.1	96.8	97.1	97.2
	3	84.1	90.8	93.7	96.0	96.8	97.1	97.2
	5	84.6	91.0	94.1	96.1	96.8	97.5	97.8
	7	84.6	90.9	94.3	96.2	97.0	97.5	97.9
보상 적용 않음	1	62.6	82.7	91.2	94.4	95.2	94.5	93.2
	2	72.5	86.9	92.4	95.1	96.3	96.5	96.5
	3	77.3	86.9	91.9	94.9	96.3	96.5	96.5
	5	77.9	87.6	92.0	95.1	96.3	97.0	97.3
	7	77.9	87.8	92.1	95.2	96.4	97.1	97.6

5. 결 론

본 연구에서는 다 모델 기반의 음성 인식시스템에서 D-JA 방식의 모델 보상을 사용하였으며, 그 성능이 기존의 다른 모델 보상방식에 비해서 우수함을 보였다. 또한 다양한 실험을 통해서 다 모델 기반의 음성인식기에서 필요로 하는 기준 HMM의 개수에 대해서 조사하였다. 다 모델 기반의 음성 인식시스템은 기본적으로 단일모델의 음성인식시스템에 비해서 우수한 성능을 나타낸다. 그러나 D-JA와 같은 모델 보상 기법을 사용함으로써 성능의 향상과 더불어 HMM 모델수의 증가

에 따른 부담을 줄일 수 있었다. 모델보상 방식을 사용함으로써 훈련시와 인식시의 잡음의 종류 및 SNR 값에서 발생하는 차이를 극복할 수 있음을 알 수 있었으며, 이를 위해서는 각 잡음종류별로 2~3개 정도의 SNR 값에 따른 모델을 구성하면 충분한 것으로 생각된다. 다 모델 기반의 음성 인식시스템은 비록 모델 개수의 증가를 통한 부가적인 부담을 인식시스템에 지우지만 기존의 여러 가지 잡음음성 인식을 위한 방법에 비해서 매우 우수한 인식성능을 보임을 알 수 있었다.

참 고 문 헌

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

[2] S. Sagayama, Y. Yamaguchi, S. Takahashi, "Jacobian adaptation of noisy speech models", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 396-403, 1997.

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

[4] H. Xu, Z. Tan, P. Dalsgaard, B. Lindberg, "Robust speech recognition based on noise and SNR classification - a multiple-model framework", *Proc. Interspeech*, pp. 977-980, 2005.

[5] 정용주, "잡음음성인식을 위한 데이터기반의 Jacobian 적응방식", *한국음향학회지*, 25권, 4호, pp. 159-163, 2006.

[6] L. E. Baum, G S. T. Petrie, N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Ann. Math., Statist.*, Vol. 41, No. 1, pp. 164-171, 1970.

[7] D. Pearce, H. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under conditions", *Proc. ICSLP*, Vol. IV, pp. 29-32, 2000.

접수일자: 2007년 5월 10일
게재결정: 2007년 6월 25일

▶ 정용주(Young-Joo Chung) : 교신저자
주소: 704-701 대구광역시 달서구 신당동 1000, 계명대학교
소속: 계명대학교 전자공학과
전화: 053) 580-5925
E-mail: yjjung@kmu.ac.kr

▶ 곽성우(Seung-Woo Kwak)
주소: 704-701 대구광역시 달서구 신당동 1000, 계명대학교
소속: 계명대학교 전자공학과
전화: 053) 580-5926
E-mail: swkwak@kmu.ac.kr