한국음향학회지 제21권 제6호 pp. 559~565 (2002)

결정적 잡음 모델을 이용한 효율적인 잡음음성 인식 접근 방법

An Efficient Approach for Noise Robust Speech Recognition by Using the Deterministic Noise Model

정 용 주*
(Yong-Joo Chung*)

*계명대학교 컴퓨터 · 전자공학부
(접수일자: 2002년 5월 14; 채택일자: 2002년 7월 25)

본 논문에서는 잡음음성 HMM (Hidden Markov Model)의 파라미터 값을 효율적으로 추정하는 새로운 방법에 대해서 제안하였다. 기존의 방법들에서 잡음음성의 HMM 파라미터값을 추정하기 위해서는 먼저 잡음음성의 생성 모델을 가정한 후, 잡음과 원래 음성의 통계 모델을 이용하여 잡음음성 HMM 파라미터값을 해석적으로 얻게 된다. 하지만 이러한 해석적 방법은 항상 단순화의 가정을 취하게 되므로 실제의 잡음음성 HMM 분포에 정확히 근접하는데 어려움을 겪게 된다. 본 연구에서는 이러한 가정을 하지 않고, 원래의 깨끗한 음성에서 얻을 수 있는 HMM의 파라미터값을 사용하고 결정적 잡음 모델을 이용함으로서 기존의 방법보다 인식시에 계산량을 줄일 수 있었을 뿐만 아니라 인식 성능의 향상도 이룰 수 있었다.
핵심용어: 음성인식, 잡음음성처리, HMM
투고분야: 음성처리 분야 (2.5, 2.6)

In this paper, we proposed an efficient method that estimates the HMM (Hidden Markov Model) parameters of the noisy speech. In previous methods, noisy speech HMM parameters are usually obtained by analytical methods using the assumed noise statistics. However, as they assume some simplication in the methods, it is difficult to come closely to the real statistics for the noisy speech. Instead of using the simplication, we used some useful statistics from the clean speech HMMs and employed the deterministic noise model. We could find that the new scheme showed improved results with reduced computation cost.
Keywords: *Speech recognition, Noisy speech processing, HMM*
ASK subject classification: *Speech signal processing (2.5, 2.6)*

I. 서 론

잡음음성 (noisy speech) 인식에 관한 연구는 실용적인 음성인식시스템의 구성에서 중요한 이슈가 되고 있으며, 이 방면의 연구자들의 노력에 의해서 최근에도 많은 연구 결과들이 발표되고 있다. 잡음음성인식을 위한 대표적 방법 중의 하나는 음성신호의 분석을 이용하여 잡음음성 신호로부터 원래음성신호를 복원하는 방식이다[1]. 그러나 최근에 들어와서 또 다른 접근 방법인 인식기모델 보상을 이용한 잡음음성인식 방식이 제안되어 좋은 결과를 보여주고 있다[2,3]. 인식기모델보상 방식은 입력음성신호 자체를 추정하는데 초점을 두지 않고 이미 훈련된 음성인식기의 모델 파라미터를 잡음의 특성에 따라서 변환 시켜

책임저자: 정용주 (yjjung@kmu.ac.kr)
704-919 대구시 달서구 신당동 1000번지
계명대학교 컴퓨터 · 전자공학부
(전화: 053-580-5925)

줌으로서 인식기의 파라미터값이 주어진 환경에 최대한 적합하도록 유도한다. 인식기모델변환 방식 중에서 대표적인 방식에는 음성/잡음 분해법 (SND: speech and noise decomposition), 병렬모델 결합법 (PMC: parallel model combination), 테일러 전개방식 (VTS: vector Taylor series), 통계적 재추정법 (STAR: Statistical re-estimation) 등이 있다.

그 중에서도 PMC 방식에서는 소량의 잡음 샘플을 이용하여 훈련된 인식기모델을 변환할 수 있는데, 인식시에 적은 계산량으로도 매우 향상된 성능을 보여주었다. 특히, PMC는 VTS 나 STAR 방식과 비교해서 소요되는 계산량이 매우 적을 뿐 아니라 적응을 위한 잡음음성데이터를 따로 필요로 하지 않는다는 면에서 큰 장점이 있는 방식이라 생각된다. 하지만 PMC 방식은 그 유도과정에서 비교적 많은 가정을 사용함으로서 그 결과의 정확도가 다소 떨어지는 문제점을 가지고 있다고 생각된다.

본 연구에서는 PMC 방식에서처럼 적은 계산량으로 처리가 가능하며 따로 적응데이터가 필요로 하지 않으면서도, PMC 방식의 단점인 확률분포에 대한 단순화 가정을 거치지 않는 새로운 인식기모델 보상방식을 제안하고자 한다. 이 제안된 방식에서는 미리 훈련된 음성 HMM이 원래 음성신호에 대한 최적의 모델을 제공한다는 생각에 기반을 두며, 잡음이 부가된 이후에도 이러한 음성 HMM의 기본구조가 크게 바뀌지 않을 것이라는 가정에 근거를 둔다. 또한 확률적 잡음모델이 아닌 결정적 잡음모델을 적용함으로서 인식기모델보상이 비교적 간단히 이루어지도록 하였다.

본 논문의 구성은 2장에서 기존에 제안된 PMC 방식을 이용한 잡음음성 인식에 관해서 간략히 소개하고 3장에서는 제안된 결정잡음 모델을 이용한 인식기모델보상 방식에 대해서 설명하며 4장에서 그 실험 결과에 대해서 비교 분석하고 5장에서 결론을 맺는다.

II. PMC 방식의 개요

본 연구에서는 제안된 방식의 성능을 기존의 방식중 PMC와 비교하고자 한다. 따라서 이번 장에서는 PMC 방식에 대해서 간략히 소개한다.

PMC 방식은 그 세부적인 구현과정에 따라서 로그 정규분포, 로그 뎃셈 (log-add) PMC 그리고 데이터 적응 PMC 방식 등이 있다. 로그 정규 분포 방식에서는 켑스트럴 영역의 HMM 파라미터들을 선형주파수 영역으로 변환하는 과정이 먼저 이루어진다. 잡음이 섞인 음성 (noisy speech) HMM모델을 만들기 위해서 원래의 깨끗한 음성 (clean speech) HMM 파라미터 값과 잡음 (noise)에 대한 HMM 파라미터 값을 선형주파수 영역에서 서로 결합하여 준다.

이를 대한 자세한 과정은 다음과 같이 요약된다.

1) 로그스펙트럼 영역의 HMM 파라미터 값을 구하기 위해서 역DCT (Discrete Cosine Transformation) 변환을 [4] 취한다.

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

여기서 μ^l과 Σ^l는 로그 스펙트럼 영역에서의 평균벡터와 공분산 행렬이며 μ^c와 Σ^c는 켑스트럼 영역에서의 값이다.

2) 로그스펙트럼 영역의 HMM 파라미터값을 선형영역으로 변환한다.

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

여기서 μ_i와 Σ_{ij}는 선형영역의 파라미터 값인 평균벡터 μ와 공분산행렬 Σ의 구성 원소이다.

3) 음성과 잡음의 HMM 파라미터 값을 결합한다.

$$\hat{\mu} = \mu + \bar{\mu},\quad \hat{\Sigma} = \Sigma + \bar{\Sigma} \qquad (3)$$

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

4) 결합된 HMM 파라미터값에 대해서 다음과 같이 로그변환과 DCT 변환을 취함으로서 최종적으로 켑스트럼 영역에서의 잡음음성의 평균벡터 $\hat{\mu}^c$와 공분산행렬 $\hat{\Sigma}^c$이 얻어진다.

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

$$\hat{\Sigma}_{ij}^l = \log(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\,\hat{\mu}_j} + 1) \qquad (4)$$

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

로그 뎃셈 (log-add) PMC 방식은 위의 식 (2), (4)에서 잡음음성 HMM 평균벡터 값을 구하는 과정에서 잡음과 음성신호HMM의 공분산 값이 충분히 작다는 가정을 함으로서 변환공식이 다음과 같이 단순화된다.

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

이와 같이 로그영역의 평균벡터값을 구한 다음 위의 식 (5)을 이용하여 켑스트럼 영역의 평균벡터를 구함으로서 로그 정규 분포 방식에 비해서 훨씬 간단하게 계산과정을 마칠 수 있다.

위의 변환과정의 식 (1)에서 보면 역DCT 변환을 취하게 된다. 그러나 DCT변환은 켑스트럼영역과 로그스펙트럼영역 사이의 변환을 설명하는데, 일반적으로 두 영역에서의 특징벡터의 공간 차원 (space dimension)이 다르므로 정확한 역변환을 수행할 수 없을 것이다. 따라서 식 (1)을 이용한 역변환에서 주어진 켑스트럼 평균벡터로부터 얻어지는 로그스펙트럼 평균벡터의 값이 원래의 DCT 변환하기 전의 값과 같지 않을 수 있음을 의미한다 (즉, $C^{-1}C \neq I$). 즉 그림 1에서 보인 바와 같이 여러 개의 로그스펙트럼 값이 동일한 켑스트럼값으로 변환이 된 후 다시 역변환을 취할 경우 다시 원래의 값으로 되돌아 간다는 보장이 없을 것이다.

이와 같은 현상은 평균벡터와 같이 약간의 변화만으로도 인식율에 미치는 영향이 매우 큰 경우에 중요한 문제가 될 것으로 생각된다. 또한 위의 변환과정의 식 (2)와 식 (4)에서는 로그스펙트럼 영역에서의 특징벡터의 분포를 가우시안 함수로 가정하고 있다. 하지만 최근의 연구결과 이러한 가정은 실제 데이터값과 차이를 나타내는 것으로 알려졌다[5].

이와 같이 PMC 방식은 계산량이 적고 또한 적응데이터가 따로 필요하지 않으면서도 우수한 성능을 나타내는 좋은 방법이지만, 변환과정에서 단순화 가정을 함으로서 실제 인식성능이 떨어지는 단점을 가지고 있다. 따라서 이러한 가정을 사용하지 않으면서도 인식시 계산량이 증가하지 않는 알고리듬을 개발하는 것이 중요하다고 생각되며 다음 장에서 제안된 알고리듬을 소개하고자 한다.

III. 결정적 잡음모델과 훈련시의 통계치를 활용한 인식 방식

3.1. 잡음음성의 생성 모델링

강인한 음성인식을 위해서는 실제 상황에서 음성신호가 어떠한 왜곡을 거치는가 하는 것을 정확히 파악하고 모델링하는 것은 중요하다. 일반적으로 음성신호를 왜곡시키는 요인으로는 채널변이 등으로 대표되는 콘벌루션 잡음과 부가잡음이 존재한다. 본 연구에서 관심을 갖고 있는 부가잡음의 영향만이 존재하는 경우에는 음성신호와 잡음신호가 시간영역상에서 단순히 더해지는 것으로 생각할 수 있다. 그것을 퓨리에 (Fourier) 변환을 거친 스펙트럼 영역에서 본다면 다음과 같이 나타낼 수 있다.

$$Y_i = X_i + N_i \tag{6}$$

여기서 Y_i는 잡음음성에 대한 멜주파수 (mel-frequency) 기반의 필터뱅크 출력중에서 i번째 출력을 나타내며,

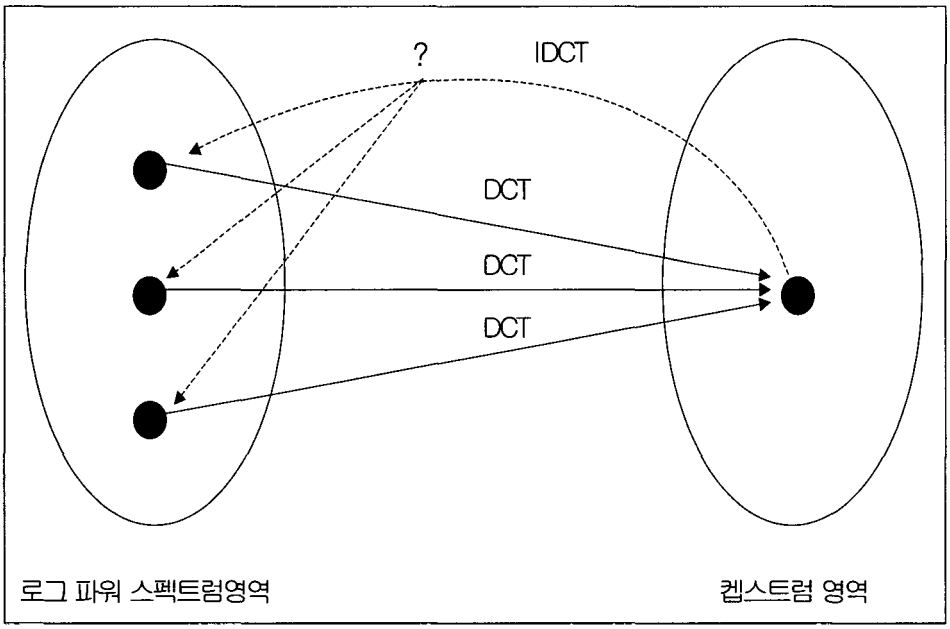

그림 1. DCT 변환과 그 역변환 (IDCT)의 관계
Fig. 1. The relationship between the DCT and inverse DCT.

X_t, N_t은 각각 음성신호와 잡음에 관한 것을 나타낸다. 여기서는 잡음과 음성신호의 스펙트럼간의 상관관계(Correlation)는 무시할 수 있다고 가정한다. 위에서 표시한 필터뱅크 출력으로부터 음성인식을 위해서 사용하는 켑스트럼 출력을 얻기 위해서는 앞절에서도 언급한 바 있는 로그 변환을 거친 후 DCT 변환을 수행해야 한다.

즉, 켑스트럼 출력 벡터 Y^c는 다음과 같이 나타난다.

$$Y^c = C \log Y \tag{7}$$

따라서, i 번째 켑스트럼 원소는 다음과 같다.

$$Y_i^c = \sum_{k=1}^{N} \log Y_k \cos\left[i(k-\frac{1}{2})\frac{\pi}{N}\right], \; i=1,2,\cdots,M \tag{8}$$

여기서 $Y=(Y_1, Y_2, Y_3, \cdots, Y_N)$이며, N은 사용된 필터뱅크 필터의 개수이다. DCT 변환 행렬 C는 $M \times N$ 행렬을 나타내고 M은 켑스트럼 특징벡터의 차수 (dimension)가 된다. 위의 식 (7)에 식 (6)의 잡음과 음성과의 선형결합관계식을 대입하면 다음과 같다.

$$\begin{aligned} Y^c &= C \log (X+N) \\ &= C \log X + C \log(i + \exp(\log N - \log X)) \end{aligned} \tag{9}$$

여기서 i는 모든 원소의 값이 1 인 단위벡터이다. 식 (9)는 잡음음성의 생성 모델을 나타낸다. PMC 방식의 경우는 잡음 N과 음성 X에 대해서 확률적 분포를 가정한다. 이렇게 함으로서 잡음음성인 Y^c에 대한 확률분포를 해석적으로 얻고자 한다. 하지만 앞절에서도 언급되었듯이 이 과정에서 야기된 단순화에 의한 오차를 극복하지 못하므로, 본 연구에서는 다소 새로운 개념으로 이 문제에 접근하려고 한다. 이에 대해서 다음 절에서 설명한다.

3.2. 결정적 잡음모델과 훈련시의 통계치를 이용한 보상방식

연속밀도 HMM의 평균벡터와 분산행렬을 추정하는 일반적인 방법은 Baum-Welch 알고리듬을 따른다[6]. 만약 비터비 (Viterbi) 알고리듬에 의해서 각 관측벡터들이 특정한 HMM 상태와 일대일 관계를 갖는 것이 알려진다면, 그 추정 공식은 아래와 같다.

$$\mu_{jk}^c = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\, X_t^c}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{10}$$

$$\Sigma_{jk}^c = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\,(X_t^c - \mu_{jk}^c)(X_t^c - \mu_{jk}^c)'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{11}$$

여기서, $\gamma_t(j,k)$는 시간 t에서 관측벡터가 상태 j를 점유하면서 mixture 성분 k에 의해서 발생될 확률을 의미한다.

$$\gamma_t(j,k) = \frac{c_{jk}\, N(X_t^c, \mu_{jk}^c, \Sigma_{jk}^c)}{\sum_{m=1}^{M} c_{jm}\, N(X_t^c, \mu_{jm}^c, \Sigma_{jm}^c)} \tag{12}$$

여기서 N은 평균벡터가 μ_{jk}^c이고 분산행렬이 Σ_{jk}^c 인 가우시안 분포를 나타낸다. 위의 식 (10)~(12)에 의해서 HMM 파라미터 값들이 추정되며, 이를 인식시에 이용함으로서 입력신호와 가장 근접한 단어나 문장을 찾아가게 된다.

위의 과정을 통해서 얻어진 HMM의 평균벡터나 분산행렬은 인식기 훈련을 위해서 준비된 음성에 적합하도록 추정된다. 하지만 인식시에 음성이 잡음에 의해서 왜곡되었을 때는 위에서 추정된 HMM 의 평균벡터나 분산행렬에는 변화가 있을 것이다. 이때 변화되는 파라미터값은 어떻게 음성신호 X_t^c가 변화되느냐에 달려있다고 생각된다. 따라서 음성신호 X_t^c가 식 (9)에서처럼 변형이 된다는 가정이 맞다면 식 (10), (11)은 다음과 같이 변형될 것이다.

$$\begin{aligned} &\hat{\mu}_{jk}^c = \\ &\frac{\sum_{t=1}^{T} \gamma_t(j,k)(C \log X_t + C \log(i + \exp(\log N - \log X_t)))}{\sum_{t=1}^{T} \gamma_t(j,k)} \\ &= \mu_{jk}^c + \frac{\sum_{t=1}^{T} \gamma_t(j,k)(C \log(i + \exp(\log N - \log X_t)))}{\sum_{t=1}^{T} \gamma_t(j,k)} \\ &= \mu_{jk}^c + E\big(C \log(i + \exp(\log N - \log X_t))\big) \\ &= \mu_{jk}^c + \mu_{jk}^n \end{aligned} \tag{13}$$

$$\begin{aligned} &\hat{\Sigma}_{jk}^c = \Sigma_{jk}^c + E \\ &\big((X_t^c - \mu_{jk}^c)(C \log(i + \exp(\log N - \log X_t)) - \mu_{jk}^n)'\big) \\ &+ E\big(\big(C \log(i + \exp(\log N - \log X_t)) - \mu_{jk}^n\big) \cdot \\ &\quad \big(C \log(i + \exp(\log N - \log X_t)) - \mu_{jk}^n\big)'\big) \end{aligned} \tag{14}$$

위의 식 (13), (14)에서 우리는 잡음에 의해서 변화된 음성신호 HMM의 평균벡터와 분산행렬에 관한 수식을 유도했다. 하지만 실제로 그 값들을 얻기 위해서는 잡음신호에 대한 정보를 가지고 있어야 한다. 예를 들어, 평균벡터를 구하기 위해서는 μ_{jk}^n를 구해야 할 것이다.

PMC에서는 잡음신호 N을 랜덤신호로 간주하여 잡음의 HMM 파라미터로부터 해석적인 방법으로 잡음음성의 HMM 파라미터값을 구하였다. 본 연구에서는 잡음신호 N을 결정신호라고 가정한다. 물론 이 가정은 약간의 오류를 가지고 있으나, 실제 잡음신호에 대한 통계치를 인식음성으로부터 충분히 정확히 얻는다는 것이 어렵기 때문에 인식성능에 미치는 영향이 크지 않으리라 생각된다. 따라서 식 (13)의 μ_{jk}^{n}를 얻는 방법에 수정을 다음과 같이 하였다.

$$\mu_{jk}^{x} = E(X_t)$$
$$\mu_{jk}^{n} = C\ \log(i + \exp(\log N - \log \mu_{jk}^{x}))$$

즉, 멜주파수 필터뱅크 출력값 X_t에 대한 평균값을 구한 다음 이를 μ_{jk}^{n}를 구하는 식에 대입한다. 이렇게 함으로서 인식기 훈련시 사용 가능한 많은 양의 음성데이터를 이용하여 μ_{jk}^{x}를 구하고, 인식시에 얻을 수 있는 잡음신호를 이용하여 효과적으로 잡음에 대한 보상을 이룰수 있을 것으로 생각된다.

공분산행렬에 대해서는 잡음평균벡터와 같이 보상을 할 수 없을 것이다. 왜냐하면 이를 수행하기 위해서는 미리 μ_{jk}^{n}값을 구해야 하는데, 이는 훈련과정에서 얻을 수 없는 정보이기 때문이다. 따라서 공분산행렬의 보상에는 잡음신호에 대한 적응데이터가 필요하다.

IV. 실험결과

4.1. 기반 인식시스템

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

그리고 본 논문에서는 인식 성능이 비교적 우수하여 많이 사용되고 있는 13차의 MFCC (mel-frequency cepstrum coefficients)을 특징벡터로서 이용하였다.

4.2. 기본 인식 실험

본 절에서는 먼저 기본인식기의 성능에 대해서 먼저 검토하고자 한다. 표 1에서 우리는 기본인식기의 인식 성능을 자동차 잡음 환경하에서의 신호대 잡음비 (SNR: signal to noise ratio (단위: dB))가 달라짐에 따라서 어떻게 변화하는지 나타내었다.

기본인식기는 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 신호대 잡음비가 20 dB 이상인 경우는 인식율의 저하가 그리 심하지는 않으나, 신호대 잡음비가 10 dB 이하인 경우는 매우 심각한 정도의 인식율의 저하를 가져옴을 알 수 있다. 따라서 신호대 잡음비가 매우 낮은 영역에서의 인식율을 향상시키는 방법은 잡음음성인식에서의 성공적 수행을 위해서 매우 중요하다 할 것이다.

한편 표 2에서는 인식시의 음성의 신호대 잡음비값이 훈련시에 사용된 음성과 동일한 신호대 잡음비값을 가질 경우 (일치된 환경)의 인식율을 나타내어 보았다.

이 경우에는 표 1과 대비하여 인식성능이 많이 향상됨을 알 수 있다. 이것은 이미 훈련시에 잡음의 영향이 HMM의 파라미터들의 값에 충분히 반영되기 때문이라고 생각된다. 따라서 이 결과는 잡음보상 인식알고리듬의 개발에

표1. 기본인식기의 잡음음성에 대한 인식율
Table 1. Recognition rates of the baseline recognizer for noisy speech.

인식환경	0 dB	10 dB	20 dB	clean
인식율 (%)	32.5	71.3	88.7	94.7

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

인식환경	0 dB	10 dB	20 dB	clean
인식율 (%)	84.4	91.8	94.3	94.7

있어서 중요한 벤치마크 결과하고 생각된다. 그러나 이러한 인식성능을 얻기 위해서는 항상 변화하는 잡음환경에 대한 충분한 양의 음성 데이터베이스를 가지고 모델을 미리 훈련시켜야 하는 어려움이 있다.

4.3. 제안된 알고리듬의 성능 분석

본 절에서는 제안된 알고리듬을 기존의 PMC 방식과 비교 검토하고자 한다.

표 3에는 기존의 PMC 알고리듬들의 결과와 함께 제안된 방식 (Direct Adaptation)의 결과를 신호대 잡음비값의 변화에 따라서 나타내었다. PMC 로그 정규 분포 방식은 2장에서 설명한 그대로의 방식을 의미하며, PMC 로그 덧셈은 로그 정규 분포 방식을 구현하기 쉽도록 간단하게 만든 방식이다. 로그 덧셈 방식에서는 분산 (variance)에 대한 보상은 이루어지지 않는다. 먼저 로그 정규 분포 방식을 평균과 분산에 대한 보상을 동시에 해주는 경우와 평균에 대해서만 보상을 해주는 경우로 나누어서 인식실험을 해보았다. 이 경우 분산을 보상해주는 경우가 그렇지 않은 경우에 비해서 오히려 인식성능이 떨어지는 것을 알 수 있었다. 이것은 두가지 경우로 설명이 될 수 있을 것이다.

첫 번째 이유로는 잡음 HMM을 구성하는 과정에서 이용되는 잡음신호의 샘플수가 너무 작아서 충분히 잡음의 통계치를 얻을 수 없다는 것이다. 이것은 PMC 과정에서 인식시의 입력음성으로부터 잡음 샘플을 추출하므로 생기는 문제이다. 본 연구에서는 음성구간 시작 전의 3 내지 5개 정도의 음성특징벡터을 추출하여 잡음 HMM을 구성하였다. 따라서 잡음의 분산값이 다소 정확히 추정되지 못한 것으로 생각된다.

두 번째 이유로는 추정된 잡음음성의 분산은 원래의 분산에 비해서 줄어드는 경향이 있으므로 실제 인식시에 화자의 바뀜이나 잡음특성의 변이 등에 대한 강인함이 줄어들 수 있는 가능성을 생각할 수 있다. 한편, 로그 덧셈 방식은 평균벡터만을 보정하는데, 그 계산량 측면에서는 로그 정규 분포 방식에 비해서 훨씬 적게 들지만, 그 성능에서는 큰 차이가 없음을 알 수 있었다. 한편 제안된 직접 적응 (DA: Direct Adaptation) 방식에서는 평균과 분산을 동시에 보정하는 방식과 평균만을 보정하는 방식 모두에 대해서 인식실험을 해 보았다. 평균에 대해서만 보상을 해 준 경우에 이전의 PMC 방식에 비해서 다소 좋은 성능을 보임을 알 수 있었다. 특히 0 dB에서는 재훈련 (Retraining)에 의한 인식율이 84.4%이고 기존의 PMC 방식이 82.7%였으나, 제안된 DA 방식은 83.6%로서 재훈련 방식을 벤치마크 인식율로 보았을 때, 기존의 PMC와 벤치마크 결과와의 차이를 50% 가까이 줄이는 효과가 있음을 알 수 있었다. 직접 적응 방식에서는 평균외에도 분산에 대해서도 보상한 경우의 인식실험을 하였는데, 성능의 변화를 별로 볼 수 없었다. 이것은 분산의 보상을 하기 위해서는 미리 적응데이터을 이용하였는데, 이때 이용된 적응데이터는 순수하게 잡음신호만으로 이루어져 있어서 잡음이 음성신호에 미치는 영향을 분산의 적응과정에 충분히 이용할 수 없는 한계가 있다고 생각된다. 그리고 적응데이터와 실제 인식시에 존재하는 잡음신호와는 다르므로 이러한 차이가 원인이 될 것으로 생각된다.

제안된 방식은 기존의 PMC에 비해서 성능이 뛰어날 뿐만 아니라, 소요되는 계산량도 훨씬 작다는 것을 알 수 있다. 특히 평균값을 변환하는 경우, 계산량이 비교적 작은 로그 덧셈 PMC 방식에 비해서도 약 1/2 정도의 계산량이 필요하였다. $M \times N$의 DCT 변환행렬을 사용하였을 경우, 로그 정규 분포 방식의 경우 역DCT 변환을 위해서는 $N \cdot M + N \cdot M + N \cdot M^2$의 곱셈이 필요하며, 선형영역/로그영역 상호 변환시에는 각각 $2 \cdot N^2$ 그리고 최종적으로 DCT 변환시에 $M \cdot N + M^2 \cdot N + M \cdot N$의 곱셈이 필요하다.

로그 덧셈 방식에서는 분산에 대한 변환이 필요치 않으므로 역DCT 변환에 $N \cdot M$의 곱셈만이 필요하고 DCT 변환시에 $M \cdot N$ 곱셈이 필요하다. 한편 제안된 직접 적응 방식에서는 역DCT 변환이 필요하지 않으므로 DCT 변환시에 $M \cdot N$의 곱셈만이 소요된다.

$M = 13$, $N = 18$의 차수를 사용한 경우 로그 정규 분포, 로그 덧셈, 직접 적응 방식에서의 곱셈의 수는 각각, 7668, 468, 234이다.

표 3. 제안된 방식 (DA)과 기존의 PMC방식들과의 성능비교
Table 3. Performance comparison between the proposed method (DA) and the conventional PMC methods.

인식환경	0 dB	10 dB	20 dB
PMC log-normal (mean only)	82.6 (%)	90.8 (%)	93.8 (%)
PMC log-normal (mean & variance)	73.9 (%)	89.8 (%)	93.1 (%)
PMC log-add	82.7 (%)	90.6 (%)	93.7 (%)
DA (mean only)	83.6 (%)	91.3 (%)	93.9 (%)
DA (mean & variance)	83.6 (%)	91.5 (%)	93.9 (%)
Retraing	84.4 (%)	91.7 (%)	93.9 (%)

V. 결 론

본 연구에서는 잡음에 강인한 음성인식을 위해서 훈련된 인식기의 HMM 파라미터값을 보상하는 새로운 방식에 대해서 제안하였다. 제안된 알고리듬은 HMM의 훈련시에 얻을 수 있는 통계치를 충분하게 활용하고 또한 결정적 잡음모델을 이용함으로서 잡음음성에 대한 HMM 파라미터값을 효율적으로 얻을 수 있었다. 이 방식은 기존의 PMC 등의 방법에서 잡음의 통계치를 기반으로 하여 잡음음성의 모델파라미터 값을 얻는 해석적 방식이 단순화의 오류를 가지고 있는 점을 극복할 수 있으며, 계산량도 기존의 방식에 비해서 줄일 수가 있었다.

참 고 문 헌

1. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, 27 (2), 113–120, 1979.
2. M. Gales and S. Young, *Parallel Model Combination for Speech Recognition in Noise*, Tech. Rep. 135, Cambridge University, June 1993.
3. P. Moreno, *Speech Recognition in Noisy Environments*, PhD Thesis, Carnegie Mellon Univ., April, 1996.
4. Davis S. B. and Mermelstein P. "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, 28, 357–366, 1980.
5. J-W. Hung, J-L. Shen and L-S. Lee, "New approaches for domain transformation and parameter combination for improved accuracy in parallel model combinatin (PMC) techniques", *IEEE Trans. Speech and Audio Processing*, 9 (8), 842–855, 2001.
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.*, 41, 164–171, Jan. 1970.
7. 정용주, "Bayesian 적응 방식을 이용한 잡음음성인식에 관한 연구", 한국음향학회지 20 (2), 21–26, 2001.

저자 약력

● 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 전자공학과
※ 주관심분야: 음성인식, 신호처리, 패턴인식