직접데이터 기반의 모델적응 방식을 이용한 잡음음성인식에 관한 연구*

A Study on the Noisy Speech Recognition Based on the Data-Driven Model Parameter Compensation

정 용 주**
Yong-Joo Chung

ABSTRACTS

There has been many research efforts to overcome the problems of speech recognition in the noisy conditions. Among them, the model-based compensation methods such as the parallel model combination (PMC) and vector Taylor series (VTS) have been found to perform efficiently compared with the previous speech enhancement methods or the feature-based approaches. In this paper, a data-driven model compensation approach that adapts the HMM(hidden Markv model) parameters for the noisy speech recognition is proposed. Instead of assuming some statistical approximations as in the conventional model-based methods such as the PMC, the statistics necessary for the HMM parameter adaptation is directly estimated by using the Baum-Welch algorithm. The proposed method has shown improved results compared with the PMC for the noisy speech recognition.

Keywords: Speech recognition, model-based compensation, PMC

1. 서 론

모델 파라미터 보상기법은 HMM에 기반한 잡음음성인식에 있어서 매우 좋은 결과를 나타냄이 알려져 있다[1][2][3]. 그중에서도 PMC는 다른 방식에 비해서 효과적인 것으로 나타난다. 특히, 계산량이 VTS 등과 같은 다른 모델 적응 방식에 비해서 비교적 작게 소요되며, 적응을 할 때에도 입력음성에 포함된 소량의 잡음 샘플을 이용하므로 매우 효율적이다. 하지만, PMC 방식에서는 해석의 편리함을 위해서 몇몇 통계적인 가정을 하고 있는데, 이러한 가정은 주로 특징벡터의 영역 변환이나 모델결합과정에서 발생하며, 결과적으로 HMM모델 보상과정에서 오차를 발생하여 인식율의 저하를 가져오는 것으로 생각된다. 특히, Hung[3] 등은 로그스펙트럼 특징벡터의 분포가 PMC 방식에서 가정한 것과는 다르게 가우시안 분포가 되지 않음을 보였으며, 이러한 제한을 완화함으로서 보다 높은 인식률을 얻을 수 있었다. 또한 PMC 방식에서는 모델파라미터 변환을 위해서, 로그스펙

* 본 연구는 한국과학재단 목적기초연구(R01-2003-000-10242-0) 지원으로 수행되었음.
** 계명대학교 전자공학과

트럼 벡터의 평균 μ^l을 켑스트럼 평균벡터 μ^c로부터 역 DCT(discrete cosine transformation)[5] 변환을 통하여 다음과 같이 얻게 된다.

$$\mu^l = C^{-1} \mu^c \tag{1}$$

하지만, 로그스펙트럼과 켑스트럼은 서로 다른 차원의 벡터공간을 형성하므로 정확한 역 변환이 일어나기가 어렵다. 즉, 서로 다른 로그스펙트럼값이 DCT 변환에 의해서 동일한 켑스트럼값으로 변환이 되는 경우가 발생할 수 있으며, 이후 위의 식(1)에 의해서 역DCT 변환을 거치면 원래의 로그스펙트럼값의 복원이 어려워지게 된다.

이와 같은 PMC 방식의 단점을 보완하고자, 우리는 특정 통계 정보값을 깨끗한 음성데이터를 이용하여 직접 추정한 후, 이를 인식시에 적용하여 HMM파라미터를 보상하는 방법을 제시한 바 있다[4]. 이 방식에서는 통계정보가 훈련 중에 미리 얻어지므로, 인식시에 요구되는 계산량이 비교적 적으며, 또한 PMC에서와 같은 통계적인 가정을 하지 않으므로, 잡음에 의한 음성변이를 좀 더 정확하게 HMM파라미터에 반영 할 수 있다는 장점이 있다.

하지만, 기존의 우리가 제안된 연구방식에서는 적응의 간편함을 위해서 다소 간의 해석적인 부정확함이 있었으며, 이것은 HMM의 평균벡터의 적응에 있어서는 별 다른 문제를 만들지는 않았으나, 공분산 행렬의 적응에 있어서는 추가적인 적응데이터의 필요성 등의 문제점을 가지고 있었다. 따라서, 본 논문에서는 기존에 우리가 제안한 연구방법이 가지고 있던 해석적 부정확성을 보완하는 연구결과를 소개하고자 한다. 이 방법은 기존 방식과 인식성능에 있어서는 큰 차이를 보이진 않았으나, 공분산 행렬의 적응에서 기존방식의 단점을 보완해 줄 수 있었다.

본 논문의 구성은 2 장에서 제안된 방식을 소개하며 기존의 연구방법과의 차이도 설명하고자 한다. 또한 3 장에서는 제안된 방식에 의한 연구 결과를 소개하며 4 장에서 결론을 맺고자 한다.

2. HMM 파라미터 변환을 위한 통계정보의 추정

먼저, 켑스트럼 영역에서의 잡음음성에 대한 왜곡 현상은 일반적으로 아래와 같은 식으로 묘사되어진다.

$$\begin{aligned} Y^c &= C \log(X+N) \\ &= C \log X + C \log(i + \exp(\log N - \log X)) \end{aligned} \tag{2}$$

여기서, X는 깨끗한 원래 음성의 파워 스펙트럼(power spectrum)을 나타내며 N은 부가잡음신호의 파워스펙트럼을 나타낸다. 또한 i는 모든 원소의 값이 1인 단위벡터을 나타내며, C는 DCT 변환을 나타낸다.

일반적으로 HMM을 기반으로 한 음성인식에서는 Baum-Welch 알고리듬[6]을 이용하여 파라미터 추정이 이루어진다. 예를 들어, 연속밀도(continuous density) HMM의 경우, 캡스트럼 특징 평균벡터 μ_{jk}^{c}는 아래의 식과 같이 추정된다.

$$\mu_{jk}^{c} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\, \boldsymbol{C} \log \boldsymbol{X}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3}$$

여기서 $\gamma_t(j,k)$는 캡스트럼 특징벡터 $C\log X_t$가 임의의 HMM 상태(state) j와 혼합성분(mixture component) k에서 발생했을 확률값을 나타낸다 [8].

기존의 잡음결정모델을 이용한 방식[4]에서는 잡음음성에 대한 평균벡터를 구하기 위해서, 위의 식(3)에서, 원래의 음성인 $C\log X_t$ 대신에 잡음음성에 해당하는 식(2)의 Y_t^c 값을 적용하였으며 그 결과는 다음과 같다.

$$\hat{\mu}_{jk}^{c} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\left(\boldsymbol{C} \log \boldsymbol{X}_t + \boldsymbol{C} \log\left(\boldsymbol{i} + \exp\left(\log \boldsymbol{N} - \log \boldsymbol{X}_t\right)\right)\right)}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{4}$$

$$= \mu_{jk}^{c} + \frac{\sum_{t=1}^{T} \gamma_t(j,k)\left(\boldsymbol{C} \log\left(\boldsymbol{i} + \exp\left(\log \boldsymbol{N} - \log \boldsymbol{X}_t\right)\right)\right)}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{5}$$

$$= \mu_{jk}^{c} + E\left(\boldsymbol{C} \log\left(\boldsymbol{i} + \exp\left(\log \boldsymbol{N} - \log \boldsymbol{X}_t\right)\right)\right) \tag{6}$$

$$= \mu_{jk}^{c} + \mu_{jk}^{n} \tag{7}$$

위의 식에서 알 수 있는 바와 같이 잡음음성에 대한 평균벡터를 구하기 위해서는 최종적으로 μ_{jk}^{n}을 추정해야 한다. 이를 구하기 위해서는 잡음 $\boldsymbol{N}$이 포함된 항에 대한 평균값을 얻을 수 있어야 한다. 하지만, 잡음 $\boldsymbol{N}$은 인식시에만 얻을 수 있는 값이므로, 실제 Baum-Welch 방식을 이용한 훈련시에는 알 수 없게 된다. 따라서, 기존의 연구에서는 아래의 식을 이용하여 원하는 결과를 얻을 수 있었다.

$$\begin{aligned} \mu_{jk}^{x} &= E(\boldsymbol{X}_t) \\ \mu_{jk}^{n} &= \boldsymbol{C} \log\left(\boldsymbol{i} + \exp\left(\log \boldsymbol{N} - \log \mu_{jk}^{x}\right)\right) \end{aligned} \tag{8}$$

즉, 훈련 중에 얻을 수 있는 깨끗한 음성 파워스펙트럼 벡터 $\boldsymbol{X}_t$에 대한 평균값을 구한 후, 이를 μ_{jk}^{n}을 구하는 식에 대입하였다. 여기에는 다소 해석적 정확성에 문제가 있는 것이 사실이지만, 실

제적인 인식실험결과 만족할 만한 결과를 얻을 수 있었다. 하지만, 분산의 적응에 있어서는 위의 식(8)과 같은 간단한 접근 방법이 가능하지 않기 때문에, 잡음에 대한 적응데이터를 훈련 중에 얻을 수 있어야 하는 문제점이 있었다. 따라서, 본 논문에서는 기존 방식에 비해서 해석적 엄밀성을 기하고 또한 분산의 적응을 위해서 따로 적응데이터가 필요하지 않는 잡음적응 알고리듬을 제안하고자 한다. 제안된 방식은 기본적으로 기존의 방식과 그 틀을 같이한다. 즉, Baum-Welch 알고리듬의 적응과정에서 원래의 음성 대신에 잡음음성을 이용한다. 그러나, μ_{jk}^{n}을 구하기 위해서 잡음 N을 포함한 식(6)에 대해서 테일러 (Taylor) 전개식을 사용하여 잡음 N이 훈련 중에 얻어 질 수 없는 한계를 극복하고자 한다. 따라서, μ_{jk}^{n}을 얻는 식은 다음과 같이 전개된다.

$$\begin{aligned}\mu_{jk}^{n} &= E(C\log(i+\exp(\log N-\log X_t)))\approx \\ &E(C\log(i+\exp(\log \hat{N}-\log X_t)))+E\left(\left[\frac{\partial C\log(i+\exp(\log \hat{N}-\log X_t))}{\partial \hat{N}}\right]\right)(N-\hat{N}) \quad (9)\\ &+\frac{1}{2}E\left(\left[\frac{\partial^2 C\log(i+\exp(\log \hat{N}-\log X_t))}{\partial \hat{N}^2}\right]\right)(N-\hat{N})(N-\hat{N})\end{aligned}$$

위 식에서 $\hat{N}$은 훈련중에 미리 가정된 잡음신호값이다. 식(9)에서 나타난 바와 같이 μ_{jk}^{n}을 구하기 위해서는 $\hat{N}$를 포함하는 항에 대한 평균값을 훈련 중에 미리 구한 후, 인식시에 위의 테일러 전개식과 인식중의 잡음값 N 등을 이용한다. 위의 평균값들은 훈련중의 많은 음성데이터들을 이용함으로 해서 그 추정치의 신뢰도가 매우 높다고 생각되며, 따라서 잡음음성의 평균 벡터값들을 안정적으로 얻을 수 있게 된다.

위의 식(9)에서 우리가 주목해야 할 것은 N과 $\hat{N}$ 사이의 차이가 테일러 전개식에 있어서 많은 오차를 주어서는 안 된다는 것이다. 그림1에서는 SNR(signal to noise ratio)이 변화함에 따라서 테일러 전개식의 근사값이 원래의 값에 대해서 상대적으로 얼마만큼의 오차를 낳는지를 백분율(%)의 단위로 나타내고 있다.

N과 $\hat{N}$ 사이의 상대적 오차값인 $RD\left(=\frac{|N-\hat{N}|}{|N|}\right)$ 에 따라서 테일러 전개식의 오차가 많은 영향을 받음을 알 수 있다. 그림에서 우리는 비교적 RD 값의 범위가 큰 경우에도 작은 오차를 얻을 수 있음을 알 수 있는데, RD 값이 0.8보다 크지 않는 경우에 오차값은 SNR이 0 dB 경우 에서도 15(%)를 넘지 않음을 알 수 있다. 또한 RD 값이 0.4 보다 작은 경우에는 전개 오차 값이 5(%)도 넘지 않음을 보여준다.

본 연구에서는 테일러 전개 오차 값을 가능한 한 많이 줄이기 위해서 미리 몇 개의 가정된 잡음 $\hat{N}$에 대해서 통계치를 구하였고, 실제로 인식시의 잡음값 N과 비교해서 가장 근사한 $\hat{N}$에 해당하는 통계정보를 이용하여 HMM 파라미터 값을 보정하였다.

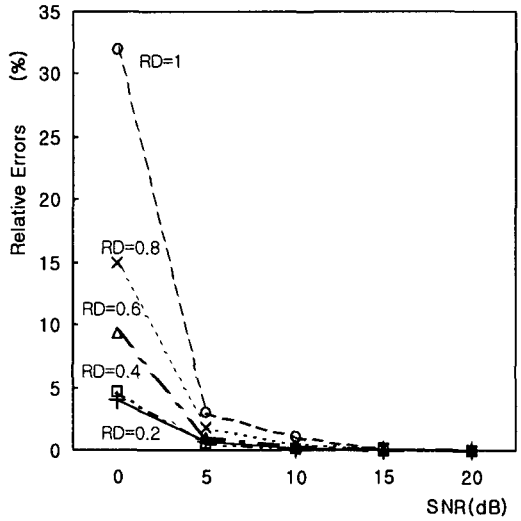

그림 1. 원래의 값과 테일러 전개식 근사값 간의 상대적인 오차(%)

잡음음성 Y_t^c에 대한 공분산 행렬도 평균벡터와 유사하게 적응할 수 있는데, 그 식은 아래와 같다.

$$\hat{\Sigma}_{jk}^{c} = E((Y_t^c - \hat{\mu}_{jk}^{c})(Y_t^c - \hat{\mu}_{jk}^{c})^T) = E(Y_t^c Y_t^{cT}) - \hat{\mu}_{jk}^{c}\hat{\mu}_{jk}^{cT} \tag{10}$$

위의 식(10)에서 $\hat{\mu}_{jk}^{c}$는 식(4)로부터 미리 얻어진 값이며, $E(Y_t^c Y_t^{cT})$는 다음과 같이 표현된다.

$$\begin{aligned} E(Y_t^c Y_t^{cT}) = E((&C\log X_t + C(\log i + \exp(\log N - \log X_t))) \\ &\cdot (C\log X_t + C\log(i + \exp(\log N - \log X_t)))^T) \end{aligned} \tag{11}$$

한편, 위의 식(11)을 다시 전개하면 다음과 같다.

$$E(Y_t^c Y_t^{cT}) = E(C\log X_t \cdot (C\log X_t)^T) \tag{12}$$

$$+ 2E(C\log X_t \cdot (C\log(i + \exp(\log N - \log X_t)))^T) \tag{13}$$

$$+ E(C\log(i + \exp(\log N - \log X_t)) \cdot (C\log(i + \exp(\log N - \log X_t)))^T) \tag{14}$$

한편, 식(12)는 깨끗한 켑스트럼 벡터의 평균치이므로 쉽게 얻어진다. 하지만, 식(13), (14)는 인식시에 발생하는 잡음 N과 관련되어 있으므로, 훈련시에 바로 구할 수 없고, 평균벡터의 적응에서와 마찬가지로 벡터 테일러 전개식을 이용하여 정리 할 수 있다.

식(13)은 다음과 같이 전개된다.

$$2E(\boldsymbol{C}\log\boldsymbol{X}_t\cdot(\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\boldsymbol{N}-\log\boldsymbol{X}_t)))^T)$$

$$\cong 2E(\boldsymbol{C}\log\boldsymbol{X}_t\cdot(\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\hat{\boldsymbol{N}}-\log\boldsymbol{X}_t)))^T) \quad (15)$$

$$+2E(\boldsymbol{C}\log\boldsymbol{X}_t\cdot(\frac{\partial\,\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\hat{\boldsymbol{N}}-\log\boldsymbol{X}_t)))^T}{\partial\hat{\boldsymbol{N}}}))(\boldsymbol{N}-\hat{\boldsymbol{N}}) \quad (16)$$

$$+2E(\boldsymbol{C}\log\boldsymbol{X}_t\cdot(\frac{\partial^2\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\hat{\boldsymbol{N}}-\log\boldsymbol{X}_t)))^T}{\partial\hat{\boldsymbol{N}}^2}))(\boldsymbol{N}-\hat{\boldsymbol{N}})(\boldsymbol{N}-\hat{\boldsymbol{N}}) \quad (17)$$

위에서 식(15)에 비해서 벡터 미분식을 포함하고 있는 식(16), (17)은 일반적으로 무시해도 될 정도의 작은 값이므로 적응시 고려하지 않아도 인식성능에 미치는 영향은 거의 없다. 한편, 식(14)의 경우에도 식(13)의 전개 때와 마찬가지로 그 값이 작게 나오는 벡터 미분항을 무시하면 다음과 같이 근사화된다.

$$\begin{aligned}&E(\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\boldsymbol{N}-\log\boldsymbol{X}_t))\cdot(\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\boldsymbol{N}-\log\boldsymbol{X}_t)))^T)\\ \cong\ &E(\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\hat{\boldsymbol{N}}-\log\boldsymbol{X}_t))\cdot(\boldsymbol{C}(\log(\boldsymbol{i}+\exp(\log\hat{\boldsymbol{N}}-\log\boldsymbol{X}_t)))^T)\end{aligned} \quad (18)$$

따라서, 공분산의 적응을 위해서는 위의 식들 중에서 식(10), (12), (15) 그리고 (18)을 결합하여 처리 하였다.

제안된 방식에서는 차분특징벡터(delta-MFCC)에 대한 파라미터 변환도 가능하다. 이것은 기존의 PMC 방식에서 차분특징벡터가 선형회귀계수(linear regression coefficient)로부터 얻어지는 경우에 해석적으로 파라미터 변환이 가능하지 않는 점과 비교하여 큰 장점이라 할 수 있다. 선형회귀계수를 이용한 경우에 차분특징벡터 $\Delta\boldsymbol{Y}_t$는 다음과 같이 얻어진다.

$$\Delta\boldsymbol{Y}_t=\mu\sum_{k=-K}^{k=K}k\boldsymbol{Y}_{t+k} \quad (19)$$

여기서 μ는 정규화 상수이다. 위의 식에서 $\Delta\boldsymbol{Y}_t$는 정적특징 벡터 $\boldsymbol{Y}_{t+k}$들의 선형결합이므로, 차분특징벡터에 대한 평균값을 얻기 위해서는 $E(\boldsymbol{C}\log(\boldsymbol{i}+\exp(\log\boldsymbol{N}-\log\boldsymbol{X}_{t+k})))$, $k=-K,\cdots,K$ 등의 통계정보를 얻으면 될 것이다. 물론 이 경우에도 정적 파라미터의 변환에서와 마찬가지로 테일러 전개식을 이용함으로서 인식시의 잡음 $\boldsymbol{N}$을 직접 얻을 수 없는 어려움을 극복 할 수 있다.

일반적으로 PMC 방식의 경우에는 차분특징벡터에 대한 평균값 $\Delta\hat{\mu}_i^l$을 추정하기 위해서는 다음과 같은 근사치를 이용한다 [2].

$$\Delta\hat{\mu}_i^l\ \approx\ \frac{\exp(\mu_i^l)}{\exp(\mu_i^l)+\exp(\hat{\mu}_i^l)}\Delta\mu_i^l \quad (20)$$

여기서, $\hat{\mu_i^l}$과 μ_i^l은 잡음음성과 원래의 깨끗한 음성에 대한 로그파워스펙트럼 벡터의 i번째 성분을 나타낸다.

그림 2에서는 제안된 전체 알고리듬의 흐름도가 나타나 있다. 여기서 우리는 전체 과정을 인식과정과 훈련과정으로 나누었다. 훈련과정에서는 가정된 몇 개의 잡음 신호값 $\hat{N}$에 대해서 인식에 필요로 하는 통계정보치를 얻게 된다. 이러한 정보는 인식과정에서 테일러 전개식을 이용할 때 사용된다. 인식시에는 입력음성이 들어오면 이로부터 잡음신호값 N을 추정하게 된다. 이것은 묵음구간을 추정한 후 이 구간으로부터 잡음신호의 평균값을 얻음으로서 이루어진다. 이때 추정된 N값과 가장 거리가 가까운 $\hat{N}$값을 선택한다. $\hat{N}$가 선택되면, $\hat{N}$에 해당하는 통계정보를 이용하고 $\hat{N}$과 N의 차이를 이용하여 테일러 전개식을 계산한다. 그리고 테일러 전개식의 오차를 검사하며, 이 오차가 선택한 임계치보다 작은 경우는 제안된 방식에 의하여 HMM의 평균값과 공분산 행렬을 추정하게 된다. 하지만, 테일러 전개식의 오차가 임계치 이상으로 큰 경우가 가끔 발생하게 되는데, 이러한 경우는 테일러 전개식을 이용한 파라미터 추정은 에러를 수반하게 된다. 이러한 경우에는 식(8)에서 제시된 기존의 방법을 이용하여 평균벡터값만을 추정하게 되며 공분산행렬은 새로이 추정하지 않는다.

3. 인식실험 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-to-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 교차 반복실험을 통해서 매번 훈련화자그룹과 인식화자그룹을 달리하여서 인식결과의 신뢰도를 높였다. 각 화자는 75 개의 단어를 1 회씩 조용한 사무실 환경에서 발성하였고 이 음성데이터는 16 kHz, 16 bit로 A/D 변환되었다. 잡음음성에 대한 실험을 위하여 실제로 자동차 내에서 발생하는 잡음을 녹음하여 A/D 변환한 것을 사용하였다. 특징벡터로서는 18 차의 멜-스케일(mel-scale)의 로그스펙트럼을 DCT 변환하여 얻은 13 차의 MFCC(mel-frequency cepstrum coefficients)을 사용하였다. 또한 선형회귀계수를 산정하는 방식을 이용한 차분특징(delta-MFCC)을 부가적으로 사용하여 전체의 특징벡터의 계수는 26 차가 되도록 하였다.

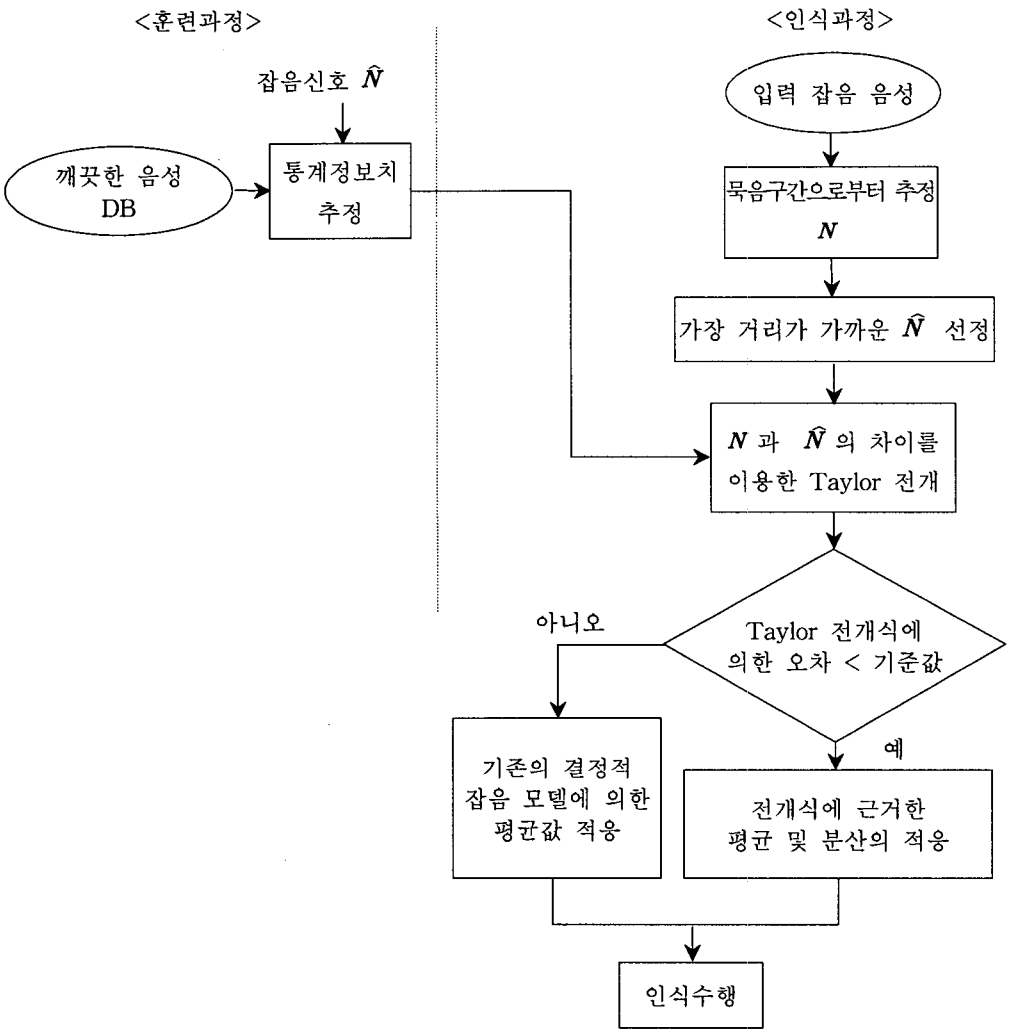

그림 2. 전체 알고리듬의 흐름도

3.2 인식실험 결과

표 1에는 각 HMM의 상태별로의 혼합성분의 개수가 6인 경우, 제안된 방식의 인식률이 SNR 값에 따라서 변화하는 것을 나타내었다. 또한 기반인식시스템과 재훈련을 한 경우 그리고 기존의 PMC 방식에 대한 인식실험 결과도 성능비교를 위해서 제시되었다.

표 1. 제안된 방식과 log-normal PMC와의 인식성능 비교

		0 dB	10 dB	20 dB
기반인식기		55.7	89.9	94.6
재훈련시		89.1	95.6	97.5
log-normal PMC	MFCC 평균	86.9(81.5)	93.7(90.2)	96.8(93.5)
	+delta-MFCC 평균	86.7	94.1	96.8
	+MFCC 공분산	84.3	94.1	96.9
제안된 방식	MFCC 평균	87.5(82.4)	93.8(90.2)	96.9(93.4)
	+delta-MFCC 평균	89.7	96.0	97.6
	+ MFCC 공분산	89.9	95.9	97.5

먼저, 기반인식기의 경우에는 아무런 보상작업이 이루어지지 않았으므로, 예상한 대로 잡음신호의 세기가 강해질수록 인식률이 급격히 저하되는 것을 알 수 있었다. 한편, 잡음음성을 이용하여 인식기를 재훈련한 경우에는 인식성능이 기반인식기에 비해서 월등함을 알 수 있다. 이것은 잡음신호에 의해서 발생하는 음성의 변이가 재훈련의 과정을 거치면서 HMM 파라미터값에 충실하게 전달되었기 때문이라고 생각된다. 따라서, 재훈련에 의한 인식결과는 파라미터변환을 이용한 음성 인식률 향상에 있어서, 하나의 벤치마크(benchmark) 성능이라고 할 수 있다. 물론 여기에는 화자나 잡음의 변이에 의한 영향이 충분히 고려되지 않을 수 있다.

제안된 연구결과는 잡음음성인식에서 널리 알려져 있는 log-normal PMC 방식과 비교되었다. 두 가지 방식을 비교 대상으로 선정한 것은 서로 간에 공통의 요소를 가지고 있기 때문이다. 먼저, 두 가지 방식은 다른 모델변환방식들에 비해서 소요되는 계산량이 비교적 적다. 그리고, 보다 중요한 것은, 두 방식이 모두 다 순수히 모델파라미터 변환에 의존한다는 것이다. 즉, 특징벡터를 생성하거나 변환하는 작업을 하지 않기 때문에 직접적으로 성능을 비교하는데 있어서 보다 객관적인 연구결과를 얻을 수가 있다.

좀더 상세하게 제안된 방식의 성능을 비교하기 위해서, HMM의 파라미터값들을 순차적으로 보상하였다. 즉. 먼저, 켑스트럼(MFCC)의 평균벡터를 보상하고 그 다음에 델타 켑스트럼(delta-MFCC)의 평균벡터를 함께 적응하였으며, 마지막으로 MFCC의 공분산 행렬에 대한 적응을 추가하였다. delta-MFCC에 대한 공분산 적응은 성능에 거의 영향을 주지 못할 것으로 생각되어 수행하지 않았다.

먼저, 켑스트럼 평균벡터에 대해서만 적응을 한 경우에, 제안된 방식은 log-normal PMC방식에 비해서 SNR 값이 0 dB에서 성능이 우수한 것으로 나타났다. 이러한 성능 향상을 좀 더 분명하게 관찰하기 위해서, log-likelihood 계산과정에서 정적 MFCC 특징 벡터만을 사용한 경우의 인식률을 괄호 안에 나타내었다. 이것은 delta-MFCC에 대한 HMM 파라미터가 아직까지 적응이 안 된 상태이므로, 이들이 log-likelihood 계산과정에 미치는 영향을 무시하기 위해서이다. 괄호 안의 인식률을 비교해보면, SNR 값이 0 dB에서, 제안된 방식이 훨씬 향상된 인식성능을 보여줌을 뚜렷이 알 수 있다. 하지만, SNR 값이 0 dB 보다 큰 경우에는 뚜렷한 성능차이를 볼 수 없었는데, 이것은 제안된 방식이 SNR 값이 낮은 경우에 특히 효과적임을 나타낸다.

한편, 기존의 log-normal PMC 방식에서는 delta-MFCC의 평균벡터를 보상한 경우에 인식성능

의 향상을 보여 주지 못했다. 이것은 식(20)에서 사용된 근사화가 정확하지 않음으로 인해서 발생하는 문제점인 것으로 생각된다. 하지만, 본 논문에서 제안된 방식에서는 delta- MFCC의 평균 벡터를 보상함으로써 큰 성능 향상을 볼 수 있었다. 제안된 방식에서는 PMC에서와 같은 통계적 가정을 사용하지 않았고 적응에 필요로 하는 통계정보를 직접 훈련 과정중에서 얻음으로서 이러한 인식성능의 향상이 가능해진 것이라 생각된다.

평균벡터의 변환과 더불어서 공분산행렬의 적응을 수행한 경우에는 제안된 방식과 log- normal PMC방식 모두에 있어서 인식성능의 뚜렷한 향상을 볼 수 없었다. 이와 같이 평균벡터의 보상과 달리 공분산 행렬의 보상에서는 인식성능이 향상되지 못하는 경우가 가끔 보고 되고 있다[2]. 일반적으로 추정된 공분산 행렬 값들이 부가잡음의 영향으로 인하여 원래의 공분산 값보다 다소 작게 나타나게 되는데, 이러한 작아진 공분산값은 화자나 잡음환경의 변이 등에 대해서 다소 강인함이 떨어지는 것으로 생각된다. 그리고 공분산의 추정에 있어서는 평균벡터의 적응에서 보다 많은 양의 적응데이터가 필요하게 된다. 특히, 이러한 경향은 log-normal PMC에서 강하게 나타나는데, log-normal PMC의 경우에는 입력음성의 묵음구간에서 추출된 적은 양의 잡음샘플을 가지고서는 신뢰성 있는 공분산을 추정하기가 매우 힘든 것으로 생각된다.

또한, 표 1의 결과를 보면 평균벡터의 보상만을 통해서도 이미 재훈련 때의 인식성능과 필적 할 만한 인식성능을 얻을 수 있는 것을 알 수 있다. 따라서, 공분산 행렬의 적응을 통해서 추가적으로 얻을 수 있는 인식성능의 향상은 그리 크지 않을 것이라고 생각된다.

표 1의 결과에서 보면, 제안된 방식에서는 공분산의 보상시에 인식성능의 변화가 거의 없었으나, log-normal 방식에서는 SNR 값이 0 dB인 경우에 인식성능이 오히려 상당히 떨어지는 것을 알 수 있었다. 이러한 공분산 행렬의 보상시에 나타나는 강인성은 제안된 알고리듬의 또 다른 장점이라고 생각된다. 이러한 강인성은 log-normal PMC에서는 공분산 행렬을 추정하기 위해서 추정 오차값이 클 수 있는 잡음신호의 분산값을 직접적으로 이용하는데 비해서, 제안된 방식에서는 잡음신호를 포함하는 다양한 통계정보를 신뢰성 있게 추정하여 이를 인식시에 사용하는데 따른 것으로 생각된다.

4. 결 론

본 논문에서는 잡음음성인식을 위한 직접데이터 기반의 HMM 파라미터 보상방식을 제안하였다. 제안된 방법에서는 훈련과정을 통해서 직접 음성신호와 잡음신호로부터 다양한 통계정보를 추정하며, 이후 이를 인식시에 이용함으로서 기존의 PMC 방법 등에서 사용되던 부정확한 통계 가정을 사용하지 않아도 되는 장점이 있다. 또한 테일러 전개식을 사용함으로서 HMM의 평균벡터 뿐만 아니라 공분산 행렬의 적응도 추가적인 적응데이터가 필요 없이 수행할 수 있었다. 이러한 직접 데이터 기반의 적응방식은 잡음에 의한 음성신호의 변이를 좀 더 정확하게 모델파라미터에 반영할 수 있는 장점이 있다고 생각된다. 전체적으로 제안된 방식은 기존의 PMC 방식에 비해서 우수한 성능을 보였으며, 특히, 차분특징 파라미터의 변환이나 공분산 행렬의 변환에 있어서는 더욱 많은 향상을 보여주었다.

참 고 문 헌

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

[2] Moreno, P. 1996. Speech Recognition in Noisy Environments. PhD Thesis. Carnegie Mellon Univ., April.

[3] 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.

[4] 정용주. 2002. "결정적 잡음 모델을 이용한 효율적인 잡음음성 인식 접근방법." *한국음향학회지, 제21권 제6호*, pp. 559-565, 8월.

[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, Jan.

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

[8] Labiner, L. R., & Juang, B.-H. 1993. Fundamentals of Speech Recognition, Prentice Hall.

접수일자: 2004. 3. 20
게재결정: 2004. 6. 15

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