연속 잡음 음성 인식을 위한 다 모델 기반 인식기의 성능 향상에 대한 연구*

Performance Improvement in the Multi-Model Based Speech Recognizer for Continuous Noisy Speech Recognition

정 용 주**
Yongjoo Chung

ABSTRACT

Recently, the multi-model based speech recognizer has been used quite successfully for noisy speech recognition. For the selection of the reference HMM (hidden Markov model) which best matches the noise type and SNR (signal to noise ratio) of the input testing speech, the estimation of the SNR value using the VAD (voice activity detection) algorithm and the classification of the noise type based on the GMM (Gaussian mixture model) have been done separately in the multi-model framework. As the SNR estimation process is vulnerable to errors, we propose an efficient method which can classify simultaneously the SNR values and noise types. The KL (Kullback-Leibler) distance between the single Gaussian distributions for the noise signal during the training and testing is utilized for the classification. The recognition experiments have been done on the Aurora 2 database showing the usefulness of the model compensation method in the multi-model based speech recognizer. We could also see that further performance improvement was achievable by combining the probability density function of the MCT (multi-condition training) with that of the reference HMM compensated by the D-JA (data-driven Jacobian adaptation) in the multi-model based speech recognizer.

Keywords: Noisy speech recognition, Hidden Markov model, model compensation, multi-model based speech recognizer

1. 서 론

음성인식시스템의 성능 저하를 야기 하는 중요한 원인중의 하나는 음성인식시스템이 작동하는 주변의 배경잡음이나 마이크나 통신선로 등에 의한 채널왜곡을 들 수 있다. 이러한 신호 왜곡 현상에 대처하기 위한 방법으로는 음질을 개선한다든지 강인한 음성 특징을 사용한다든지 하는 방법 외에도 음성인식기의 HMM(hidden Markov model) 파라미터 값을 보상해주는 방법들이 사용되고

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

있다(Gales, 1993; Moreno, 1996; Hung, 2001).

그러나 이와 같은 다양한 방법에도 불구하고 음성인식시스템의 성능은 다양한 잡음환경에 충분한 대처가 되지 못하고 있는 실정이다. 최근에는 잡음환경에 보다 강인한 음성인식시스템을 구현하기 위해서 단일 HMM 대신에 다수의 HMM을 사용하는 방식이 채택되고 있다(Xu, 2005). 이와 같은 다 모델 기반의 음성인식시스템에서는 잡음종류별과 신호대잡음비(SNR: Signal to Noise Ratio) 별로 다수의 HMM을 미리 훈련하여 인식시에 이들 중 잡음환경에 가장 가까운 모델을 선택함으로서 훈련환경과 인식환경의 차이가 자연스럽게 최소화 되도록 한다.

기존의 연구에서 우리는 다 모델 기반의 음성인식시스템에서 데이터 기반의 Jacobian 적응방식이 매우 효과적임을 보였다(정용주, 2007). 이러한 다 모델 기반의 HMM 파라미터 적응방식을 위해서는 입력음성에 가장 적합한 기준 HMM 을 선택하기 위해서 입력 잡음음성신호의 SNR을 계산하고 입력 잡음음성신호에 포함된 잡음의 종류를 추정하는 작업을 수행하였다. SNR을 계산하기 위해서는 VAD(voice activity detection) 을 사용하여 잡음구간을 추정하는 방식을 사용하여 SNR 값의 신뢰성을 높이도록 하였고 잡음의 종류를 추정하기 위해서는 잡음종류별로 GMM(Gaussian mixture model)을 구성하여 잡음분류가 가능하도록 하였다. GMM을 이용한 잡음분류는 매우 만족스러운 결과를 나타내었으나, VAD 기반의 SNR 추정방식은 연속음성인식의 경우에 다소 많은 계산을 요구하여 때때로 매우 부정확한 SNR 값을 추정하는 경우가 자주 발생하였다. 따라서 본 연구에서는 기존의 방식에서 잡음종류 추정과 SNR 값 추정을 각각 따로 수행함으로서 발생하는 오류를 줄이고 계산의 간편함을 도모하기 위해서 위의 두 과정을 한 번의 계산으로 마치도록 하는 방안을 제안하였다. 여기서는 훈련과정에서 각각의 잡음종류별과 SNR 별로 별도의 single Gaussian 모델을 만든 후, 인식과정에서의 입력 잡음음성신호에 존재하는 잡음신호의 평균과 분산 값을 이용하여 입력 잡음신호와 가장 거리가 가까운 single Gaussian 모델을 찾아내는 방법을 이용함으로서 잡음의 종류와 SNR 인식을 동시에 수행할 수 있도록 하였다. 이를 위해서는 Gaussian 확률 분포들 간의 KL(Kullback-Leibler) distance를 이용 하였다.

한편 본 연구에서는 기준 HMM을 훈련과정에서 생성하는 다양한 방법들에 대해서 논의하고 성능분석을 통하여 서로 다른 방법에 의해서 생성된 기준 HMM 간의 결합을 통하여 보다 우수한 인식 성능을 나타낼 수 있음을 보였다. 한편 기존의 다 모델 기반의 음성인식시스템은 고립단어 인식 실험에 매우 유용함을 보임을 알 수 있었는데, 본 연구에서는 Aurora 2 데이터베이스에 있는 연속 숫자음 잡음음성을 이용하여 모델보상 방식 기반의 다 모델 인식시스템의 효율성에 대해서 고찰하였다.

다음 장에서는 모델보상 방식 기반의 다 모델 인식시스템에 대해서 간략히 소개하며 3장에서는 Aurora 2 데이터베이스를 이용한 다 모델 기반 구조의 인식시스템의 성능에 대한 실험 결과를 논의하고 4 장에서 결론을 맺고자 한다.

2. 모델 보상 방식기반의 다 모델 인식시스템의 개요

2.1 기준 HMM의 훈련 방식에 따른 인식시스템의 형태

음성인식시스템의 구성은 훈련과정과 인식과정으로 나누어진다고 할 수 있다. 훈련과정에서는 인식과정에서 사용될 HMM을 만드는 과정이라 할 수 있으며, 이때 만들어진 HMM은 훈련된 HMM 또는 기준 HMM이라 불리어진다. 훈련시 만들어지는 기준 HMM은 인식성능을 결정하는데 있어서 매우 중요하며 훈련과정과 인식과정에서 발생되는 음성신호간의 음향학적인 특성의 차이, 배경잡음 등의 차이 그리고 화자의 발성에 미치는 여러 가지 요인의 차이 등에 강인한 특성을 갖도록 하는 것이 매우 바람직 할 것이다. <그림 1>에는 일반적으로 음성인식시스템에서 훈련과정에서 기준 HMM을 만드는 방법과 이를 인식시에 활용하는 몇 가지 방안에 대한 개괄적인 구조가 나타나 있다.

가장 일반적으로 많이 사용되는 구조는 <그림 1(a)>처럼 기준 HMM의 구성시에 잡음에 오염되지 않은 깨끗한 음성신호만을 사용하는 것이다. 이러한 경우에 있어서 가장 큰 문제점은 인식과정에서 잡음에 오염된 음성신호에 대한 대처능력이 현저히 떨어진다는 점이다. 이를 보완하기 위해서 오염된 음성신호를 개선시킨다든가 기준 HMM의 파라미터를 오염된 음성신호에 적합하게 보상하는 방안들이 주로 이용된다. 기준 HMM을 구성하는 또 다른 방법은 <그림 1(b)>에서처럼 기준 HMM의 구성시에 미리 잡음에 오염된 음성신호를 사용하는 경우이다. 이 경우에는 미리 인식과정의 잡음환경을 예상하여 그에 해당하는 잡음 음성신호를 수집하여 훈련과정에서 활용하는 방식을 사용한다. 이러한 인식시스템의 경우에도 인식과정의 잡음환경이 미리 예상한 경우와 다를 경우가 존재하며 이에 대처하기 위해서는 HMM의 파라미터를 보상하는 방식을 사용할 수 있다. <그림 1(c)>에 나온 방식은 <그림 1(b)>와 비교하여 훈련과정에서 미리 잡음에 오염된 음성신호를 사용한다는 유사점이 있으나 인식과정에서 발생하는 잡음환경을 미리 구체적으로 예측하지 않는다는 차이점이 있다. 여기서는 몇 가지 종류의 잡음과 다양한 SNR을 모두 고려한 잡음 음성신호를 이용하여 기준 HMM을 구성하게 된다. 따라서 기준 HMM의 파라미터 값들은 특정한 잡음종류나 SNR 값에 전적으로 의존하지 않아서 실제 인식과정에서의 잡음신호의 다양한 특성에 대비할 수 있도록 되어 있다. 이 방식에서는 앞의 방식에서 사용되는 음질개선이나 HMM 파라미터 보상 방식들을 추가적으로 사용하지는 않는다. 이러한 훈련 과정을 MCT(multi-condition training)라 부른다. 마지막으로 <그림 1(d)>는 비교적 최근에 제안된 방식으로서 <그림 1(b)>와 <그림 1(c)>의 장점을 취합하여 이루어진다. <그림 1(b)>에서는 훈련과정에서 고려되지 않은 잡음 신호가 인식과정에서 입력될 경우에는 인식성능의 저하가 많이 발생하는 문제가 있고 반면에 <그림 1(c)>에서는 다양한 잡음신호의 특성이 복합적으로 HMM의 파라미터에 반영되었기 때문에 특정한 잡음신호에 대해서는 분별력이 다소 떨어진다는 단점이 있다. 따라서 <그림 1(d)>의 방식에서는 잡음종류별과 SNR 별로 다수의 기준 HMM을 독립적으로 생성하고 인식과정에서의 잡음신호와 가장 유사한 기준 HMM을 선택하여 인식을 수행하는 방식을 취하고 있다. 이렇게 함으로서 <그림 1(b)>의 경우와 같이 단독 기준 HMM을 사용할 경우의 인식성능이 저하되는 문제와 <그림 1(c)>의 경우처럼 기준HMM이 다소 방대한 음향학적 스펙트럼을 포함함으로서 생기는 분별력의 저하를 모두 극복할 수 있는 장점이 있다. 또한 이러한 방식에서는 HMM 파라미터 보상방식을 사용함으로서 추가적인 인식성

능의 향상을 꾀할 수 도 있다. 이러한 방식의 인식시스템은 일반적으로 다 모델기반 구조의 음성인식시스템이라 한다.

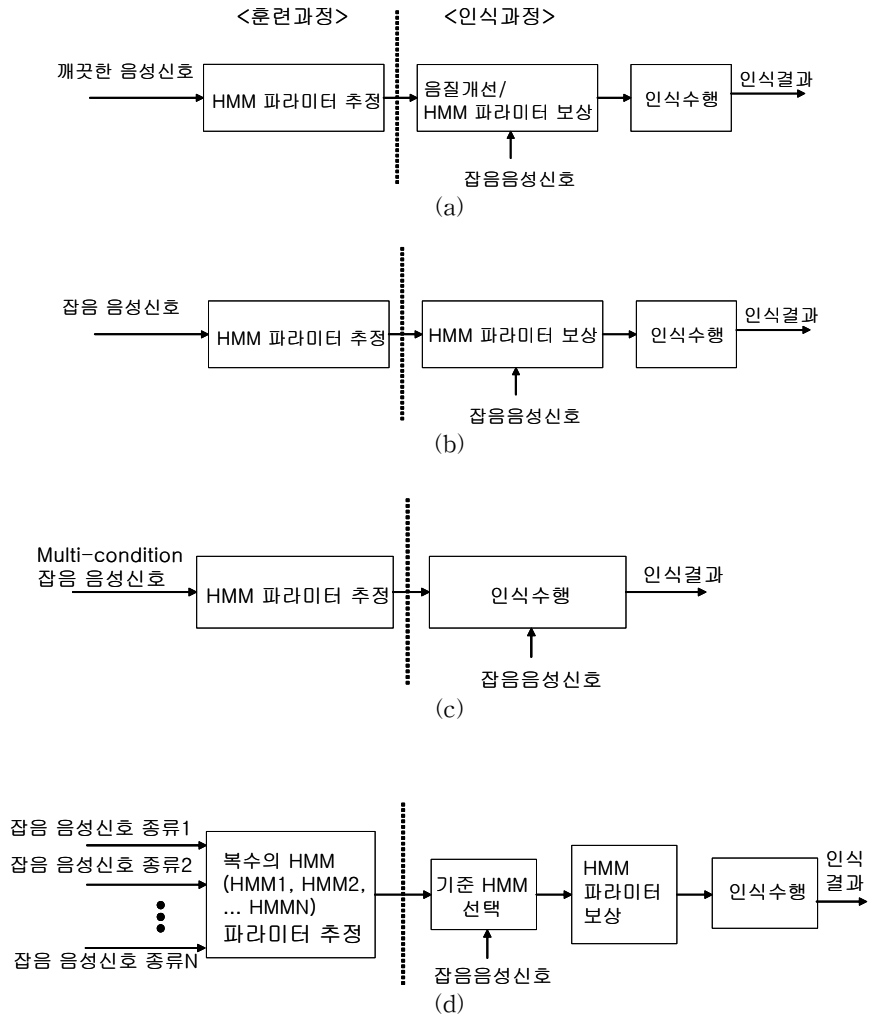

그림 1. 기준 HMM의 구성방식에 따른 인식기의 다양한 형태

위에서 제시된 여러 가지 형태의 음성인식시스템 구성 중에서 본 연구에서 다 모델 기반 구조의 음성인식시스템의 성능개선에 중점을 두고 연구하였다. 특히, 다 모델 기반 구조에서는 여러 종류의 기준 HMM중 잡음음성신호와 가장 근접한 것을 선택하는 것이 중요한데, 이를 위해서는 인식 잡음음성신호의 SNR 값을 추정하고 포함된 잡음신호의 종류를 찾아내는 과정이 필요하다. 또한 이러한 과정을 통해서 선택된 기준 HMM을 인식에 그대로 사용하는 것 보다는 모델 보상방식을 적용하여 인식성능의 향상을 이끌어 내는 과정이 필요하게 된다. 2.2절과 2.3절에서는 이러한 과정에 대해서 보다 상세한 설명을 하도록 하겠다.

2.2 SNR 값과 잡음 종류의 추정을 이용한 기준 HMM의 선택

2.2.1 SNR 값의 추정

SNR 값을 추정하기 위해서 잡음음성신호의 샘플벡터 $\boldsymbol{x}$ 는 다음과 같이 표현된다고 가정한다.

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n} \tag{1}$$

여기서 $\boldsymbol{s}$ 와 $\boldsymbol{n}$ 는 각각 원래의 깨끗한 음성과 잡음신호의 샘플값을 나타낸다.
식(1)의 잡음음성신호 $\boldsymbol{x}$ 의 SNR 값은 일반적으로 다음과 같이 추정된다.

$$\widehat{SNR} = 10\ \log\frac{\widehat{\sigma_s^2}}{\widehat{\sigma_n^2}} \tag{2}$$

여기서 잡음신호 $\boldsymbol{n}$ 의 전력 $\widehat{\sigma_n^2}$ 과 원래 음성신호 $\boldsymbol{s}$ 의 전력 $\widehat{\sigma_s^2}$이 추정되어야 하며, 이를 위해서는 VAD를 이용하여 잡음음성신호의 음성구간과 비음성구간이 분리되어야 한다. 비음성구간을 이용한 잡음신호의 전력은 다음과 같이 추정된다.

$$\widehat{\sigma_n^2} = \frac{1}{l_n}\sum_{l=0}^{l_n} n^2(l) \tag{3}$$

여기서 l_n 은 분리된 잡음구간의 샘플 수이다. 위식을 이용하여 원래 음성신호의 전력 $\widehat{\sigma_s^2}$은 다음과 같이 잡음음성신호의 전력과 잡음신호의 전력의 차이로서 구해진다.

$$\widehat{\sigma_s^2} = \widehat{\sigma_x^2} - \widehat{\sigma_n^2} \tag{4}$$

위의 식 (3)과 (4)를 식(2) 에 대입함으로서 원하는 SNR 값을 얻게 된다.
위와 같은 SNR 값의 추정방식에서는 잡음음성신호로부터 잡음구간의 추정이 중요하며 SNR 값이 낮은 경우 잡음구간 추정에서 오류가 발생할 가능성이 높다.

2.2.2 잡음신호의 종류 인식

잡음신호의 종류 인식을 위해서는, 일반적으로 잡음신호의 특징벡터 $\boldsymbol{N}$ 은 GMM으로 모델링되며 이는 M 개의 단일모드 가우시안 밀도함수의 가중 선형결합의 형태를 띠고 있으며, 아래의 식과 같다.

$$p(\boldsymbol{N}) = \sum_{i=1}^{M} w_i\, p_i(\boldsymbol{N}) \ = \ \sum_{i=1}^{M} w_i \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\{-\frac{1}{2}(\boldsymbol{N}-\boldsymbol{\mu_i})'(\Sigma_i)^{-1}(\boldsymbol{N}-\boldsymbol{\mu_i})\} \tag{5}$$

각 성분의 가중치 w_i 는 $\sum_{i=1}^{M} w_i = 1$ 의 조건을 만족하며 μ_i, Σ_i 는 각각 가우시안 밀도함수의 D 차원의 평균벡터와 공분산 행렬을 나타낸다. 잡음신호 종류의 인식을 위해서는 잡음종류별로 위와 같은 GMM을 미리 훈련과정에 얻어야 하며 이를 위해서는 expectation-maximization (EM) 방식에 기반한 최대우도 추정방식을 사용하여 식(5)의 각각의 가우시안 밀도함수의 평균과 공분산 그리고 가중치를 얻게 된다.

2.2.3 Kullback-Leibler(KL) distance 기반의 잡음신호 분류

기존의 다 모델기반 구조의 음성인식시스템에서 최적의 기준 HMM을 선택하기 위해서는 앞서 설명한 바와 같이 SNR 값 추정과 잡음신호의 종류를 분류하는 두 가지 과정을 거치게 된다. 그러나 SNR 값 추정방식은 VAD의 정확성 등이 요구되는 부담이 있다. 그리고 SNR 값 추정과 잡음신호의 종류 인식을 각각 수행함으로서 번거로움과 함께 인식오류가 높아지게 된다. 따라서 본 연구에서는 이와 같은 두 가지 과정을 하나로 합하여 다 모델기반 구조의 음성인식시스템에서 기준 HMM을 선택하는데 있어서 효율성을 높이도록 하였다. 이를 위해서는 훈련과정에서 각 SNR 별 그리고 잡음의 종류별의 단일모드 가우시안 밀도함수를 추정한다. 이러한 각각의 단일모드 가우시안 밀도함수들은 인식과정의 잡음신호로부터 얻은 가우시안 밀도함수와의 KL distance 값을 통하여 잡음신호와의 유사도를 측정 받게 된다. 두 개의 단일 모드 가운시안 확률밀도 함수 $N_1(\mu_1, \Sigma_1), N_2(\mu_2, \Sigma_2)$ 간의 KL distance 값은 다음식과 같다(Juang, 1984).

$$KLD(N_1, N_2) = \frac{1}{2}\sum_{i=1}^{D}\left[\log\left(\frac{(\Sigma_1)_{ii}}{(\Sigma_2)_{ii}}\right) + \frac{\left((\mu_2)_i - (\mu_1)_i\right)^2}{(\Sigma_1)_{ii}} + \left(\frac{(\Sigma_2)_{ii}}{(\Sigma_1)_{ii}} - 1\right)\right] \quad (6)$$

2.3 다 모델 기반 인식기에서의 모델 보상 방식

다 모델 기반 구조의 인식기는 잡음종류와 SNR 별로의 기준 HMM을 구성하여 인식환경에서의 잡음신호 변이를 효과적으로 대처하지만 보다 나은 성능향상을 위해서는 기준 HMM의 파라미터 값을 인식과정의 잡음신호를 이용하여 보상하는 방법이 많이 사용된다.

이때 사용되는 가장 대표적인 모델 보상방식으로는 JA과 이를 개선한 data-driven JA 방식이 있다. 이를 간단히 소개하면 다음과 같다.

2.3.1 Jacobian adaptation (JA)

JA 방식에서는 어느 특정한 잡음신호에 기반한 기준 HMM 파라미터 값을 인식과정의 잡음신호에 맞도록 변환한다. 켑스트럼 영역에서 부가잡음 신호 N 에 대하여 원래의 깨끗한 음성신호 S 는 다음과 같이 변환된다.

$$X = C\left[\log\{\exp(C^{-1}S) + \exp(C^{-1}N)\}\right] \quad (7)$$

여기서 $\boldsymbol{X}$는 잡음음성신호를 나타내며 $\boldsymbol{C}$는 DCT(Discrete Cosine Transformation)을 나타낸다. 이 경우 잡음신호가 $\boldsymbol{N}$ 에서 $\widetilde{\boldsymbol{N}}$ 로 변할 경우의 잡음음성신호 $\boldsymbol{X}$의 변이는 다음식과 같이 나타낼 수 있다.

$$\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \frac{\partial \boldsymbol{X}}{\partial \boldsymbol{N}}(\boldsymbol{N} - \widetilde{\boldsymbol{N}}) \tag{8}$$

식(8)의 양변에 평균치를 취하면 잡음신호의 변이에 의한 기준 HMM의 평균파라미터 값의 변이를 얻을 수 있다.

2.3.2 Data-driven Jacobian adaptation (D-JA)

Data-driven JA (D-JA) 방식에서는 JA에서 기준 HMM을 모델결합방식을 이용하여 얻는 대신에 잡음음성을 이용하여 직접 훈련하는 방식이 채택되었다. 또한 D-JA 방식에서는 Jacobian 행렬을 얻기 위하여 Baum-Welch 알고리듬에 기반한 추정방식을 사용하였다.

$$E\{\boldsymbol{X_t}\} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\boldsymbol{X_t}}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{9}$$

여기서 $\gamma_t(j,k)$는 켑스트럼 특징벡터 $\boldsymbol{X_t}$ 가 HMM의 상태 j의 혼합성분 k에 의해서 발생될 확률을 의미하며 T는 특징벡터의 길이를 나타낸다. 식(9)에 식(8)을 대입하면 기준 HMM의 평균 파라미터 값을 다음식과 같이 얻을 수 있다.

$$E\{\widetilde{\boldsymbol{X}}_{\boldsymbol{t}}\} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)(\boldsymbol{X_t} + \frac{\partial \boldsymbol{X_t}}{\partial \boldsymbol{N_t}}(\boldsymbol{N_t} - \widetilde{\boldsymbol{N}}_{\boldsymbol{t}}))}{\sum_{t=1}^{T} \gamma_t(j,k)} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\boldsymbol{X_t}}{\sum_{t=1}^{T} \gamma_t(j,k)} + \frac{\sum_{t=1}^{T} \gamma_t(j,k)\frac{\partial \boldsymbol{X_t}}{\partial \boldsymbol{N_t}}}{\sum_{t=1}^{T} \gamma_t(j,k)}\Delta \boldsymbol{n} \tag{10}$$

여기서 잡음신호의 차이 $\Delta \boldsymbol{N}(=\boldsymbol{N_t} - \widetilde{\boldsymbol{N}}_{\boldsymbol{t}})$의 값이 시간에 대한 평균치로서 대체가 가능하다고 가정된다.

3. 실험 결과

본 장에서는 Aurora 2 데이터베이스를 이용하여 본문에서 설명된 여러 가지 인식시스템의 형태에 따른 인식성능의 변화를 비교 검토 하였다. Aurora 2 데이터베이스의 인식용 실험 set은 set A (훈련과정에 알려진 4 개의 부가잡음(subway, babble, car, exhibition)음성 으로 이루어짐), set B (미리 알려지지 않은 4 개의 부가잡음(restaurant, street, airport, train-station)음성으로 이루어짐)

그리고 set C(부가잡음 외에도 컨벌루션 잡음을 포함한 1 개의 알려진 잡음(subwayM)음성과 1 개의 미리 알려지지 않은 잡음(streetM)음성으로 이루어짐)로 이루어져 있다. 훈련을 위해서 Aurora 2 데이터베이스에는 clean training 과 multi training을 위한 훈련용 데이터가 따로 제공되는데, 특히, multi training 훈련용 데이터를 이용하는 경우를 MCT(multi-condition training)라 부른다. 이와 대비하여 clean training을 이용한 경우는 향후 CCT(clean-condition training)라 부르기로 하겠다. 한편, 다 모델 기반 구조의 인식시스템을 구성하기 위해서는 각 SNR 별과 잡음종류별로 기준 HMM을 만들어 주어야 하며 이를 위해서는 Aurora 2 데이터베이스에서 제공하는 잡음음성파일생성 프로그램을 이용하여 추가적인 잡음음성신호를 생성하였다.

인식실험에서 사용한 특징은 Aurora project에서 제공한 DSR front-end 2.0 버전을 사용하여 추출하였으며, 13 차의 켑스트럼 계수와 그들의 delta, acceleration 계수를 포함한 39차의 특징벡터들로 이루어져 있다. 연속 숫자음 모델링을 위해서 사용된 HMM의 구조는 Aurora 2 데이터베이스에서 제안된 형태를 따르고 있으며(Pearce, 2000), 자체 개발된 인식엔진을 사용 하였다.

다 모델 기반 구조의 인식시스템에서는 인식과정에서의 잡음신호에 대해서 잡음분류를 하게 되는데 <표 1>에는 잡음의 종류에 대한 분류결과를 나타내고 있다. 비교를 위하여 기존에 사용하던 GMM 방식과 본 논문에서 사용된 KL distance 측정 방식을 함께 나타내었다(정용주, 2007).

표 1. GMM 방식과 KL distance 측정 방식에 근거한 잡음신호의 분류 결과

결과 / 테스트잡음	car(%)		babble(%)		exhibition(%)		subway(%)	
	KL	GMM	KL	GMM	KL	GMM	KL	GMM
car	**98.3**	**99.2**	1.7	0.8	0.0	0.0	0.0	0.0
babble	1.9	0.4	**98.1**	**99.6**	0.0	0.0	0.0	0.0
exhibition	0.0	0.1	0.1	0.0	**99.8**	**99.9**	0.1	0.0
subway	0.0	0.0	0.0	0.0	0.1	0.3	**99.9**	**99.7**

<표 1>의 결과를 보면 기존에 사용되던 GMM 방식이 잡음 종류의 분류 성능 면에서 다소 나은 것을 알 수 있으나, KL distance 추정방식도 매우 높은 인식성능을 보여서 다 모델 기반 구조의 인식시스템에서 잡음분류를 위해서 사용하기에 충분하다고 생각된다.

<표 2>에는 set A 데이터에 대해서 CCT, MCT, PMC(parallel model combination), Multi-model(JA), Multi-model(D-JA), Multi-model(Base) 등을 적용한 인식결과가 나타나 있다. Multi-model(JA) 는 다 모델 기반 구조의 인식시스템에서 JA을 통해서 모델 보상을 한 경우를 의미하며, Multi-model(D-JA)는 D-JA를 통해 모델 보상을 하며, Multi-model(Base)는 다 모델 기반 구조의 인식시스템에서 모델보상을 적용하지 않은 경우를 말한다. <표 2>의 결과를 통해서 우리는 다 모델기반의 구조를 가진 인식기들이 전반적으로 높은 성능을 보임을 알 수 있다. Multi-model(Base)의 경우에는 기존에 모델 보상 방식으로 많이 사용되던 PMC 방식에 비해서 단어인식오류(word error rate)를 50% 감소시키는 것으로 나타났다. 또한 다 모델 기반 구조의 인식기 중에서도 Multi-model(Base)의 경우가 모델 보상을 해준 경우인 Multi-model(JA)나 Multi-model(D-JA) 비해서 오히려 향상된 성능을 보임을 알 수 있다. 이것은 set A의 경우에는 훈련과정을 통해서 이미 잡음의 종류가 알려져 있었으므로, 인식시와 훈련시에 잡음의 차이에 의한 HMM 파라미터 값의 변

이가 심하지 않아서 모델보상 방식이 그리 효과가 없기 때문이기도 하며 한편으론 잡음이 시간적으로 계속 변하는 과정에서, 모델 보상을 위해 추출된 잡음신호의 평균값이 입력 잡음음성에 전반에 걸쳐 나타나는 잡음신호의 통계적 특성을 잘 나타내지 못하여, 이를 근거로 한 모델 보상방식에서 약간의 오류가 발생하기 때문인 것으로 파악된다. 한편, MCT 방식도 매우 좋은 인식성능을 보임을 알 수 있는데, 훈련과정에서 다양한 잡음음성을 포함하여 인식기의 강인성이 매우 높아진 것으로 생각된다. 한편 우리는 <표 2>에서 Multi-model(D-JA)의 경우에 SNR 값과 잡음의 종류를 추정하지 않고 미리 알려진 경우에 대해서도 인식실험을 수행하였는데, 제안된 방식을 이용하여 추정한 경우와 거의 인식성능의 차이가 없음을 알 수 있었다. 또한 기존의 VAD를 이용한 SNR 값 추정과 GMM 기반의 잡음종류를 추정하는 방식에 의한 인식실험에서도 제안된 방식과의 인식성능 차이가 거의 없음을 알 수 있었다. 이는 비록 SNR 값 추정이나 잡음종류의 추정에서 오류가 다소 발생하더라도 D-JA 방식의 모델 보상을 통해서 충분히 보완이 되기 때문이라 생각된다.

표 2. Set A 데이터에 대해서 다양한 모델 보상 방식을 적용한 결과(Word accuracy(%)).

모델보상	set A				
	subway	babble	car	exhibition	평균
CCT	69.34	49.41	55.89	62.70	59.34
PMC	81.02	80.66	75.23	80.28	79.30
MCT	86.27	87.74	85.01	86.63	86.41
Multi-model(Base)	91.28	87.06	90.79	91.91	90.23
Multi-model(JA)	86.25	83.77	83.15	83.06	84.06
Multi-model(D-JA)	90.52	85.26	91.16	91.10	89.51

<표 3>에서는 위의 모델 보상 방식에 대한 결과를 Aurora 2 데이터베이스 전체에 적용한 결과를 나타낸다. 평균 인식율은 관례대로 0.4*A+0.4*B+0.2*C로 나타내었다.

표 3. Aurora 2 데이터베이스 전체에 대해서 다양한 모델 보상 방식을 적용한 결과(Word accuracy(%)).

모델보상	set A	set B	set C	평균
CCT	59.34	55.17	67.53	59.31
PMC	79.30	81.18	78.02	79.79
MCT	86.41	86.78	83.78	86.03
Multi-model(Base)	90.26	85.15	87.92	87.74
Multi-model(JA)	84.06	84.32	86.90	84.73
Multi-model(D-JA)	89.51	86.73	89.21	88.33

<표 2>의 결과와 대체적으로 비슷한 경향이 set B와 set C에 대해서도 나타나는 것으로 보이지만 set B와 set C의 경우에는 Multi-model(D-JA)가 가장 우수한 인식성능을 보임을 알 수 있다. 이것은 set A에서는 잡음의 종류가 훈련과정에서 미리 알려져 모델보상의 효과가 별로 없었지만, set B와 set C의 인식환경은 훈련과정에서 고려 될 수 없으므로 모델 보상이 효력을 발휘한 것으로 보인다.

위의 인식결과에서 우리는 다 모델 기반 구조의 인식기와 더불어 MCT 훈련의 결과가 상당히

우수한 것으로 나타남을 알 수 있다. MCT 훈련을 통해서는 여러 종류의 잡음음성 데이터가 HMM 파라미터에 녹아 있으므로 다양한 종류의 잡음환경에 대해서도 비교적 인식성능이 우수한 것으로 보인다. 특히, MCT를 이용한 인식과정에서는 기준 HMM을 선택하기 위한 SNR과 잡음종류를 추정하는 과정이 따로 필요 없으므로, 이로 인한 오류를 예방할 수 있다는 장점이 있다고 생각된다.

훈련과정의 근본적인 차이로 인하여 MCT 훈련을 통해서 얻은 기준 HMM과 다 모델 기반 구조의 인식기에서 얻어진 기준 HMM이 상호 보완 작용을 할 수 있으리라 생각된다. 우리는 MCT 훈련에서 얻어진 기준 HMM의 확률밀도 함수 f_{MCT}와 Multi-model(D-JA)에서 얻어진 $f_{Multi(D-JA)}$를 상호 결합하여 인식율을 높이고자 하였다. 결합된 확률밀도 함수는 다음과 같다.

$$f_{combine}(\boldsymbol{X}) = \alpha f_{Multi(D-JA)}(\boldsymbol{X}) + (1-\alpha) f_{MCT}(\boldsymbol{X}) \tag{11}$$

<표 4>에는 식(10)의 α 값의 변환에 따른 결합 확률밀도 함수를 사용한 경우의 Aurora 2 데이터베이스에 대한 인식율이 나타나 있다.

표 4. Aurora 2 데이터베이스 전체에 대해서 결합 확률밀도 함수를 적용한 경우의 결과 (Word accuracy(%)).

	set A	set B	set C	평균
$\alpha = 0.3$	90.26	88.71	88.92	89.37
$\alpha = 0.4$	90.32	88.65	88.99	89.38
$\alpha = 0.5$	90.33	88.58	89.04	89.37
$\alpha = 0.6$	90.30	88.51	89.10	89.34
$\alpha = 0.7$	90.30	88.40	89.12	89.30

<표 4>의 결과에서 α 값의 상당한 변화폭에도 불구하고 인식성능은 견조한 흐름을 보여 주고 있으며, 또한 <표 3>의 결과와 비교해 보면 확률밀도함수를 결합하기전인 Multi-model(D-JA)와 MCT 각각의 결과에 비해서도 우수한 성능을 보임을 알 수 있다. 따라서 다 모델기반 구조의 인식 시스템과 MCT는 상호간에 보완하는 특성이 있음을 확인할 수 있었다.

4. 결 론

본 연구에서는 다 모델 기반 구조의 인식기를 Aurora 2 데이터베이스에 대하여 적용하여 연속 잡음음성에 대한 인식성능의 향상을 꾀하였으며 기존의 연구에서 사용하던 VAD 기반의 SNR 추정과 GMM 방식의 잡음종류 분류 방식 대신에 KL distance 기반의 SNR과 잡음종류 동시 분류 방식을 제안하였다. 제안된 방식은 비교적 그 구현이 간단하며 SNR 추정시의 오류 등을 방지할 수 있는 장점이 있었으며, 잡음 종류의 분류 성능에서도 기존의 방식과 비슷한 성능을 나타냄을 알 수 있었다. 또한 다 모델 기반 구조의 인식기에서 D-JA 방식의 모델보상을 적용한 HMM의 확률밀도

함수와 MCT 훈련된 HMM의 확률밀도 함수를 결합하여 상호 보완해 줌으로서 보다 높은 인식성능의 향상을 이룰 수 있었다.

참 고 문 헌

정용주, 곽성우. 2007. “다 모델 방식과 모델보상을 통한 잡음환경 음성인식.” *말소리* 62호, 98-111.

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.,* 41, 164-171.

Boll, S. 1979. “Suppression of acoustic noise in speech using spectral subtraction.” *IEEE Trans. Acoust., Speech, Signal Processing*, 27(2), 113-120.

Ephraim, Y. & Malah, D. 1984. “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator.” *IEEE Trans. on ASSP.* 32(6), 1109-1121.

Gales, M. & Young, S. 1993. “Parallel model combination for speech recognition in noise.” *Tech. Rep. 135*, Cambridge University.

Moreno, P. 1996. “Speech Recognition in Noisy Environments.” *PhD Thesis*, Carnegie Mellon University.

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,* 9(8), 842-855.

Juang, B. H. & Rabiner, L. R. 1984. “A probabilistic distance measure for hidden Markov models.” *AT&T Technology Journal,* 391-408.

Pearce, D. & Hirsch, H. 2000. “The Aurora experimental framework for the performance evaluation of speech recognition systems under conditions.” *Proc. ICSLP 2000,* IV, 29-32.

Sagayama, S., Yamaguchi, Y. & Takahashi, S. 1997. “Jacobian adaptation of noisy speech models.” *IEEE Workshop on Automatic Speech Recognition and Understanding*, 396-403.

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.*, 977-980.

접수일자: 2008. 4. 30
게재결정: 2008. 6. 9

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