Bayesian 적응 방식을 이용한 잡음음성 인식에 관한 연구

A Study on Noisy Speech Recognition Using a Bayesian Adaptation Method

정 용 주*
(Yong-Joo Chung*)

*계명대학교 컴퓨터 전자공학부
(접수일자: 2000년 11월 15일; 채택일자: 2001년 1월 12일)

본 논문에서는 잡음에 강인한 음성인식을 위해서 expectation-maximization (EM) 방식을 이용하여 잡음의 평균값을 추정하는 새로운 알고리듬을 제안하였다. 제안된 알고리듬에서는 온라인상의 인식용 음성이 직접 Bayesian 적응을 위해서 사용되며, 또한 훈련데이터를 이용하여 잡음의 평균값에 대한 사전 (prior) 분포를 알아낸 후 Bayesian 적응시에 이용한다. 잡음음성의 모델링을 위해서는 PMC (parallel model combination) 방식을 이용하였고, 제안된 방식을 이용하여 자동차 잡음 환경하에서 인식 실험을 수행한 결과, 기존의 PMC 방식에 비해서 향상된 인식성능을 보임을 알 수 있었다.
핵심용어: 잡음 음성인식 (noisy speech recognition), Bayesian 적응, EM 알고리듬
투고분야: 음성처리 분야 (2.5, 2.6)

An expectation-maximization (EM) based Bayesian adaptation method for the mean of noise is proposed for noise-robust speech recognition. In the algorithm, the on-line testing utterances are used for the unsupervised Bayesian adaptation and the prior distribution of the noise mean is estimated using the off-line training data. For the noisy speech modeling, the parallel model combination (PMC) method is employed. The proposed method has shown to be effective compared with the conventional PMC method for the speech recognition experiments in a car-noise condition.
Key words: *Noisy speech recognition, Bayesian adaptation, EM algorithm*
Subject classification: *Speech signal processing* (2.5, 2.6)

I. 서 론

최근에 들어와서는 음성인식시스템의 성능이 많이 향상되어서 몇몇의 상용제품도 볼 수 있게 되었다. 그러나, 실제 환경에서는 음성인식 제품이 좀 더 나은 성능을 보이기 위해서 반드시 극복해야 할 몇가지 중요한 문제들이 있다. 그들 중에 하나가 잡음에 대한 강인성이다.

특히, 움직이는 자동차 내부에는 항상 엔진 소음 등에 의한 부가잡음이 발생하며 이것이 음성을 방해함으로 인해서 훈련된 모델과 그에 해당하는 실제 잡음음성 간에는 차이가 발생하게 된다. 이러한 차이는 결국 음성인식 시스템의 인식 오류를 증가시키게 될 것이다. 이러한 차이를 보정하기 위한 다양한 방법들이 제안되어 왔다. 최근에는 HMM (hidden Markov model)에 근거한 잡음음성 인식을 위하여, 모델 파라미터 적응 방식들이 제안되어서 성공적인 결과들을 보여주고 있다. 그러한 방법들에서는 HMM 파라미터들이 주로 ML (maximum likelihood) 방식에 의해서 보정되었다[1].

Bayesian 적응 방식은 HMM을 이용한 음성인식에서 많이 이용되어 왔다. 이 방식은 추정하고자 하는 파라미터들에 관한 정보를 미리 알고 있을 경우에 비교적 적은 양의 데이터를 이용하여 주어진 파라미터를 추정하는데 있어서 효과적임이 입증되었다. 특히, Bayesian 방식을 이용한 화자적응은 많은 효과를 보임을 알 수 있었는데[2] 이러한 접근 방식은 잡음 모델 파라미터를 추정하는데에도 적용할 수 있으리라 생각된다.

PMC 방식은 잡음음성 인식을 위한 모델 파라미터 적응방법의 하나로서 인식성능의 향상에 성공적으로 적용되어 왔다[3]. 이 방식에서는 잡음음성을 위한 HMM을 구성하기 위해서 원래의 음성 HMM과 잡음의 HMM을 단순히 결합한다. 그러나, 이 방식에서는 잡음 HMM의 파라미터를 추정하기 위해서 입력음성으로부터의 묵음 구간을 추출할 필요가 있다. 그러나 입력음성으로부터 추

책임저자: 정용주 (yjjung@kmu.ac.kr)
704-701 대구시 달서구 신당동 1000번지
계명대학교 컴퓨터 전자공학부
(전화: 053-580-5925; 팩스: 053-580-5165)

출된 짧은 구간의 묵음에서는 신뢰성있는 잡음에 대한 통계값을 추정하기가 쉽지 않을 것이다. 또한 잡음의 특성이 순간적으로 바뀌게 되면 추출된 묵음구간은 전체 잡음에 대한 정확한 정보를 제공하는데 실패할 것으로 생각된다. 따라서 PMC 방식에서 잡음 HMM의 파라미터를 추정하기 위해서는 Bayesian 접근 방식을 이용하는 것이 유리할 것으로 생각된다. 이것은 잡음에 관한 사전 정보를 이용하게 해 줄 뿐만 아니라 온라인상의 입력음성의 전 구간을 잡음파라미터의 추정에 이용할 수 있기 때문이다. 그러나 Bayesian 접근 방식을 이용하기 위해서는 효율적인 구현 알고리듬이 요구되며, 본 논문에서는 이러한 동기에 근거하여 Bayesian 방식에 의한 잡음평균 추정 알고리듬을 제시하고자 한다.

본 논문에서는 2장에서 제안된 알고리듬을 자세히 설명하고 3장에서 실험결과를 소개하며 끝으로 결론을 맺고자 한다.

II. Bayesian 적응 방법

이번 장에서는 HMM에 근거한 음성인식에서의 잡음평균의 Bayesian 적응방식에 대해서 소개한다. Bayesian 방식은 추정하고자 하는 파라미터에 관한 사전 분포를 사용한다는 점에서 ML 방식과 크게 차이가 난다. 길이가 T 인 잡음음성의 벡터 시퀀스 $Y=\{ y_1, y_2, \ldots, y_T \}$가 주어지고 $P(Y|\Lambda)$가 그들의 확률밀도함수 (probability density function)라 하자. Λ가 위의 분포를 특징짓는 파라미터의 집합을 의미한다면, Λ에 대한 ML 추정은 아래식을 풀면 얻어진다.

$$\frac{\partial P(y_1, y_2, \ldots, y_T | \Lambda)}{\partial \Lambda} = 0 \qquad (1)$$

만약 Λ에 대한 사전 분포인 $P_0(\Lambda)$가 알려져 있다면, Λ에 대한 Bayesian 적응은 MAP (maximum a posteriori) 추정에 의해서 실행되어질 수 있으며, Bayes 정리를 이용하여 아래와 같은 방정식의 해를 구하면 된다.

$$\frac{\partial P(y_1, y_2, \ldots, y_T | \Lambda) P_0(\Lambda)}{\partial \Lambda} = 0 \qquad (2)$$

Bayesian 적응 방식을 잡음평균의 추정에 적용하기 위해서는 먼저 생각해야될 두 가지 문제가 있다. 그 첫 번째는 잡음음성의 확률밀도함수인 $P(Y|\Lambda)$가 정해져야 한다. 확률밀도함수는 가급적 실제의 잡음 오염과정을 정확히 묘사할 수 있도록 정해져야 하겠다. 이를 위해서 본 논문에서는 간단하면서도 비교적 정확한 방법인 PMC 방식을 이용하고자 한다. 두 번째 고려할 문제는 HMM에 근거한 Bayesian 적응 방식을 적용하는 구체적 방법에 관한 것이다. 일반적으로, 식 (2)의 직접적인 해를 구하기가 쉽지 않기 때문에 본 논문에서는 EM 방식을 제안하고자 한다. 다음절에서는 위에서 언급한 내용들에 대해서 좀 더 구체적으로 설명하고자 한다.

2.1. 잡음음성의 확률밀도함수에 대한 고찰

PMC 방식에서는 잡음음성 y_l^j 은 다음과 같은 불일치 함수 F 에 의해서 특징지워진다[3].

$$y_l^j = F(x_l^j, n_l^j) = \log(\exp(x_l^j) + \exp(n_l^j)) \qquad (3)$$

여기서 x_l^j 과 n_l^j 은 각각 음성과 잡음의 로그 (log) 영역에서의 스펙트럼을 나타낸다.

연속밀도함수 HMM (continuous density HMM)에서 각 state 당 여러개의 mixture 성분을 가지고 있는 경우, 만약 어떤 음성벡터가 특정한 mixture 성분에 의해서 발생한다고 가정할 때, 이러한 성질이 잡음에 의해서도 영향을 받지 않는다고 가정하면, 잡음음성의 평균 $\hat{\mu}_k^j$은 아래의 식으로서 근사화될 수 있을 것이다[3].

$$\hat{\mu}_k^j = E\{F(x_l^j, n_l^j)\} = \log(\exp(\mu_k^j) + \exp(\mu_n^j)) \qquad (4)$$

여기서 μ_k^j 은 어떤 k 번째 mixture 성분에 해당하는 음성의 평균벡터이며 μ_n^j은 잡음의 평균벡터이고 이때 잡음은 1개의 mixture 성분으로 구성되었다고 가정한다. 그런데 일반적으로 음성인식에서는 위에서 언급한 로그-스펙트럼보다는 mel-frequency cepstrum coefficients (MFCC)를 음성 특징벡터로서 더 많이 사용하고 있다. 로그-스펙트럼 x_l^j 로부터 MFCC x_l^c 로의 변환은 코사인변환 (cosine transformation) 행렬 C 에 의해서 이루어진다[4]. 코사인변환은 선형적이므로 이 변환은 잡음 및 음성의 평균벡터에 대해서도 적용될 수 있다.

$$x_l^c = C x_l^j \qquad (5)$$

$$\mu_k^c = C \mu_k^j \qquad (6)$$

$$\mu_n^c = C \mu_n^j \qquad (7)$$

식 (4) 및 식 (5)~(7)을 이용하고 음성 HMM에서 잡음에 의해서 공분산행렬 Σ_k^c 이 변하지 않는다고 가정을 하면 MFCC를 이용한 잡음음성의 확률밀도함수는 아래와 같이 주어진다.

$$\begin{aligned} P(y_l^c, k \mid \mu_k^c, \Sigma_k^c, \mu_n^c) &= N(C\log(\exp(C^{-1}\mu_k^c) \\ &\quad + \exp(C^{-1}\mu_n^c)), \Sigma_k^c) \\ &= a_k \exp(-\tfrac{1}{2}[y_l^c - C\log(\exp(C^{-1}\mu_k^c) \\ &\quad + \exp(C^{-1}\mu_n^c))] \\ &\quad \cdot \Sigma_k^{c-1}[y_l^c - C\log(\exp(C^{-1}\mu_k^c) \\ &\quad + \exp(C^{-1}\mu_n^c))] \end{aligned} \qquad (8)$$

위에서 L은 MFCC의 차원이 되고 α_k는 $\frac{1}{(\sqrt{2\pi})^L |\Sigma_k^c|^{\frac{1}{2}}}$ 과 같다.

2.2. EM 알고리듬에 의한 잡음평균의 적응 알고리듬

잡음평균의 Bayesian 적응을 위한 목적 함수는 아래와 같다.

$$\log P(Y^c | \mu_n^c) P_o(\mu_n^c) \tag{9}$$

위의 식에서 HMM의 다른 파라미터들은 표기의 편리상 빠뜨렸다.

μ_n^c 에 대한 식 (9)의 최대화를 위해서는 음성인식에서 많이 쓰이는 segmental k-means 알고리듬이 채택되었으며 그 과정은 2단계로 나누어질 수 있다.

1) 먼저, 주어진 잡음평균 값 μ_n^c 에 대하여 최적의 상태-시퀀스 (state-sequence) $\hat{S}=\{s_1, s_2, \cdots, s_T\}$을 아래의 식에서 구한다.

$$\hat{S} = \arg\max_S \log P(Y^c, S | \mu_n^c) P_o(\mu_n^c) \tag{10}$$

2) 식 (10)에서 얻은 상태-시퀀스를 이용하여 아래와 같이 잡음평균의 추정치를 구한다.

$$\hat{\mu}_n^c = \arg\max_{\mu_n^c} \log P(Y^c, \hat{S} | \mu_n^c) P_o(\mu_n^c) \tag{11}$$

위의 두가지 과정은 반복적으로 수행되어 수렴하게 된다. 식 (10)에 의한 분할 (segmentation) 과정은 Viterbi 알고리듬에 의해서 가능하게 된다. 그러나 일반적으로 식 (11)에 대한 해는 직접적으로 구하기 어렵기 때문에 본 논문에서는 EM 방식을 사용할 것을 제안한다. EM 방식은 아래와 같은 새로운 보조함수 H 를 필요로 한다[5].

$$\begin{aligned} H(\mu_n^c, \overline{\mu}_n^c) &= Q(\mu_n^c, \overline{\mu}_n^c) + \log P(\overline{\mu}_n^c) \\ &= E[\log P(Y^c, \hat{S}, K | \overline{\mu}_n^c) | Y^c, \mu_n^c] + \log P_o(\overline{\mu}_n^c) \end{aligned} \tag{12}$$

여기서 K 는 어느 mixture 성분이 관측 데이터 벡터 Y^c 를 발생시켰느냐를 나타내는 비관측 데이터를 나타낸다. 새로운 보조함수를 이용하면 식 (11)에 대한 해는 아래와 같은 식의 해로 변한다.

$$\begin{aligned} \arg\max_{\mu_n^c} [& \sum_{t=1}^{T} \sum_{k=1}^{Kmax} P(s_t, k | y_t^c, \mu_n^c) \\ & \log P(y_t^c, s_t, k | \overline{\mu}_n^c) + \log P_o(\overline{\mu}_n^c)] \end{aligned} \tag{13}$$

여기서 $Kmax$는 연속밀도 HMM에서의 각 상태의 mixture 갯수이다. 사전분포 $P_o(\mu_n^c)$는 잡음평균에 대한 conjugate prior인 가우시안 (Gaussian) 밀도 함수로 정하였다[2]. 즉,

$$\begin{aligned} P_0(\mu_n^c) = \alpha_n \exp(-\frac{1}{2}(\mu_n^c - \mu_{n,i}^c) \\ \Sigma_{n,i}^{c\,-1} (\mu_n^c - \mu_{n,i}^c)^T) \end{aligned} \tag{14}$$

여기서 $\mu_{n,i}^c$ 와 $\Sigma_{n,i}^c$ 는 사전 (prior) 평균과 사전 분산이다. α_n 은 $\frac{1}{(\sqrt{2\pi})^L |\Sigma_{n,i}^c|^{\frac{1}{2}}}$ 와 같다. 이러한 사전 파라미터값들은 Bayesian 적응을 하기 전 훈련 데이터로부터 미리 얻어진다.

개선된 잡음 평균값을 구하기 위해서는 식 (14)를 식 (13)에 적용하고 $\overline{\mu}_n^c$ 에 대해서 식 (13)을 미분한 다음 그 미분값을 0 으로 하는 $\overline{\mu}_n^c$ 값을 구하면 된다. 복히, $\overline{\mu}_n^c$ 에 대한 $Q(\mu_n^c, \overline{\mu}_n^c)$ 의 미분값은 아래의 식과 같이 구할 수 있다.

$$\begin{aligned} \frac{\partial}{\partial \overline{\mu}_n^c} Q(\mu_n^c, \overline{\mu}_n^c) = [& \sum_{t=1}^{T} \sum_{k=1}^{Kmax} P(s_t, k | y_t^c, \mu_n^c) \cdot \\ & (y_t^c - C\log(\exp(C^{-1}\mu_k^c) \\ & + \exp(C^{-1}\overline{\mu}_n^c))) \Sigma_k^{c\,-1} \cdot] \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\partial}{\partial \overline{\mu}_n^c} (& C\log(\exp(C^{-1}\mu_k^c) \\ & + \exp(C^{-1}\overline{\mu}_n^c))) \end{aligned}$$

한편 W_k 를 다음과 같이 정의하면,

$$\begin{aligned} W_k = & \frac{\partial}{\partial \overline{\mu}_n^c} (C\log(\exp(C^{-1}\mu_k^c) \\ & + \exp(C^{-1}\overline{\mu}_n^c)))^T \end{aligned} \tag{16}$$

$\overline{\mu}_n^c$ 에 대한 식 (13)의 미분값은 아래와 같이 구해진다.

$$\begin{aligned} \frac{\partial}{\partial \overline{\mu}_n^c} H(\mu_n^c, \overline{\mu}_n^c) &= \frac{\partial}{\partial \overline{\mu}_n^c} Q(\mu_n^c, \overline{\mu}_n^c) \\ &\quad + \frac{\partial}{\partial \overline{\mu}_n^c} \log P_o(\overline{\mu}_n^c) \\ &= [\sum_{t=1}^{T} \sum_{k=1}^{Kmax} P(s_t, k | y_t^c, \mu_n^c) \\ &\quad (y_t^c - \overline{\mu}_n^c - C\log\frac{(\exp(C^{-1}\mu_k^c) + \exp(C^{-1}\overline{\mu}_n^c))}{\exp(C^{-1}\overline{\mu}_n^c)}) \\ &\quad \cdot \Sigma_k^{c\,-1} W_k - \overline{\mu}_n^c \Sigma_{n,i}^{c\,-1} + \mu_{n,i}^c \Sigma_{n,i}^{c\,-1}] \end{aligned}$$

따라서 위의 미분식을 0 으로 하면 개선된 잡음평균값 $\overline{\mu}_n^c$은 아래와 같이 얻을 수 있다.

$$\overline{\mu}_n^c = [\sum_{i=1}^{I}\sum_{k=1}^{K_{max}} P(s_i,k \mid y_i^c, \mu_n^c) \cdot$$

$$(y_i^c - C\log\frac{(\exp(C^{-1}\mu_k^c)+\exp(C^{-1}\overline{\mu}_n^c))}{\exp(C^{-1}\overline{\mu}_n^c)})(17)$$

$$\Sigma_k^{c\,-1} W_k + \mu_{n,i}^c \Sigma_{n,i}^{c\,-1}] \cdot$$

$$[\sum_{i=1}^{I}\sum_{k=1}^{K_{max}} P(s_i,k \mid y_i^c, \mu_n^c)\ \Sigma_k^{c\,-1} W_k + \Sigma_{n,i}^{c\,-1}]^{-1}$$

한편, 위의 식에서는 정형화된 해를 얻기 위해서 W_k 를 구할때 전 단계에서 구한 잡음 평균 값을 이용한다. 잡음 평균에 대한 Bayesian 적응 과정은 최종적으로 온라인상의 인식 음성에 대하여 적용되나 전체적으로 다음의 3가지 과정으로 이루어져 있다.

1) 사전 파라미터를 구하는 과정: 잡음의 사전 평균벡터 $\mu_{n,i}^c$ 와 사전 공분산 행렬 $\Sigma_{n,i}^c$ 를 훈련 데이터를 이용하여 먼저 구한다. 이는 다수의 잡음 샘플을 이용하여 잡음의 평균벡터와 분산을 ML (maximum likelihood) 방식으로 구함으로서 행해진다.
2) 초기 Bayesian 적응: 약간의 훈련 데이터를 이용하여 초기 Bayesian 과정을 수행한다. 이 과정을 통하여 어느 정도 잡음의 평균값을 추정하여, 실제 온라인 적응시 수렴속도를 빠르게 한다. 이것은 인식시 온라인 과정에서 많은 시간을 소모할 수 없기 때문에 필요하다.
3) 온라인 적응과정: 과정 1)에서 얻은 잡음평균에 대한 사전 파라미터값과 과정 2)에서 얻은 잡음평균의 초기치를 이용하여 실제의 온라인 테스트 음성을 이용하여 Bayesian 적응을 수행한다.

위의 과정에서는 모두 비감독 (unsupervised) 적응을 수행하므로 적응데이터가 어떤 단어인가 하는 정보들은 필요없게 된다.

그림 1에는 제안된 방식의 Bayesian 적응 과정에 대한 개요도가 그려져 있다.

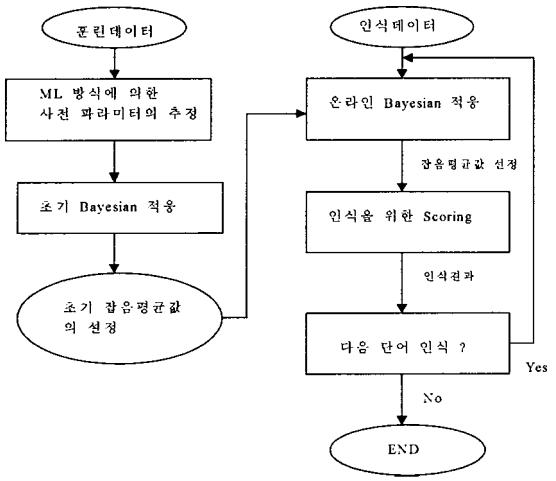

그림 1. 제안된 Bayesian 적응 방식의 개요도
Fig. 1. Flowchart of the proposed Bayesian adaptation method.

III. 인식 실험결과

본 장에서는 잡음평균을 Bayesian 적응하는 제안된 방식의 성능을 고립단어 인식실험을 통하여 알아보았다.

3.1. 데이터베이스와 기본 인식시스템

본 연구에서는 단순한 left-to-right 형식의 3개의 상태로 이루어진 연속밀도 HMM을 기본인식기로 사용하였으며 각 상태는 6개의 mixture 성분으로 이루어지도록 하였다. 인식 대상 어휘는 75개의 고립단어들을 이용하였으며 이 단어들은 음향학적으로 발생 빈도가 고르게 분포되도록 구성되어져 있다. 기본인식단위는 32개의 PLU (phoneme like unit)를 사용하였다. 기본인식시스템은 15명의 화자가 75단어를 한번씩 발성한 것을 이용하여 Baum-Welch 알고리듬에 의해 훈련되었다[6]. 또한, 1명의 훈련 화자로부터 발성된 75개의 단어를 초기 Bayesian 적응을 위하여 사용하였으며, 온라인상의 적응 및 인식실험을 위하여 20명의 화자로부터의 1495단어를 이용하였다. 인식실험을 위해서 사용된 잡음음성을 만들기 위해서 기존의 깨끗한 음성에 인위적으로 자동차 환경의 잡음을 부가적으로 더해주었으며 이때 다양한 신호대 잡음비 (SNR)를 이루도록 고려하였다. 또한 13차의 MFCC를 특징벡터로 이용하여 인식실험을 수행하였다.

3.2. 기본인식시스템의 성능

본 절에서는 기본인식시스템의 성능에 대해서 논의하고자 한다. 표 1에는 원래의 깨끗한 음성으로 훈련된 기본인식시스템을 다양한 SNR 환경하에서 인식테스트를 수행했을 경우의 성능을 나타내고 있다. 표 1에서 알 수 있듯이 인식율은 SNR이 10dB 이하로 내려가면 상당히 저하된다. 이러한 인식율의 저하 현상의 원인은 음성이 부가잡음에 의해서 심하게 왜곡됨에 따라서 원래의 훈련된 HMM들과의 불일치가 심해진 결과일 것이다. 한편, 표 2에는 기본인식시스템을 테스트 환경과 같은 조건의 SNR에서 재 훈련하였을 때의 인식 실험 결과를 나타내고 있다. 예상한 대로 인식결과는 상당히 향상됨을 알 수 있다. 이것은 음성의 왜곡현상이 충분한 훈련에 의해서 HMM에 반영되었기 때문에 가능한 것으로 생각되며, 본 실험에서의 궁극적인 목표치가 될 것이다.

표 1. 기본인식기의 잡음음성에 대한 인식율
Table 1. Recognition rates of the baseline recognizer for noisy speech recognition.

인식환경	0 dB	10 dB	20 dB	cleam
인식율(%)	26.4	64.4	86.5	92.9

표 2. 학습과 인식환경이 동일한 경우의 기본 인식기의 인식율
Table 2. Recognition rates of the baseline recognizer when the testing condition is the same as that of the training.

인식환경	0 dB	10 dB	20 dB	cleam
인식율(%)	83.2	90.8	92.2	92.9

3.3. 제안된 알고리듬의 성능 비교

표 3에는 제안된 Bayesian 적응방식과 기존의 PMC 방식과의 인식결과를 비교하고 있다. 또한 좀 더 객관적 성능 비교를 위해서, 기존에 널리 알려진 Codeword-Dependent Cepstral Normalization (CDCN)[8] 방식과 주파수차감법[9]에 대한 실험결과도 함께 나타내었다.

전반적으로 주파수차감법이나 CDCN 방식에 비해서 제안된 Bayesian 적응방식이 다소 높은 인식율을 나타냄을 알 수 있었다. 특히, CDCN 방식은 SNR이 20dB 인 경우 상당히 인식성능이 떨어짐을 보여 주었는데, 이는 CDCN 방식은 원래 Discrete HMM을 이용한 인식을 위해서 제안되었고[8], 본 연구에서는 인식엔진을 연속밀도 HMM을 사용하였으므로 코드워드에 대한 최적화 등을 이루지 못한 관계로 이러한 예상 밖의 저조한 인식율이 나온 것으로 보인다.

온라인 Bayesian 적응은 각 단어 수준에서 행하여지며, 적응을 거친후 같은 단어를 인식실험에 사용한다. 온라인 Bayesian 적응에서는 많은 계산시간을 소비할 수 없기 때문에, 잡음평균벡터의 초기치를 미리 정하기 위해서 초기 Bayesian 적응을 미리 수행한다. 이를 위해서는 1명의 훈련화자로 부터의 잡음음성을 이용한다. 초기 Bayesian 적응을 통해서 얻은 잡음평균의 초기값을 이용하여 실제 온라인 Bayesian 적응에서는 EM 알고리듬을 1회 반복 수행 하였다. 표 3에는 기존의 PMC 방식과 더불어 초기 Bayesian 적응 결과 및 온라인 적응 결과를 SNR 값이 변화함에 따라서 나타내었다.

표 3. 제안된 Bayesian 적응 방식의 성능비교
Table 3. Performance comparison between the proposed bayesian adaptation method and the conventional methods.

인식환경	0 dB	10 dB	20 dB
주파수차감법	61.0 (%)	82.8 (%)	88.4 (%)
CDCN	81.7 (%)	84.6 (%)	84.7 (%)
PMC	79.8 (%)	87.4 (%)	89.9 (%)
초기 Bayesian 적응 후	79.3 (%)	87.6 (%)	90.9 (%)
온라인 Bayesian 적응 후	82.3 (%)	88.0 (%)	91.1 (%)

초기 Bayesian 결과는 기존의 PMC 방식에 비해서 별반 차이가 없었다. 그러나 온라인 적응 후에는 상당한 인식성능의 향상이 나타남을 알 수 있었다. 그 이유는 온라인 적응에서는 인식하고자 하는 단어를 적응에 이용함으로서 실제 잡음이 미치는 영향을 직접적으로 보상할 수 있기 때문일 것이다. 이와 같이 Bayesian 적응방식은 사전의 잡음정보만을 이용하는 기존의 방식에 비해서 우월한 성능을 나타낼 수 있다고 보여진다. 또한 이러한 Bayesian 방식의 장점은 미리 얻어진 사전 잡음정보에 에러가 존재할 때 더욱더 크게 나타나리라 보여진다. 실제로 기존의 PMC 방식에서는 입력음성중의 묵음구간을 검출하여 잡음음성의 통계정보를 추정한다. 그런데, 묵음구간의 선정은 항상 에러를 유발할 가능성이 있고 또한 잡음의 특성이 갑작스럽게 변하게 되면, 추정된 묵음구간에서의 잡음의 특성은 전체 잡음의 특성과는 차이가 발생할 수 있을 것이다. 이러한 잡음정보 추정의 에러에 의한 인식율의 저하는 PMC 방식에서는 매우 심각할 것으로 생각되나 Bayesian 적응에서는 전체 입력음성을 적응에 이용하므로 비록 잡음에 관한 사전정보에 에러가 있다고 하더라도 인식율의 감소는 그리 크지 않을 것으로 생각된다. 이와 같은 예상을 직접 인식실험을 통해서 알아보기 위해서, 표 4에서는 잡음의 통계정보를 추정할 때의 SNR 값과 실제 인식시의 잡음의 SNR 값이 틀릴 경우의 PMC 방식에서의 인식율을 나타내었다. 또한 비교를 위해서 Bayesian 적응에서 사전 잡음 파라미터 (평균 벡터 $\mu^{c}_{n,i}$ 와 공분산 행렬 $\Sigma^{c}_{n,i}$)를 추정할 때의 SNR 값과 실제 온라인 상에서 적응할 때의 SNR 값이 틀린 경우에 인식율을 나타내었다. 이러한 인식율의 비교를 통하여 두 방식에서의 잘못 추정된 잡음정보가 인식성능의 저하에 미치는 영향을 간접 비교할 수 있으리라 생각된다.

표 4. 잡음 통계 추정시와 인식시의 SNR 값이 서로 다를 경우의 제안된 Bayesian 적응 방식과 PMC 방식의 인식결과
Table 4. Recognition rates of the proposed Bayesian adaptation method and the PMC method when the SNRs of the estimated noise statistic are different from those of the testing speech.

	잡음통계 추정 SNR 값	인식시의 SNR 값		
		0 dB	10 dB	20 dB
Bayesian 적응	0 dB	82.3 (%)	85.7 (%)	78.0 (%)
	10 dB	81.3 (%)	88.0 (%)	88.2 (%)
	20 dB	74.6 (%)	88.3 (%)	91.1 (%)
PMC	0 dB	79.8 (%)	67.9 (%)	57.7 (%)
	10 dB	73.2 (%)	87.4 (%)	81.5 (%)
	20 dB	44.9 (%)	86.2 (%)	89.9 (%)

두 방식에서는 모두 잘못된 잡음정보의 추정에 의해서 전반적으로 인식성능의 저하가 일어남을 알 수 있었다. 그러나 Bayesian 적응방식에서는 PMC 방식에 비해서는 훨씬 적은 인식성능의 저하가 생김을 알 수 있다. 예를 들면, 테스트 입력음성의 SNR 값이 0dB인 경우에 사전 잡음파라미터 값이 각각 10dB와 20dB에서 추정 되었다면 인식율은 각각 81.3% 와 74.6%로 나타난다. 이것은 동일한 SNR 환경일때의 인식율인 82.3%에 비해서는 다소 낮아진 것이다. 그러나 PMC인 경우에는 잡음의 통계

정보가 10dB와 20dB에서 얻어진 경우에 인식율이 각각 72.8%와 44.9%로 나타남을 알 수 있다. 이러한 차이는 테스트 음성의 SNR 값이 10dB와 20dB인 경우에도 비슷하게 나타남을 알 수 있으며, 이는 Bayesian 방식이 사전 잡음 파라미터 값의 오류에 대하여 상대적으로 더 강인함을 나타낸다.

IV. 결 론

본 논문에서는 잡음에 강인한 음성인식을 위하여 잡음 평균의 Bayesian 적응 방식에 관하여 제안하였다. 제안된 방식에서는 PMC 방식을 이용하여 잡음음성을 모델링하였으며, 잡음음성 HMM에 근거하여 잡음의 평균벡터를 Bayesian 방식으로 적응하였다. 한편 Bayesian 적응을 위해서는 EM 방식에 근거한 반복적 추정알고리듬이 제안되었으며, 자동차 잡음 환경하에서의 인식 실험 결과들을 통해서 제안된 Bayesian 적응 방식이 상당히 효과적임을 알 수 있었다. 이러한 인식 성능의 향상은 Bayesian 적응 방식이 잡음평균에 대한 사전 정보를 이용할 수 있을 뿐만 아니라 테스트 단어를 직접 적응에 이용할 수 있는 구조를 가진데에 기인한다고 보여진다. 또한 Bayesian 적응방식은 PMC 방식에 비해서 사전정보의 오류에 강인한 특성을 가진 것을 인식 실험을 통해서 알 수 있었다. 본 논문에서 제안된 Bayesian 방식은 PMC 방식에 근거하여 인식 성능의 향상을 이루었으나, 다른 방식의 모델 변환을 이용한 잡음음성 인식에서도 그 적용이 가능하리라고 생각된다.

감사의 글

본 연구는 1999년도 계명대학교 비사연구기금으로 수행되었습니다. 지원에 감사드립니다.

참 고 문 헌

1. P. J. Moreno, *Speech Recognition in Noisy Environments, Ph. D. Dissertation*, Carnegi Mellon University, 1996.
2. Chin-Hui Lee, Chih-Heng Lin and Biing-Hwang Juang, "A study on speaker adaptation of the parameters of the continuous density hidden Markov models," IEEE Trans. on Signal Processing, vol. 39, no. 4, April, 1991.
3. M. J. F. Gales, *Model Based Techniques for Noise Robust Speech Recognition*, Ph. D. Dissertation, University of Cambridge, 1995.
4. Davis S. B. and Mermelstein P., "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions ASSP, vol. 28, pp. 357-366, 1980.
5. A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Stat. Society, Series B, vol. 39, pp. 1-38, 1977.
6. L. E. Baum, G. S. T. Petrie and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math., Statist., vol. 41, pp. 164-171, Jan. 1970.
7. 장유현, 정용주, 박성현, 은종관, "잡음환경에서의 음성인식을 위한 모델파라미터 변환 방식에 관한 연구," 한국음향학회지 제16권 제5호, pp. 112-121, 1997.
8. Alejandro Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.
9. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions. ASSP, vol. 27, pp. 113-120, 1979.

▲ 정 용 주 (Yong-Joo Chung)
1988년 2월 : 서울대학교 전자공학과 졸업 (공학사)
1995년 8월 : 한국과학기술원 전기 및 전자공학과 졸업 (공학박사)
1995년 9월～1999년 2월 : LG정보통신(주) 중앙연구소 재직
1999년 3월～현재 : 계명대학교 컴퓨터전자공학부

※ 주관심분야 : 음성인식, 멀티미디어 신호처리, 패턴인식