자동차 잡음환경 고립단어 음성인식에서의 VTS와 PMC의 성능비교*

Performance Comparison between the PMC and VTS Method for the Isolated Speech Recognition in Car Noise Environments

정 용 주** · 이 승 욱***
Yong-Joo Chung · Seung-wook Lee

ABSTRACT

There has been many research efforts to overcome the problems of speech recognition in noisy conditions. Among the noise-robust speech recognition methods, model-based adaptation approaches have been shown quite effective. Particularly, the PMC (parallel model combination) method is very popular and has been shown to give considerably improved recognition results compared with the conventional methods. In this paper, we experimented with the VTS (vector Taylor series) algorithm which is also based on the model parameter transformation but has not attracted much interests of the researchers in this area. To verify the effectiveness of it, we employed the algorithm in the continuous density HMM (Hidden Markov Model). We compared the performance of the VTS algorithm with the PMC method and could see that the it gave better results than the PMC method.

Keywords: Speech Recognition, Model-based Compensation, PMC, VTS

1. 서 론

최근에 음성인식 기술의 꾸준한 향상으로 인하여 음성인식 제품의 실용화가 점차 이루어지는 것을 볼 수 있다. 하지만, 아직까지 음성인식 제품들이 좀 더 많은 이용자들의 관심을 끌기 위해서는 해결해야 하는 문제가 많은 것도 사실이다. 그 중에서도 자동차 환경에서 발생하는 잡음이나 전화선로/마이크로폰 등과 같은 전송채널에 의한 음성신호 왜곡 현상으로 인한 인식성능의 저하 문제는 아직까지 완전히 해결하지 못한 분야 중의 하나이다. 이와 같은 음성신호의 왜곡에 의한 인식성능의 저하를 방지하기 위한 연구는 지금까지 다양하게 이루어져 왔다. CMN(Cepstral mean normalization)과 같은 방식에서는 음성신호왜곡을 보완할 수 있는 음성특징을 이용하여 좋은 결과를 이룰 수 있었다[1][2]. 또한, 왜곡된 음성신호를 개선하는 방법에 대해서도 많은 연구가 이루어졌다[3][4][5]. 이러한 방법 중에서 가장 대표적인

* 본 연구는 2002년도 계명대학교 비사연구기금으로 이루어졌음.
** 계명대학교 전자공학과
*** 동명전자(주)

방법은 주파수 차감법(Spectral subtraction)이다. 여기서는 잡음의 스펙트럼을 추정한 후, 이를 잡음음성신호의 스펙트럼에서 차감하여 원래 음성의 스펙트럼을 얻고자 한다. 이 방법은 정확한 잡음 스펙트럼의 추정이 필요하다는 단점이 있다. 따라서, 추정된 잡음스펙트럼의 오차를 보상하기 위한 다양한 방식이 제안되었다.

최근에 이르러서는 앞에서의 방식에 비해서 좀 더 획기적으로 잡음음성의 인식성능을 향상시키고자 하는 방법들이 제시되었다. 이러한 방식은 기본적으로 음성인식을 위한 HMM (hidden Markov model)의 모델 파라미터의 변형을 통해서 이루어지므로, 모델기반 방식이라고 불리워진다[6][7]. 이러한 방식 중에서 가장 널리 알려져 있는 방식은 PMC와 VTS이다. 이 두 방식은 서로 다른 연구논문에서 각각 제시되어 연구결과가 발표된 경우는 많았으나, 동일한 조건 하에서의 상호 인식성능을 자세히 비교한 경우는 국내외적으로 발표된 경우가 별로 찾아볼 수 없었다. 따라서 본 논문에서는 동일한 조건 하에서 VTS 방식과 PMC 방식을 구현하여, 상호 인식결과를 비교하고, 그 결과에 대해서 논의 하고자 한다. 본 논문의 구성은 2 장에서 각각 VTS 방식과 PMC 방식에 대해서 간략히 소개하고 3 장에서 2 가지 방식에 의한 연구결과를 소개하며 마지막으로 4 장에서 결론을 맺고자 한다.

2. 모델변환 방식(VTS/PMC)의 개요

잡음음성인식을 위한 모델변환 방식에서는 잡음음성 생성의 원인을 가정하고 수학적으로 분석한 후 이를 HMM의 각 상태별의 가우시안 분포를 나타내는 평균값과 분산값에 적절히 반영되도록 한다. 이를 위해서는 음성신호 $\boldsymbol{x}$가 잡음 $\boldsymbol{n}$, $\boldsymbol{h}$에 의해서 어떤 영향을 받는가를 적절히 분석해야 한다. 일반적으로 잡음음성 $\boldsymbol{y}$의 생성원인을 파악하기 위해서 가정되는 구체적인 상황은 그림 1과 같다.

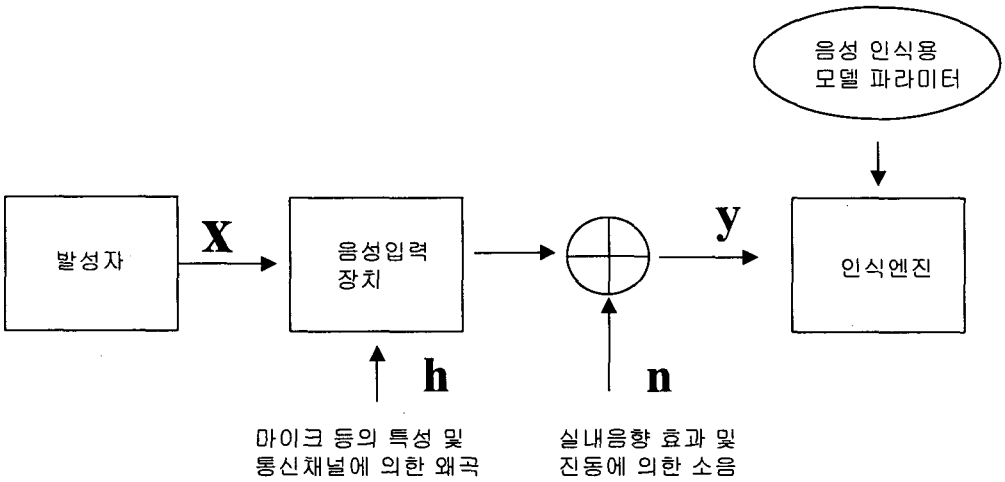

그림 1. 잡음음성 생성에 관한 개요도

모델변환 방식 중 PMC 방식에서는 이러한 잡음신호가 HMM의 파라미터 값에 미치는 영향을 분석하기 위해서 잡음음성 $\boldsymbol{y}$로부터 추출되는 cepstrum 특징 벡터의 확률분포를 해석적으로 유도하였으며 이를 위해서 잡음음성의 생성과정을 설명하는 불일치함수를 사용하였다[6]. VTS 방식에서는 잡음음성의 생성 함수를 Taylor 전개를 이용한 근사화함으로서, 로그스펙트럼영역에서의 특징벡터의 확률분포를 유도하였다. 다음에서는 이 두 가지 방식에 대해서 좀 더 상세히 소개하고자 한다.

2.1 PMC 방식의 개요

PMC 방식에서는 잡음이 섞인 음성(noisy speech)에 대한 HMM모델을 만들기 위해서 원래의 깨끗한 음성(clean speech)에 대한 HMM 파라미터 값과 잡음(noise)에 대한 HMM 파라미터 값을 선형주파수 영역에서 서로 결합하여 준다. 이를 위한 자세한 과정은 다음과 같이 요약된다.

1) 로그스펙트럼 영역의 HMM 파라미터 값을 구하기 위해서 역DCT (inverse discrete cosine transformation) 변환을 취한다[8].

$$\mu^{l} = C^{-1}\mu^{c} \quad , \quad \Sigma^{l} = C^{-1}\Sigma^{c}(C^{-1})^{T} \tag{1}$$

여기서 $\mu^{l} = E(x)$과 $\Sigma^{l} = E((x-\mu^{l})(x-\mu^{l})')$는 로그 스펙트럼 영역에서의 평균벡터와 공분산 행렬이며 μ^{c}와 Σ^{c}는 켑스트럼 영역에서의 그 값이다.

2) 로그스펙트럼 영역의 HMM 파라미터 값을 선형영역으로 변환한다.

$$\mu_i = \exp\left(\mu_i^{l} + \frac{\Sigma_{ii}^{l}}{2}\right), \quad \Sigma_{ij} = \mu_i \mu_j [\exp(\Sigma_{ij}^{l}) - 1] \tag{2}$$

여기서 μ_i와 Σ_{ij}는 선형영역에서의 평균벡터 μ와 공분산행렬 Σ의 구성 원소이다. 위식을 도출하기 위해서는, 로그스펙트럼특징 벡터가 가우시안 분포를 가진다는 가정이 필요하다.

3) 음성과 잡음의 HMM 파라미터 값을 결합한다.

$$\hat{\mu} = \mu + \bar{\mu} \quad , \quad \hat{\Sigma} = \Sigma + \bar{\Sigma} \tag{3}$$

여기서 $\hat{\mu}$와 $\hat{\Sigma}$는 선형영역에서의 잡음음성의 평균벡터와 공분산 행렬이고, $\bar{\mu}$와 $\bar{\Sigma}$는 잡음에 관한 것이다.

4) 결합된 HMM 파라미터 값에 대해서 다음과 같이 로그변환과 DCT 변환을 취함으로서

최종적으로 켑스트럼 영역에서의 잡음음성의 평균벡터 $\hat{\mu}^c$와 공분산행렬 $\hat{\Sigma}^c$이 얻어진다.

$$\hat{\mu}_i^l = \log(\hat{\mu}_i) - \frac{1}{2}\log(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2}+1), \quad \hat{\Sigma}_{ij}^l = \log(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i \hat{\mu}_j}+1) \tag{4}$$

$$\hat{\mu}^c = C\ \hat{\mu}^l, \quad \hat{\Sigma}^c = C\ \hat{\Sigma}^l\ C^T \tag{5}$$

Log-add PMC 방식은 위의 식(2), (4)에서 잡음음성 HMM 평균벡터 값을 구하는 과정에서 잡음과 음성신호 HMM의 공분산 값이 충분히 작다는 가정을 함으로서 변환공식이 다음과 같이 단순화되도록 한다.

$$\hat{\mu}_i^l = \log(\exp(\mu_i^l) + \exp(\bar{\mu}_i^l)) \tag{6}$$

이와 같이 로그영역의 평균벡터값을 구한 다음 위의 식 (6)을 이용하여 켑스트럼 영역의 평균벡터를 구함으로서 원래의 정상적인 방법인 log-normal 방식에 비해서 훨씬 간단하게 계산과정을 마칠 수 있다.

다음에는 본 논문에서 주로 관심 있게 다룬 VTS 방식에 대해서 소개한다.

2.2. VTS 방식의 이론

먼저, 잡음음성이 생성되는 생성함수를 가정한다. 깨끗한 음성을 나타내는 로그스펙트럼상의 특징벡터 $\boldsymbol{x}$와 잡음신호에 의해서 왜곡된 음성신호의 로그스펙트럼 벡터 $\boldsymbol{y}$는 다음과 같은 관계를 가진다고 가정되어진다.

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{a}) \tag{7}$$

여기서, 함수 $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{a})$는 잡음음성의 생성을 나타내는데, 이를 구체적으로 얻기 위해서는 생성과정에 대한 구체적인 설명이 필요하다. 여기서 $\boldsymbol{a}$는 생성과정을 나타내는 환경변수이다. 즉, 그림 1과 같은 잡음생성 과정에서는 다음과 같이 DFT (Discrete Fourier Transform)을 이용한 잡음음성의 스펙트럼 $Y(k)$값을 얻을 수 있다.

$$Y[k] = X[k]\,|H[k]|^2 + M[k] \tag{8}$$

여기서, $X[k]$는 원래음성의 스펙트럼이고 $|H[k]|^2$는 음성 채널에 대한 스펙트럼을 나타낸다. $M[k]$는 잡음에 대한 스펙트럼을 나타내며, k는 DFT에서의 주파수 성분중의 하나를 나타낸다. 위의 수식에서 양변에 log 함수를 취하면 다음과 같이, 식 (7)에서 언급된 함수 $\boldsymbol{g}()$를 구할 수 있을 것이다.

$$g(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}) = \boldsymbol{h} + \log(\boldsymbol{i} + \exp(\boldsymbol{n} - \boldsymbol{x} - \boldsymbol{h})) \tag{9}$$

여기서,

$$\begin{aligned} \boldsymbol{h} &= \log(|H[k]|^2) \\ \boldsymbol{n} &= \log(N[k]) \\ \boldsymbol{x} &= \log(x[k]) \end{aligned} \tag{10}$$

이며, $\boldsymbol{i}$는 모든 원소의 값이 1 인 단위벡터이다. 한편 각 벡터는 차원이 L이라고 가정한다.

위와 같이 로그스펙트럼의 영역에서 잡음 $\boldsymbol{n}$ 및 원래의 음성 $\boldsymbol{x}$의 특징벡터를 이용하여 잡음음성의 특징벡터 $\boldsymbol{y}$를 수식으로 표현할 수 있었는데, 이로부터 잡음음성 $\boldsymbol{y}$에 대한 확률분포를 구하는 것이 필요하다. 하지만, $\boldsymbol{y}$에 대한 확률분포는 우리가 일반적으로 기대하는 가운시안 분포가 아님이 수식 상으로 분명히 알 수가 있다. 이것은 식 (7)의 함수 $\boldsymbol{g}()$가 식 (9)에서 보다시피, 잡음과 음성신호에 대한 다소 복잡한 함수 형태를 가지고 있기 때문이다. 이와 같은 문제를 해결하기 위해서 VTS에서는 함수 $\boldsymbol{g}()$를 Taylor series 전개를 이용하여 근사화하며, 이를 이용하여 $\boldsymbol{y}$에 대한 확률분포를 얻을 수 있다.

2.2. VTS 방식을 이용한 잡음음성인식

단계 1: 잡음음성 $\boldsymbol{y} = \boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_T$가 인식실험을 위해서 주어져 있고, 이에 대한 원래의 깨끗한 음성 $\boldsymbol{x}_t$ ($t = 1, \cdots, T$)에 대한 확률분포를 다음과 같이 가정한다.

$$p(\boldsymbol{x}_t) = \sum_{k=0}^{K-1} w_k N(\boldsymbol{x}_t, \mu_k, \Sigma_k) \tag{11}$$

그리고 잡음과 음성신호 그리고 채널스펙트럼에 대해서 초기값을 임의로 가정한다.

단계 2: 함수 $\boldsymbol{g}()$를 1차의 Taylor series를 이용하여 전개한 후, 잡음음성 $\boldsymbol{y}$에 대한 평균벡터와 분산행렬을 구한다. 함수 $\boldsymbol{g}()$에 대한 Taylor 전개식은 다음과 같다.

$$\begin{aligned} \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h}) &= \boldsymbol{x} + \boldsymbol{g}(\mu_x, \boldsymbol{n}_o, \boldsymbol{h}_0) + \nabla_x \boldsymbol{g}(\mu_x, \boldsymbol{n}_o, \boldsymbol{h}_0)(\boldsymbol{x} - \mu_x) \\ &+ \nabla_n \boldsymbol{g}(\mu_x, \boldsymbol{n}_o, \boldsymbol{h}_0)(\boldsymbol{n} - \boldsymbol{n}_0) + \nabla_h \boldsymbol{g}(\mu_x, \boldsymbol{n}_o, \boldsymbol{h}_0)(\boldsymbol{h} - \boldsymbol{h}_0) \end{aligned} \tag{12}$$

여기서, $\boldsymbol{n}_0$, $\boldsymbol{h}_0$, μ_x는 단계 1에서 정한 각각 잡음과 채널스펙트럼 그리고 음성신호의 초기값이며, Taylor 전개식의 기준점이 된다. 한편, 위의 전개식에 근거한 잡음음성의 각 mixture 별 평균벡터 $\mu_{k,y}$와 공분산 행렬 $\Sigma_{k,y}$은 아래와 같다.

$$\mu_{k,y} = \mu_{k,x} + \nabla_h g(\mu_{k,x}, n_o, h_0)(h - h_0) + \nabla_n g(\mu_{k,x}, n_o, h_0) \quad (13)$$
$$(n - n_0) + g(\mu_{k,x}, n_o, h_0)$$

$$\Sigma_{k,y} = (I + \nabla_x g(\mu_{k,x}, n_o, h_0))\ \Sigma_{k,x}(I + \nabla_x g(\mu_{k,x}, n_o, h_0))^T \quad (14)$$

단계 3: 주어진 잡음음성 y와 식 (13)과 식 (14)에서 얻어진 각 mixture 별로의 평균값과 분산값을 이용하여, 식 (15)의 log-likelihood 함수값을 최대로 하는 잡음 벡터 n과 채널벡터 h를 구한다. 이때 EM (Estimate-Maximize) 방식을 채용한다[10].

$$L(y = y_0, y_1, \cdots, y_T) = \sum_{t=0}^{T} \log(p(y_t | h, n)) \quad (15)$$

단계 4: 단계 3에서 구한 n과 h를 초기값 n_0 및 h_0와 비교하여 수렴이 되었는지 확인한다. 수렴이 되면, 단계 5로 넘어가고 그렇지 않으면, 단계 2로 가서 반복 수행한다. 이때, 단계 2 수행 전에 n과 h 값을 이용하여 n_0와 h_0값을 대체한다.

단계 5: 위의 과정에서 얻어진 잡음음성에 대한 평균벡터 값과 공분산 행렬 값을 이용하여, MMSE(Minimum Mean Squared Error) 방식에 근거하여, 원래의 음성벡터 x를 잡음음성 벡터 y로부터 추정한다. 구체적인 관계식은 다음과 같다.

$$x_{MMSE} = E(x|y) = \int_x x\, p(x|y)dx = y - \int_x g(x,n,h)\, p(x|y)dx \quad (16)$$
$$= y - \sum_{k=0}^{K-1} P[k|y]\ g(\mu_{k,x}, n, h)$$

위의 식에서 새로이 추정된 음성벡터 x_{MMSE}를 이용하여 인식실험을 수행한다. 이때 사용되는 HMM 파라미터는 기존의 깨끗한 음성에서 얻어진 것을 사용한다. 전체 과정을 종료한다. 위의 마지막 단계에서 우리는 새로운 특징벡터 x_{MMSE}를 반드시 구할 필요는 없을 것이다. 대신에 식 (13)과 식 (14)을 이용하여, 변형된 HMM 파라미터를 구하고 이를 이용하여, 주어진 잡음음성 y에 대한 인식 수행을 할 수도 있을 것이다. 그러나 이러한 경우에는 시간차분(delta) 음성특징에 대한 파라미터 변환 값을 구하기 어렵다는 문제점이 있다.

2.3. 연속밀도 HMM 에서의 VTS 방식의 적용

원래의 VTS 방식은 semi-continuous HMM에서 구현이 이루어졌으나, 본 연구에서는 연속밀도 HMM에서 VTS 알고리듬을 적용하는 방식을 취하였다. VTS 방식에서의 가장 기본적인 가정은 식 (11)에서와 같이 음성신호의 분포를 가우시안 함수의 mixture로 가정하는 것이다. 이때 중요한 것은 가운시안 분포를 선택하는 문제이다. 원래 VTS 방식은 semi-continuous HMM에서 제안되었으므로 가운시안 분포는 미리 정해진 코드북으로 이루어졌다. 예를 들면,

코드북의 코드워드 개수가 256이면, 음성신호를 구성하는 가운시안 분포는 256 개로 이루어진다. 연속밀도 HMM에서도 이와 같이 코드북을 설계하여, 이를 이용한 가우시안 분포를 가정할 수 있다. 하지만, 연속밀도 HMM에서의 각 상태별로 가운시안 분포가 주어져 있으므로, 이를 이용하면 새로운 코드북의 구성 없이도 VTS 방식을 적용할 수 있을 것이다. 이때 식 (11)의 수식에서 의미하는 mixture의 개수 K는 전체 HMM의 각 상태의 mixture의 개수를 합한 것과 같다.

한편 VTS 방식은 unsupervised 적응 형태로 적용이 가능하다. unsupervised 방식에서는 음성신호의 분포를 모든 가능한 가운시안 분포들의 결합으로 근사화할 수 있다. 다시 말해서, 모든 HMM의 각 상태별 가운시안 분포를 동시에 전부다 고려하는 것이다. 즉, 식 (11)는 다음과 같이 변형된다.

$$p(\boldsymbol{x}_t) = \sum_{p=0}^{P-1} \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} w_{k,s,p} N(\boldsymbol{x}_t, \mu_{k,s,p}, \Sigma_{k,s,p}) \tag{17}$$

여기서 P, S는 HMM에 의해서 모델링되는 기본 모델단위인 음소의 개수와 각각의 음소모델에 존재하는 상태의 개수가 된다.

기존의 VTS 알고리듬에서는 특징벡터로서 로그스펙트럼값을 사용하였는데, 일반적으로 최근의 음성인식시스템에서는 켑스트럼을 많이 사용하고 있다. 따라서 본 연구에서는 일단, VTS 알고리듬을 로그스펙트럼에 적용한 후, 이로부터 켑스트럼 벡터를 얻은 경우에 대해서도 인식실험을 수행하고자 한다. 켑스트럼 벡터를 얻는 방법은 로그스펙트럼으로부터 아래와 같은 변환식을 이용한다.

$$\boldsymbol{x}^c_{MMSE} = \boldsymbol{C}\, \boldsymbol{x}_{MMSE} \tag{18}$$

여기서 행렬 C는 DCT (Discrete cosine transformation) 변환을 나타낸다.

3. 고립단어 인식을 통한 성능개선 결과

3.1 기반 인식시스템의 개요

본 실험에서 사용된 기반인식기(baseline recognizer)는 연속밀도 HMM으로 구성되어 있으며, 32 개의 PLU (phoneme like unit)을 기본 구성 단위로 하였다. 또한 각각의 HMM은 단순한 left-right 연결 형태로 결합되어 있는 3 개의 상태(state)로 이루어져 있다. 인식실험시 사용된 데이터베이스는 한국과학기술원에서 제공한 75 개의 고립단어들로 이루어져 있으며 이들 단어는 음향학적으로 고르게 분포되도록 선정되어져 있다. 전체 80 명분의 음성데이터베이스 중 학습을 위하여 60 명의 화자가 이용되었으며 인식 실험을 위해서는 학습에 참여하지 않은 20 명을 택하였다. 또한, 4 회의 반복실험을 통해서 매번 훈련화자그룹과 인식화자그룹을 달리하여서 인식결과의 신뢰도를 높였다. 각 화자는 75 개의 단어를 1 회씩 조용한 사

무실 환경에서 발성하였고 이 음성데이터는 16 khz, 16 bit로 A/D 변환되었다. 실제 환경에서 발생하는 잡음음성에 대한 실험을 위하여 실제로 자동차 내에서 발생하는 잡음을 녹음하여 A/D 변환한 것을 사용하였다. 특징벡터로서는 VTS 알고리듬의 적응을 위해서는 18 차의 멜-스케일(mel-scale)의 로그스펙트럼을 사용하였으며, PMC 알고리듬을 위해서는 13 차의 MFCC (mel-frequency cepstrum coefficients)을 사용하였다. 또한 로그스펙트럼과 MFCC의 각각에 대해서 regression 계수를 산정하는 방식을 이용한 차분특징(delta features)을 부가적으로 사용하여 전체의 특징벡터의 계수는 각각 36 차와 26 차가 되도록 하였다.

3.2. 기반인식시스템의 인식성능

본 절에서는 먼저 기반인식기의 성능에 대해서 검토하였다. 표1에서 우리는 기반인식기의 인식 성능을 자동차 잡음 환경 하에서의 신호대 잡음비(SNR, signal to noise ratio)가 달라짐에 따라서 어떻게 변화하는지 나타내었다.

표 1. 자동차 잡음환경에서 SNR값이 변함에 따른 기반인식기의 인식률(%)의 변화(괄호 안의 수는 차분특징벡터를 사용한 경우의 인식률)

인식환경 / 음성특징	0 dB	10 dB	20 dB	CLEAN
mel-scale 로그스펙트럼	17.8(31.3)	54.1(76.4)	80.5(92.7)	93.3(97.2)
MFCC	32.5(55.7)	71.3(89.9)	88.7(94.6)	94.7(98.3)

위의 기반 인식기의 성능을 통해서 우리는 잡음의 세기가 강해질수록 인식률의 저하가 매우 심해짐을 알 수 있었다. 기반인식기는 잡음의 영향을 받지 않은 원래의 깨끗한 음성을 이용하여 Baum-Welch 알고리듬을 이용하여 학습되었으며 잡음에 대한 영향을 고려하기 위한 어떠한 보상 작업도 하지 않았다. 표 1의 결과에서, 입력음성의 SNR이 20 dB 이상인 경우는 인식률의 저하가 그리 심하지는 않으나, SNR이 10 dB 이하인 경우는 매우 심각한 정도의 인식률의 저하를 가져옴을 알 수 있다. 따라서 SNR이 매우 낮은 영역에서의 인식률을 향상시키는 방법은 잡음음성인식에서의 성공적 수행을 위해서 매우 중요하다 할 것이다. 또한, 두 가지 서로 다른 종류의 특징벡터를 이용한 실험결과 mel-scale의 로그스펙트럼에 비해서 DCT 변환을 취한 후의 값인 MFCC를 이용한 경우에 인식률이 더 나은 것을 알 수 있었다.

표 2에서는 인식과 훈련시의 환경이 동일한 경우(동일한 SNR 값)의 인식실험 결과를 나타내고 있다. 이 경우에는 표 1과 대비하여 인식성능이 많이 향상됨을 알 수 있다. 이것은 이미 훈련을 통해서 잡음의 영향이 HMM의 파라미터들의 값에 충분히 반영되기 때문이라고 생각된다. 따라서 이 결과는 잡음보상 인식알고리듬의 개발에 있어서 중요한 벤치마크 결과라고 생각된다. 그러나, 이러한 인식성능을 얻기 위해서는 항상 변화하는 잡음환경에 대한 충분한 양의 음성 데이터베이스를 가지고 모델을 미리 훈련시켜야 하는 어려움이 있다.

한편, 로그스펙트럼과 MFCC의 두 가지의 경우 각각에 대해서 표 1과 인식률을 비교해 보면, 로그스펙트럼의 경우에 인식률의 향상이 더 많이 이루어짐을 알 수 있었다. 이것은 로그

스펙트럼 특징의 특성이 잡음에 대해서 MFCC보다 강인성이 많이 떨어진다 하더라도, 이를 재훈련을 통해서 충분히 보상해 줄 경우, 두 가지 특징이 나타내는 인식성능에는 큰 차이가 없다는 것을 말해준다.

표 2. 인식과 훈련시에 환경이 동일한 경우(동일한 SNR값)의 기반인식기의 인식률(%) 비교 (괄호 안의 수는 차분특징벡터를 사용한 경우의 인식률)

인식환경 / 음성특징	0 dB	10 dB	20 dB	CLEAN
mel-scale 로그스펙트럼	82.9(89.4)	90.5(95.5)	92.5(97.1)	93.3(97.2)
MFCC	84.4(89.1)	91.8(95.6)	94.3(97.5)	94.7(98.3)

3.3. 모델변환 방식을 이용한 잡음음성인식 실험

VTS 방식은 기본적으로 모델변환방식을 이용한 알고리듬이나, 여기서 머물지 않고 변환된 모델을 이용하여, 음성특징인 로그스펙트럼값의 변환을 시도한다. 이렇게 함으로서 단순히 정적인(static) 특징벡터 뿐만 아니라, 동적인 특징벡터에 대해서도 쉽게 적용할 수 있는 장점이 있다. 표 3에서는 VTS 방식에서, 특징벡터 변환을 한 경우와 특징벡터변환을 거치지 않고 단순히 HMM의 모델변환만을 한 경우의 인식성능을 비교하고 있다. 여기서는 두 가지 방식에 대한 비교의 단순화를 위하여 차분특징 벡터는 사용하지 않은 인식결과를 보여주고 있다. 또한, 비교를 위해서 log-add PMC 방식의 결과도 함께 나타내었다.

표 3. VTS 방식에서 특징벡터변환을 수행한 경우와 단순히 모델파라미터 변환만을 수행한 경우의 인식률(%) 비교(차분특징벡터는 사용하지 않음)

적응방법 / 인식환경	특징벡터 변환시	모델 파라미터 변환시	PMC (log-add)
0 dB	84.4	81.2	82.7
10 dB	91.3	88.7	90.6
20 dB	93.2	91.8	93.7

위의 결과에서 알 수 있듯이, VTS 방식에서는 단순히 모델파라미터의 변환만을 시도하는 것보다는 이를 이용하여, 로그스펙트럼값을 변환해주는 것이 더욱 인식성능을 높이고 있음을 알 수 있다. 이러한 원인은 아마도 모델파라미터의 변환시 발생하는 일부분의 에러는 인식률에 직접적으로 나쁜 영향을 미치지만, 특징벡터 변환을 할 경우에는 이러한 일부분의 영향이 바로 나타나지 않고 전체적인 방향성이 나타나므로 보다 강인한 인식성능을 나타내기 때문이라 생각된다. 또한, PMC 방식과 비교했을 경우, 모델파라미터만 변환할 경우는 PMC 방식보다도 성능이 떨어졌으나, 특징벡터의 변환을 한 경우는 오히려 PMC 방식보다도 인식성능이 향상됨을 알 수 있었는데, 이것도 PMC 방식이 모델파라미터만 변화시키므로 특징벡터 변환 VTS 방식에 비해서 강인성이 떨어지기 때문이라 생각된다.

아래의 표 4에서는 VTS 방식과 PMC 방식을 이용한 잡음음성인식의 성능을 보여주고 있다. PMC 방식에서는 log-normal 방식과 log-add 방식을 이용한 결과를 나타내었으며[9], VTS 방식에서는 로그스펙트럼을 재 추정한 경우와 이로부터 켑스트럼을 재 추정한 경우에 대해서 각각 인식률을 나타내었다. 인식실험을 위해서 정적특징 외에도 차분특징벡터를 사용하였다. 이 결과에서 우리는 VTS 방식이 PMC 방식에 비해서 전반적으로 향상된 결과를 보여주는 것을 알 수 있었다. 특히, 기존의 VTS에서와 같이 로그스펙트럼을 특징벡터로 사용한 경우는 PMC와 비교해서 인식성능의 향상을 뚜렷이 볼 수 없었으나, 로그스펙트럼으로부터 유도된 켑스트럼을 사용한 경우에 인식율이 많이 향상되는 것을 볼 수 있었다. 이러한 결과는 기존의 음성인식시스템에서 캡스트럼 값이 음성특징벡터로서 많이 사용되는 것과 일치하는 것이다. 특히, 켑스트럼 특징를 이용한 경우의 VTS 결과를 보면, 표 2에서 보여준 재훈련의 결과보다도 오히려 더 우수하게 나타나는데, 이것은 VTS 방식이 순수하게 unsupervised 방식이라는 것을 감안하면 상당히 괄목할 만한 인식률의 향상이라고 보여진다.

표 4. 잡음음성인식 환경에서 VTS와 PMC 방식의 인식률(%) 성능비교

	VTS		PMC	
	로그스펙트럼	켑스트럼	log-normal	log-add
0 dB	86.4	90.4	86.9	87.1
10 dB	94.9	96.4	93.7	93.8
20 dB	96.7	97.8	96.8	96.8

4. 결 론

본 논문에서는 잡음음성인식을 위한 HMM 모델파라미터 기반의 적응기법으로서 최근에 많은 주목을 받고 있는 PMC 방식과 VTS 방식의 성능을 동일한 조건 하에서 비교하였다. VTS 방식은 알고리듬의 구현이 다소 번거롭고 실시간 수행시에 계산량이 많다는 단점 때문에 PMC 방식에 비해서 다소 관심을 적게 받았다. 하지만, 본 논문에서는 연속분포 HMM을 이용한 인식실험에서 VTS 방식과 PMC 방식을 구현하여 인식실험 한 결과, 자동차 잡음환경하에서, VTS 방식이 PMC 방식에 비해서 인식성능이 전반적으로 우수함을 확인할 수 있었다. 특히, VTS에서 단순히 모델파라미터만을 변환시키는 경우보다는 음성특징을 직접 변환시키므로서 많은 성능의 향상이 이루어짐을 알 수 있었으며, 음성특징으로서 기존에 VTS에서 사용되던 로그스펙트럼 대신에 켑스트럼을 사용함으로서 재훈련에 의한 경우보다도 인식성능이 나아짐을 확인할 수 있었다.

참 고 문 헌

[1] Liu, F. et al. 1993. "Efficient cepstral normalization for robust speech recognition." *Proceedings of ARPA Human Language Technology Workshop.*

[2] Hermansky, H. 1994. "RASTA processing of speech." *IEEE Trans. on Speech and Audio Processing*, Vol. 2.

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

[4] Lockwood, P. & J. Boudy. 1991. "Experiments with a nonlinear spectral subtraction: Hidden Markov models and the projection for robust speech recognition in cars." *Eurospeech.*

[5] Alejandro, Acero. 1993. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers.

[6] Gales, M. & S. Young. 1993. "Parallel model combination for speech recognition in noise." *Tech. Rep.*, 135, Cambridge University.

[7] Moreno, P. 1996. *Speech Recognition in Noisy Environments.* Ph.D Thesis, Carnegie Mellon University.

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

[9] 정용주. 2003. "An efficient model parameter compensation method for robust speech recognition." *말소리*, 45.

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

접수일자: 2003. 7. 21.
게재결정: 2003. 8. 31.

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

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