한국음향학회지 제23권 제6호 pp. 473~480 (2004)

시변 잡음에 대처하기 위한 다중 모델을 이용한 PCMM 기반 특징 보상 기법

PCMM-Based Feature Compensation Method Using Multiple Model to Cope with Time-Varying Noise

김 우 일*, 고 한 석*
(Wooil Kim*, Hanseok Ko*)

*고려대학교 전자컴퓨터공학과

(접수일자: 2004년 2월 9일; 채택일자: 2004년 8월 20일)

본 논문에서는 잡음 환경에서 강인한 음성 인식을 위하여 음성 모델을 기반으로 하는 효과적인 특징 보상 기법을 제안한다. 제안하는 특징 보상 기법은 병렬 결합된 혼합 모델 (PCMM)을 기반으로 한다. 기존의 PCMM 기반의 기법은 시간에 따라 변하는 잡음 환경을 반영하기 위하여 매 음성 입력마다 복잡한 과정의 혼합 모델 결합이 필요하다. 제안하는 기법에서는 다중의 혼합 모델을 보간하는 방법을 채용함으로써 시간에 따라 변하는 배경 잡음에 대응할 수 있다. 보다 신뢰성 있는 혼합 모델 생성을 위하여 데이터 유도 기반의 방법을 도입하고, 실시간 처리를 위하여 프레임에 동기화된 환경 사후 확률 예측 과정을 제안한다. 다중 모델로 인한 연산량 증가를 막기 위하여 혼합 모델을 공유하는 기법을 제안한다. 가우시안 혼합 모델 사이에 통계학적으로 유사한 요소들을 선택하여 공유에 필요한 공통 모델을 생성한다. Aurora 2.0 데이터베이스와 실제 자동차 주행 환경에서 수집된 음성 데이터베이스에 대한 성능 평가를 실시한다. 실험 결과로부터 제안한 기법이 모의 환경과 실제 잡음 환경에서 강인한 음성 인식 성능을 가져오고 연산량 감소에 효과적임을 확인한다.

핵심용어: 음성 인식, 모델 기반 특징 보상, 병렬 결합된 혼합 모델, 다중 모델 보간, 혼합 모델 공유

투고분야: 음성처리 분야 (2.5)

In this paper, we propose an effective feature compensation scheme based on the speech model in order to achieve robust speech recognition. The proposed feature compensation method is based on parallel combined mixture model (PCMM). The previous PCMM works require a highly sophisticated procedure for estimation of the combined mixture model in order to reflect the time-varying noisy conditions at every utterance. The proposed schemes can cope with the time-varying background noise by employing the interpolation method of the multiple mixture models. We apply the 'data-driven' method to PCMM for more reliable model combination and introduce a frame-synched version for estimation of environments posteriori. In order to reduce the computational complexity due to multiple models, we propose a technique for mixture sharing. The statistically similar Gaussian components are selected and the smoothed versions are generated for sharing. The performance is examined over Aurora 2.0 and speech corpus recorded while car-driving. The experimental results indicate that the proposed schemes are effective in realizing robust speech recognition and reducing the computational complexities under both simulated environments and real-life conditions.

Keywords: Speech recognition, Model-based feature compensation, Parallel combined mixture model, Multiple model interpolation, mixture sharing

ASK subject classification: Speech signal processing (2.5)

책임저자: 고 한 석 (hsko@korea.ac.kr)
136-701 서울시 성북구 안암5가-1
고려대학교 전자컴퓨터공학과
(전화: 02-3290-3239; 팩스: 02-3291-2450)

I. 서 론

음성 인식 시스템의 성능을 하락시키는 주요한 원인은 인식에 사용되는 음향 모델 파라미터의 훈련을 위해 사용되는 음성 데이터베이스와 비교하여 실제 인식 환경에서 입력되는 음성 데이터의 음향학적 특성이 달라진다는 것이다. 음향학적 차이를 일으키는 요인으로 인식 시스템 주변의 배경 잡음에 의한 음성 신호의 오염, 마이크 및 전송 선로와 같이 음성 신호의 전달과정에서 생기는 채널의 변화 또는 왜곡, 화자나 발음 방법의 변화, 인식 어휘의 변화 등을 들 수 있다. 실제 환경에서의 음성 인식 성능을 최대화하려는 노력의 하나로서 이와 같은 요인에 의해 발생되는 인식 환경과 훈련 환경의 차이를 줄이기 위한 연구와 기술 개발이 최근 수년간 집중되고 있다.

훈련 환경과 인식 환경의 음향학적 차이를 줄이기 위한 방법은 음성 인식 과정에서 어느 단계에 적용하는지에 따라 구분될 수 있다. 음성 입력 단계에서 입력된 신호로부터 직접 잡음을 제거하거나 특징 추출 단계에서 잡음 요소로 판단되는 부분을 제거 혹은 보상해 주는 방법은 전처리 단계에서 적용하는 방법으로 분류할 수 있다. 주파수 차감법 (Spectral subtraction)이나 다양한 필터를 기반으로 하는 잡음 억압 및 음성 강화 기술, 켑스트럼 평균 정규화 (CMN, Cepstral Mean Normalization), 각종 특징 보상 (Feature compensation)과 같은 기법들은 전처리 단계에서 잡음 요소를 제거함으로써 깨끗한 환경에서 훈련된 음향 모델에 인식 환경의 음향학적 특성을 일치시키고자 하는 시도이다[1]. 인식에 사용되는 음향 모델 파라미터를 변화된 잡음 환경에 맞게 갱신시켜주는 방법은 훈련이나 인식 단계에서 처리하는 기법이다. MAP (Maximum A Posteriori) 적응 (Adaptation)이나 MLLR (Maximum Likelihood Linear Regression) 적응 기법은 소량의 데이터로 변화된 환경과 일치하도록 음향 모델을 변화시키고자 하는 기술이며, 병렬적 모델 결합 (PMC, Parallel Model Combination) 기법은 깨끗한 음성 모델과 잡음 모델을 독립적으로 이용하여 잡음 환경에 맞는 오염 음성 모델을 생성해내는 기술이다 [2-4].

본 논문에서는 특징 보상 기술의 하나인 병렬 결합된 혼합 모델 (Parallel Combined Mixture Model, PCMM) 기반의 보상 기법에 대해 관심을 갖는다[5,6]. PCMM 기법은 음향 모델을 기반으로 하는 특징 보상 기법의 하나로서 GMM (Gaussian Mixture Model)을 음향 모델로 이용한다. PCMM 기법에서는 PMC 기법을 도입하여 별도의 훈련 과정 없이 오염 음성의 GMM을 생성함으로써 보상에 이용하는 보정 인자를 예측한다. 본 논문에서는 한정적인 잡음 요인으로 인해 예상이 가능한 잡음 환경에 효과적이고 효율적인 PCMM 기반의 특징 보상 기법을 제안한다. 잡음 환경이 예상 가능함에 따라, 오프라인에서 병렬적 모델 결합 기법에 의해 생성된 오염 음성의 GMM 모델을 이용함으로써 온라인 상에서의 복잡한 연산 과정을 피하고자 하였다. 보다 신뢰성 있는 모델 생성을 위하여 데이터 유도 기반의 병렬 모델 결합 기법을 도입하였고, 다중 모델 보간을 적용함으로써 잡음 환경의 변화에 적응적으로 대처할 수 있게 하였다. 또한, 다중 모델을 사용함으로써 발생되는 연산량의 증가를 막기 위하여 혼합 모델을 공유하는 기법을 제안하였다.

본 논문은 다음과 같이 구성된다. II장에서는 본 논문에서 기본 기술로 사용하고 있는 PCMM 기반의 특징 보상 기법에 관하여 설명하고, 본 논문에서 제시하는 기법들의 필요성을 기술한다. III장에서는 본 논문에서 도입한 모델 보간 기법과 변형된 형태에 대해서 설명하고, IV장에서는 제안한 혼합 모델 공유 기법에 관하여 상세히 기술한다. V장에서 제안하는 기법의 성능 평가를 위해 실시한 실험과 그 결과에 관해 고찰하고 VI장에서 결론을 맺는다.

II. PCMM 기반의 특징 보상 기법

PCMM 기반의 특징 보상 기법은 음성 특징 분포의 모델링을 위해 GMM을 이용하는 모델 기반의 특징 보상 기법을 기초로 하여 개발되었다. 음성 모델을 기반으로 하는 특징 보상 기법은 Acero에 의해 처음으로 소개되었으며, 그 후 Moreno가 RATZ (Multivariate Gaussian Based Cepstral Normalization)라는 명칭으로 데이터 유도방식의 기법을 제안한 후 유사한 방식들의 근간이 되어 왔다[7,8]. RATZ에서는 잡음 환경 하에서의 음성 특징 벡터 즉 켑스트럼 (Cepstrum)의 분포가 통계적으로 어떻게 변화하는지를 훈련 데이터를 이용하여 추정한 뒤, 이를 이용하여 실제 잡음 환경에서 오염된 입력 음성 특징을 보상하는 과정으로 이루어진다.

깨끗한 음성 신호의 켑스트럼 $\mathbf{x}$의 분포는 다음과 같이 K개의 가우시안 요소로 이루어진 혼합 모델 형태로 추정할 수 있다.

$$p(\mathbf{x}) = \sum_{k=1}^{K} a_k N_{\mathbf{x}}(\boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}) \tag{1}$$

RATZ에서는 식 (1)과 같이 구성된 음성 분포가 잡음 환경에서 평균과 분산이 보정 인자 (Correction factors)에 의해 변화되는 것으로 가정하며, PCMM 기법에서도 동일한 가정을 한다. 이러한 가정 하에서 오염된 음성 $\mathbf{y}$의 분포는 다음과 같이 나타낼 수 있다.

$$p(\mathbf{y}) = \sum_{k=1}^{K} a_k N_{\mathbf{y}}(\boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{r}_k, \boldsymbol{\Sigma}_{\mathbf{x},k} + \mathbf{R}_k) \tag{2}$$

여기에서 $\mathbf{r}_k$과 $\mathbf{R}_k$은 각각 평균과 분산에 대한 보정 인자이다. 입력된 오염된 음성 특징 벡터는 MMSE (Minimum Mean Squared Error) 기반의 예측방법에 의해 깨끗한 음성 특징으로 복구된다.

$$\begin{aligned}\hat{\mathbf{x}}_{MMSE} &= E\{\mathbf{x} \mid \mathbf{y}\} = \int_{\mathbf{x}} \mathbf{x} \cdot p(\mathbf{x} \mid \mathbf{y}) d\mathbf{x} \\ &\cong \mathbf{y} - \sum_{k=1}^{K} \mathbf{r}_k \, p[k \mid \mathbf{y}]\end{aligned} \tag{3}$$

RATZ 기법에서는 잡음 환경에서 깨끗한 음성 모델의 변화를 나타내는 보정 인자 $\mathbf{r}_k$과 $\mathbf{R}_k$을 잡음 환경에서 수집된 오염된 음성 데이터베이스에 대한 훈련 과정으로부터 얻을 수 있다. 반면에, PCMM 기반의 기법에서는 훈련 과정 없이 깨끗한 음성 모델과 잡음 모델의 결합에 의해 예측된 오염된 음성 모델 분포로부터 얻을 수 있다[6].

오염된 음성의 가우시안 혼합 모델을 얻기 위해서는 병렬 모델 결합법 (Parallel Model Combination, PMC)이 도입된다. PMC 기법은 깨끗한 음성 모델과 잡음 모델을 독립적으로 이용하여 오염된 음성 모델을 생성하는 방법으로 오염된 음성 데이터베이스를 이용한 부가적인 훈련 과정이 필요하지 않는 뛰어난 장점을 가진다[4]. 기존의 PCMM 기법에서는 온라인 상에서의 혼합 모델 결합을 위하여 로그 정규 가정법 (Log-normal approximate)이 적용되었다.

본 논문에서는 작은 연산량을 요구하는 전처리 기법의 제안을 목표로 하므로 온라인 상에서의 모델 예측을 적용하지 않고, 오프라인에서 훈련된 모델을 이용하였다. 보다 신뢰성 있는 모델 예측을 위하여 데이터 유도 (Data-driven) 방식의 모델 결합법을 적용하였다. 데이터 유도 방식의 모델 결합법은 각 음향 모델로부터 인공적으로 발생시킨 음성 특징 벡터와 잡음 특징 벡터를 선형 스펙트럼 상에서 더함(Addition)으로써 오염 음성 특징을 발생시키고, 이로부터 오염된 음성 모델을 예측하는 방법이다. 이 방식은 로그-정규 가정법에서 불가피하게 발생되는 예측 손실을 줄임으로써 보다 정확한 결합 모델을 생성할 수 있다.

RATZ에서 가정 했던 평균과 분산의 이동을 고려하면, PCMM 기법에서 사용되게 되는 보정 인자를 다음과 같이 구할 수 있다.

$$\begin{aligned}\mathbf{r}_k &= \tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k \\ \mathbf{R}_k &= \tilde{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\end{aligned} \tag{4}$$

여기에서 $\tilde{\boldsymbol{\mu}}_k$과 $\tilde{\boldsymbol{\Sigma}}_k$은 데이터 유도 방식의 모델 결합으로부터 얻은 오염된 음성의 가우시안 혼합 모델의 평균과 분산을 나타낸다.

기존의 PCMM 기법에서는 시간에 따라 특성이 변하는 배경 잡음에 대처하기 위하여 잡음 모델에 대한 적응 (Adaptation) 기법이 적용되었다. 적응된 잡음 모델을 오염 음성 모델에 반영하기 위해서는 적응된 시점에서 모델 결합을 수행해야 하는데, 이는 선형 스펙트럼, 로그 스펙트럼, 켑스트럼으로의 변환 과정에서 오는 상당한 계산량을 요구하게 된다. 본 논문에서는 예상할 수 있는 배경 잡음을 반영하는 다중 (Multiple)의 모델을 이용함으로써 복잡한 적응 과정이 필요없는 효율적인 PCMM 기반의 특징 보상 기법을 제안한다. 미리 훈련되어진 다중의 모델을 이용함으로써 온라인 상에서의 모델 결합 과정 없이 시간에 따라 변하는 잡음 환경에 적응적으로 대처하여 효과적인 특징 보상을 수행할 수 있게 한다.

III. 다중 모델의 보간

단일의 잡음 모델을 이용하는 특징 보상 기법에서는

인식 과정에 수행되는 잡음 환경이 고정적이고, 미리 예상할 수 있음을 가정하여 미리 훈련되어진 모델로부터 보정 인자를 추정하여 이를 이용하게 된다. 하지만, 실제 환경에서는 시간, 장소에 따라 잡음 환경이 수시로 바뀌기 때문에 이러한 가정이 불가능하다. 다중 모델 기법에서는 입력된 오염 음성에 대해 예상 가능한 E개의 잡음 환경의 사후 확률을 각각 계산하여 현재의 잡음 환경과 가장 유사한 잡음 모델을 예측하게 된다[8]. 발생이 가능할 것으로 기대되는 잡음 환경을 반영하는 다중의 모델을 이용함으로써 시간에 따라 변하는 잡음 환경에 대처할 수 있다.

본 논문에서는 매 프레임별 (Frame-by-frame) 처리를 위해 사후 확률의 예측 과정을 식 (5)와 같이 프레임에 동기화된 형태로 수정하였다. 시간 t까지의 입력 특징 벡터 $\mathbf{Y}_1^t=[\mathbf{y}_1,\mathbf{y}_1,...,\mathbf{y}_t]^T$ 가 주어졌을 때, $\mathbf{Y}_1^t$에 대한 i번째 환경의 사후 확률은 다음과 같다.

$$P[i\,|\,\mathbf{Y}_1^t]=\frac{P[i]p(\mathbf{Y}_1^{t-1}\,|\,i)p(\mathbf{y}_t\,|\,i)}{\sum_{e=1}^{E}P[e]p(\mathbf{Y}_1^{t-1}\,|\,e)p(\mathbf{y}_t\,|\,e)} \tag{5}$$

$$p(\mathbf{Y}_1^{t-1}\,|\,i)=p(\mathbf{Y}_1^{t-2}\,|\,i)p(\mathbf{y}_{t-1}\,|\,i)=\prod_{\tau=1}^{t-1}p(\mathbf{y}_\tau\,|\,i) \tag{6}$$

식 (3)으로부터 깨끗한 특징 벡터는 다음과 같은 식으로 프레임에 동기화된 형태로 복구될 수 있다.

$$\hat{\mathbf{x}}_{MMSE}\cong\mathbf{y}_t-\sum_{e=1}^{E}P[e\,|\,\mathbf{Y}_1^t]\sum_{k=1}^{K}\mathbf{r}_{e,k}P[k\,|\,e,\mathbf{y}_t] \tag{7}$$

식 (7)에서 $\mathbf{r}_{e,k}$는 e번째 환경을 반영하는 혼합 모델의 k번째 가우시안 요소의 평균에 대한 보정 인자를 나타내며 식 (4)로부터 구할 수 있다.

주행 중인 자동차 실내 환경과 같이 배경 잡음의 종류가 어느 정도 한정적일 것으로 가정할 수 있는 상황에서는 다중 모델 기법이 적응 기법이나 온라인상에서의 예측 기법보다 계산의 복잡도 측면에서 효율적일 수 있다. 깨끗한 환경을 다중 모델 중 하나로 가정한다면, SNR이 높은 환경으로 바뀌는 상황에도 깨끗한 환경에서 가지는 기본 인식 성능을 유지할 수 있다. 뿐만 아니라, 깨끗한 환경의 모델과 잡음 모델간의 보간 효과로 인해 시간에 따라 변하거나 예상할 수 없는 SNR 상황에 대해 적절하게 적응할 수 있는 장점을 갖게 된다.

IV. 혼합 모델의 공유

GMM을 기반으로 하는 특징 보상 기법을 구현하기 위해 필요한 연산량은 확률 값을 계산해야 할 가우시안 요소의 개수에 크게 좌우된다. 본 논문에서 사용하는 다중의 모델 기법에서는 모델의 보간에 이용되는 환경 모델의 개수가 계산의 복잡도를 결정하게 된다. 환경 모델의 개수를 늘일수록 다양한 환경에 적응적으로 대처할 수 있는 성능을 가지지만 그 만큼 계산량이 늘어나게 된다.

본 논문에서는 계산량을 줄이기 위하여 다중의 환경을 나타내는 가우시안 혼합 모델 사이에서 통계학적으로 유사한 부분을 공유하는 기법을 제안하고자 한다. Kullback-Leibler 거리[9]를 이용하여 각각의 혼합 모델에서 서로 유사한 가우시안 요소를 선택하고, 일종의 스무딩 과정을 통해 공유를 위한 공통 모델을 추정한다. 유사한 요소의 선택은 다음과 같은 알고리즘에 의해 이루어진다.

$$\begin{aligned}
&\text{Step 0}: \mathbf{D}=\{d_1,d_2,...,d_K\},\mathbf{C}_S=\phi\\
&d_k=\sum_{e=1}^{E}kl_dist(g_{1,k},g_{e,k}),\ 1\le k\le K\\
&\text{Step 1}: k=\arg\min d_i\in\mathbf{D}\\
&\text{Step 2}: \mathbf{C}_S=\mathbf{C}_S\cup\{k\},\mathbf{D}=\mathbf{D}-\{d_k\}\\
&\text{Step 3}: \text{if } N(\mathbf{C}_S)=K_S, \text{ then stop.}\\
&\qquad\text{Else, then go to Step 1}
\end{aligned} \tag{8}$$

식 (8)에서 d_k는 첫 번째 환경 혼합 모델의 k번째 가우시안 요소 $g_{1,k}$로부터 각 혼합 모델의 k번째 요소의 Kullback-Leibler 거리의 합을 나타내며, $N(\cdot)$는 원소의 개수를 말한다. 알고리즘이 종료하게 되면 최종적으로 공유할 가우시안 요소에 해당하는 K_S개의 인덱스를 포함하는 집합 $\mathbf{C}_S$를 얻게 된다. 공유에 사용되는 스무딩된 가우시안 요소의 파라미터는 다음과 같이 예측할 수 있다.

$$\tilde{\boldsymbol{\mu}}_{Sk}=\frac{1}{E}\sum_{e=1}^{E}\tilde{\boldsymbol{\mu}}_{e,k},\ \tilde{\Sigma}_{Sk}=\frac{1}{E}\sum_{e=1}^{E}\tilde{\Sigma}_{e,k},\ k\in\mathbf{C}_S \tag{9}$$

따라서 집합 $\mathbf{C}_s$에 포함되는 가우시안 요소의 우도 (Likelihood) 함수는 공유 요소로 대신할 수 있다.

$$p(\mathbf{y}\mid e,k)=\begin{cases}p(\mathbf{y}\mid\tilde{\boldsymbol{\mu}}_{sk},\tilde{\boldsymbol{\Sigma}}_{sk}) & \text{if } k\in\mathbf{C}_s\\ p(\mathbf{y}\mid\tilde{\boldsymbol{\mu}}_{e,k},\tilde{\boldsymbol{\Sigma}}_{e,k}) & \text{otherwise}\end{cases} \tag{10}$$

집합 $\mathbf{C}_s$에 포함된 인덱스에 해당하는 보정 인자는 다음과 같이 구할 수 있다.

$$\mathbf{r}_{e,k}=\begin{cases}\mathbf{r}_{sk}=\tilde{\boldsymbol{\mu}}_{sk}-\boldsymbol{\mu}_k & \text{if } k\in\mathbf{C}_s\\ \mathbf{r}_{e,k}=\tilde{\boldsymbol{\mu}}_{e,k}-\boldsymbol{\mu}_k & \text{otherwise}\end{cases} \tag{11}$$

공유기법을 도입함으로써 K개의 가우시안 요소를 가지는 E개의 잡음 환경 혼합 모델에 대해서 매 프레임마다 $E\times K$ 번의 가우시안 확률 계산을 $K_s+E(K-K_s)$ 번의 계산으로 줄일 수 있게 된다. 따라서 $(E-1)K_s$ 만큼의 계산량 단축 효과를 갖게 된다. 그림 1은 혼합 모델의 공유 과정을 그림으로 나타낸 것이다.

V. 실험 및 결과

5.1. Aurora 2.0에 대한 성능 평가

객관적인 성능 평가를 위해서 ELRA (European Language Resources Association)의 Aurora 2.0에서 제공하는 평가 방식을 따랐다. Aurora 2.0에서의 평가 방식의 주요 특징은 다음과 같다[10].

1) 영어 음성, 연속 숫자음 인식, 11단어+묵음 구간(Silence)+짧은 휴지(Short pause)

2) ETSI (European Telecommunications Standards Institute) 표준의 DSR(Distributed Speech Recognition) 방식의 특징 추출[11]

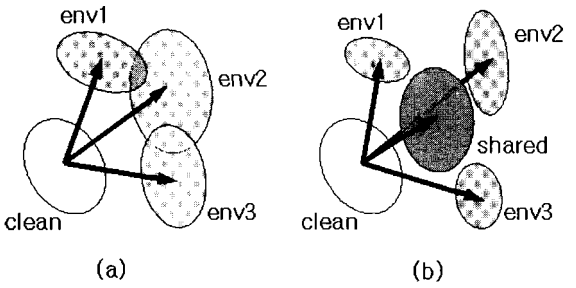

그림 1. 혼합 모델 공유의 개념
Fig. 1. Concept of the mixture sharing.

3) 13차 static 특징(c1~c12+로그 에너지) 추출 후 인식 단에서 미분계수 추출(총 39차) : 본 논문의 실험에서는 PCMM 구현의 편의를 위하여 로그 에너지 대신 캡스트럼의 0차 계수를 사용하였다.

4) 3-mixture, 16-state의 단어 모델, 2종류의 silence 모델

Aurora 2.0에서 제공하는 Clean-condition Training, Multi-condition Testing 방식에 따라 음향 모델은 깨끗한 환경에서 수집된 8,840개의.음성 데이터를 이용하여 훈련하였다. Multi-condition Testing을 위한 데이터베이스에는 각각 4 종류의 배경 잡음 환경으로 이루어진 Set A와 Set B가 있으며, 채널 왜곡이 반영된 2종류의 잡음 환경의 Set C가 있다. 모든 잡음 환경 테스트 데이터는 7가지의 신호 대 잡음비 (SNR)에 따라 각각 1,001개의 샘플로 구성되어 있다.

우선 기존의 대표적인 전처리 알고리즘에 대해 잡음 환경에서 베이스라인 시스템의 성능을 비교하여 그 결과를 표 1에 나타내었다. 대표적 전처리 알고리즘으로 가장 일반적으로 사용되는 주파수 차감법 (Spectral Subtraction, SS)과 캡스트럼 정규화 (Cepstral Mean Normalization, CMN) 기법을 선택하였으며 Aurora 2.0의 자동차 잡음 환경에 대해 실시하였다. 주파수 차감법에서는 배경 잡음을 추정하기 위해 250msec의 시간 지연을 갖는 최소 통계 (Minimum statistics) 기법을 적용하였다[12]. PCMM1과 PCMM2는 테스트 환경과 동일한 SNR을 갖도록 오염 음성 모델의 결합과정을 적용한 PCMM 기반의 특징 보상 기법을 나타낸다. 모델의 결합을 위해서 PCMM1에서는 로그-정규 가정법을 적용하였고, PCMM2에서는 데이터 유도 기법을 적용하였다. 깨끗한 음성 특징 분포의 모델링은 128개의 가우시안 요소를 가지는 혼합 모델을 이용하였으며, 이는 베이스라인 시스템의 HMM 훈련에 사용된 동일한 데이터베이스의 훈련을 통해 얻었다. 각각의 잡음 모델은 하나의 요소를 가지는 가우시안 분포로 추정하였다.

표 1로부터 테스트 환경과 일치하는 SNR을 갖도록 결합 모델을 생성하여 적용한 PCMM 기반의 특징 보상 기법이 주파수 차감법이나 CMN 또는 둘의 조합에 의한 전처리 기법보다 잡음 환경에서 우수한 성능을 가지는 것을 알 수 있다. 이러한 결과는 본 논문에서 기초로 하고 있는 음성 모델 기반의 특징 보상 기법이 음성 인식 과정의 직접적인 대상이 되는 캡스트럼 영역을 깨끗한 음성 특징의 분포로 이동시킴으로써 일반적인 전처리 기법

표 1. Aurora2.0의 자동차 잡음에 대한 베이스라인의 단어 인식률(%).
Table. 1. Word accuracy for baseline system to car noise condition in Aurora2.0(%).

	Baseline	SS	SS+CMN	PCMM1	PCMM2
Clean	98.84	98.63	98.87	98.84	98.84
20dB	96.42	97.38	97.76	97.91	97.94
15dB	87.62	93.98	95.53	97.11	97.08
10dB	61.71	81.42	86.22	93.86	93.86
5dB	26.87	50.16	59.59	81.81	82.20
0dB	10.38	17.66	25.05	53.06	53.92
-5dB	8.41	5.99	14.43	21.77	22.10
평 균	56.60	68.12	72.83	84.76	85.00

보다 음성 인식의 성능 향상 측면에서 우수함을 입증한다. PCMM2가 PCMM1보다 다소 높은 성능을 가지는 결과로부터 데이터 유도에 의한 모델 결합 방식이 로그-가정법보다 신뢰성 있는 음향 모델을 생성할 수 있음을 알 수 있다. 이 후의 모든 실험에서는 PCMM 기법에서의 모델 결합을 위해 데이터 유도 기법을 적용하였다.

베이스라인 시스템의 성능 평가와 동일한 조건 하에 본 논문에서 제안한 기법의 성능을 비교하였다. 다중 모델의 보간을 위해서는 17dB, 7dB, -2dB의 SNR을 가지는 3개의 오염 음성 혼합 모델을 생성하였다. 깨끗한 음성 모델을 또 하나의 환경 모델로 간주한다면 다중 모델의 개수는 4개가 된다. 객관적인 비교를 위해서 다음과 같이 제안한 기법의 조합에 대한 성능을 평가하였다.

1) IP : 다중 모델 보간 PCMM 기반 특징 보상

2) SS+IP : 주파수 차감법 + 다중 모델 보간 PCMM

3) SS+IP+CMN : 주파수 차감법 + 다중 모델 보간 PCMM + 켑스트럼 정규화

4) SS+IP64+CMN : 주파수 차감법 + 64개의 가우시안 요소를 공유하는 다중 모델 보간 PCMM + 켑스트럼 정규화

5) SS+IP96+CMN : 주파수 차감법 + 96개의 가우시안 요소를 공유하는 다중 모델 보간 PCMM + 켑스트럼 정규화

표 2는 본 논문에서 제안한 기법의 성능 비교 결과를 나타낸 것이다. 표의 결과는 제안한 기법이 표 1에 나타난 기존의 기법들의 결과에 비해 잡음 환경에서 보다 효과적인 성능을 가지는 것을 보여준다. 다중 모델의 보간을 이용하는 PCMM 기법이 테스트 환경과 일치하는 SNR을 갖는 단일 모델을 사용한 기법과 거의 유사한 성능을 나타내었다. 이러한 결과는 여러 종류의 SNR을 가지는 모델들의 보간 과정이 매 발음 마다 변화하는 SNR 환경에 적응적으로 대처하여 특징을 보상하는데 효과적임을 나타낸다.

표 2. Aurora2.0의 자동차 잡음에 대한 제안한 기법의 단어 인식률(%).
Table. 2. Word accuracy for the proposed schemes to car noise condition in Aurora2.0(%).

	IP	SS+IP	SS+IP+CMN	SS+IP64+CMN	SS+IP96+CMN
Clean	98.84	98.84	98.87	98.87	98.87
20dB	97.88	97.88	98.18	98.03	97.32
15dB	97.35	97.17	97.91	97.73	96.93
10dB	93.29	94.57	95.35	94.75	94.42
5dB	81.78	86.46	87.56	86.28	87.06
0dB	53.80	63.41	65.37	62.90	64.54
-5dB	22.79	29.17	28.18	24.52	25.56
평 균	84.82	87.90	88.87	87.94	88.05

SS+IP는 주파수 차감법을 이용하여 잡음을 제거한 후에 PCMM 기반의 특징 보상 기법을 적용한 것으로서, 특징 보상 기법만을 적용한 결과보다 우수한 성능을 가지는 것을 확인할 수 있었다. 이러한 현상은 주파수 차감법을 통해 특징 보상의 입력이 되는 음성이 향상된 SNR을 갖기 때문이다. 또한, 주파수 차감법을 통해 향상되는 SNR은 그 변화를 예측하기가 쉽지 않은데, 다중 모델의 보간 과정이 SNR 변화에 효과적으로 대처할 수 있음을 보여주는 결과이다.

다중 혼합 모델의 일부를 공유하는 기법(IP64, IP96)의 결과는 공유하지 않는 경우에 비해 다소 낮은 성능을 보여주었지만, 연산량을 줄이는 효과를 감안하면 만족할 만한 수준으로 판단할 수 있다. 계산량과 성능 사이의 정량적 관계는 뒤의 실험에서 다루고자 한다.

표 3과 표 4는 Aurora 2.0에 포함되어 있는 깨끗한 환경 훈련 (Clean-condition Training), 다중 환경 테스트 (Multi-condition Testing)의 모든 세트들에 대해 성능 평가를 실시한 결과를 나타낸다. 표 4의 결과는 제안한 특징 보상 기술이 다양한 배경 잡음 뿐 아니라 Set C에서와 같이 채널 왜곡이 존재하는 환경에서도 표 3의 기존의 전처리 기술에 비해 우수한 성능을 나타내는 것을 확인해 준다.

표 5는 제안한 혼합 모델 공유 기법의 성능과 연산량 감소와의 관계를 나타내는 결과이다. 첫 번째 열의 괄호 안에 있는 값은 모델 공유 기법을 적용하지 않은 SS+IP+CMN의 성능 향상율을 100%로 보았을 때의 상대적인 향상율을 나타낸다. 64개의 가우시안 요소를 공

표 3. Aurora2.0의 깨끗한 환경 훈련 및 다중 환경 테스트의 모든 세트에 대한 베이스라인 시스템의 단어 인식률 (%).
Table. 3. Word accuracy for the baseline system to all sets of clean-condition training and multi-condition testing in Aurora2.0 (%).

	Set A	Set B	Set C	Avg.
	59.59	57.18	66.81	60.07
	67.70	65.00	74.85	68.05
	73.32	76.72	74.44	74.90

유하는 SS+IP64+CMN의 경우, 계산량에서는 25%의 감소를 얻는 반면에 성능면에서는 3.16%의 하락을 가져온다. 96개를 공유하는 경우에는 4.18%의 성능 하락에 비해 37.5%의 계산량 단축을 얻을 수 있었다. 따라서, 공유 전의 성능을 만족할만 수준으로 유지하면서도 상당한 계산량 단축을 가져오는 것을 정량적으로 확인할 수 있었다.

5.2. 실제 자동차 주행 환경에 대한 성능 평가

제안한 기법이 실제 상황에서 효과적으로 동작하는지를 알아보기 위하여 실제 자동차 주행 중에 녹음한 음성 데이터베이스에 대한 성능 평가를 실시하였다. 데이터베이스는 SITEC (Speech Information Technology & Industry Center)에서 배포한 Car01과 CarNoise01을 사용하였다[13]. Car01은 80km/h의 속도로 달리는 승용차 안에서 한국어 고립 단어 음성을 수집한 것이고, CarNoise01은 다양한 주행 조건에서 잡음 신호만을 녹음한 것이다.

Car01 데이터베이스에 포함되어 있는 548 어휘의 한국어 단어를 대상으로 인식 평가를 실시하였다. 548 어휘는 자동차 안에서 사용될 수 있는 음성 명령어로 이루어져있다. 헤드 □ 마이크로 녹음된 4,348 음성 샘플을 깨끗한 음성으로 가정하여 HMM 훈련에 사용하였고, 운전자의 햇빛 가리개에 설치된 방향성 마이크로 수집된 1,096 음성 샘플을 테스트 용 오염 음성 샘플로 사용하였다.

표 6은 Cat01 데이터베이스에 대해 베이스라인 시스템과 대표적인 일반적 전처리 기법에 대한 성능 평가 결과를 나타낸 것이며, 표 7은 본 논문에서 제안한 기법에 대한 결과이다. PCMM은 평가 환경인 Car01과 동일한 단일 잡음 환경 모델을 사용한 것으로, CarNoise01 데이터베이스에 포함된 80km/h의 주행 속도에서 수집된 잡음 샘플로 훈련한 것이다. 다중 모델 보간을 이용한 PCMM (IP)의 구현을 위해서는 CarNoise01에 있는

표 4. Aurora2.0의 깨끗한 환경 훈련 및 다중 환경 테스트의 모든 세트에 대한 제안한 기법의 단어 인식률 (%).
Table. 4. Word accuracy for the proposed schemes to all sets of clean-condition training and multi-condition testing in Aurora2.0 (%).

	Set A	Set B	Set C	Avg. (Relative Imp.%)
	85.35	83.75	70.53	81.75 (52.56)
	86.43	84.00	78.07	83.79 (58.41)
	87.72	85.66	83.71	85.96 (64.31)
	86.72	84.82	82.75	85.17 (62.28)
	86.33	84.63	82.59	84.90 (61.62)

표 5. Aurora2.0의 모든 세트에 제안한 기법의 성능과 계산해야할 가우시안 요소 개수 감소의 관계.
Table. 5. Relationship of performance and reduction in the Gaussian number on the testing of all set in Aurora2.0.

	성능 향상 율 (%)	계산해야할 가우시안 요소 갯수
	64.31	512
	62.28 (96.84%)	384 (75.00%)
	61.62 (95.82%)	320 (62.50%)

50km/h, 80km/h, 100km/h의 주행 조건에서 수집된 잡음 샘플을 이용하여 3가지 오염 음성 모델을 생성하여 사용하였다. 표 7의 결과와 같이 평가 환경이 알려져 있지 않다고 가정할 때 다중 모델 보간을 이용할 경우, 평가 환경과 동일한 환경 모델을 사용한 경우와 동일한 성능 평가를 얻을 수 있었으며 이는 다중 모델 보간 기법이 실제 상황에서 발생할 수 있는 잡음 환경 변화에 매우 효과적일 수 있음을 입증하는 결과이다. 표 8의 결과는 제안한 혼합 모델의 공유 기법이 실제 자동차 주행 환경에서도 성능을 유지하면서 계산량을 단축하는데 상당한 도움이 되는 것을 보여준다.

VI. 결 론

본 논문에서는 시간에 따라 변하는 잡음 환경에 효과적으로 대처할 수 있는 효율적인 특징 보상 기법을 제안하였다. 제안한 특징 보상 기법은 음성 분포의 모델을 이용하는 PCMM 기반의 특징 보상 기법을 근간으로 한다. 온라인 상에서의 모델 결합으로 인한 계산의 복잡도

표 6. 실제 주행 환경 데이터베이스 Car01의 채널 4에 대한 베이스 라인 시스템의 단어 인식률 (%).
Table. 6. Word accuracy for the baseline system to the real car-driving condition, channel 4 of Car01 database (%).

Clean (ch1)	Noisy (ch4)	SS (ch4)	SS+CMN (ch4)
94.16	58.76	82.94	88.96

표 7. Car01의 채널 4에 대한 제안한 기법의 단어 인식률 (%).
Table. 7. Word accuracy for the proposed schemes to the channel 4 of Car01 database (%).

PCMM	IP	SS+IP+CMN	SS+IP64+CMN
88.96	88.96	91.33	91.24

표 8. Car01의 채널4에 제안한 기법의 성능과 계산해야할 가우시안 요소 개수 감소의 관계.
Table. 8. Relationship of performance and reduction in the Gaussian number on channel 4 of Car01 database.

	성능 향상 율 (%)	계산해야할 가우시안 요소 갯수
SS+IP+CMN	78.98	512
SS+IP64+CMN	78.76 (99.72%)	384 (75.00%)

를 줄이기 위하여 오프라인에서 병렬 결합으로 생성된 모델을 이용하는 방법을 적용하였다. 예상 가능한 잡음 환경을 반영하는 다중의 혼합 모델을 사용함으로써 시변 잡음 환경에 적응적으로 대응할 수 있게 하였으며, 실시간 처리를 위하여 프레임에 동기화된 형태의 환경 사후 확률 계산 과정을 도입하였다. 다중의 모델 사용으로 인해 발생하는 계산량 증가를 막기 위하여 유사한 혼합 모델을 공유하는 기법을 제안하였다. Aurora2.0과 실제 자동차 주행 환경에서 수집된 데이터베이스에 대해 성능 평가를 실시하여 제안한 특징 보상 기법이 음성 인식 성능 향상과 계산량 단축 측면에서 모두 효과적임을 확인하였다.

감사의 글

본 논문은 두뇌한국21사업을 통해 수행된 연구 결과의 일부입니다.

참 고 문 헌

1. X. Huang, A. Acero, and H. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.
2. J. L. Gauvain and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. on Speech and Audio Processing, 2(2), pp.291-298, April. 1994.
3. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," Computer Speech and Language, 9, pp. 171-185, 1995.
4. M. J. F. Gales and S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEEE Trans. on Speech and Audio Processing, 4(5), pp.352-359, Sep. 1996.
5. W. Kim, S. Ahn, and H. Ko, "Feature Compensation Scheme Based on Parallel Combined Mixture Model," Proc. Eurospeech 2003, pp.677-680, Sep. 2003.
6. 김우일, 이홍규, 권오일, 고한석, "병렬 결합된 혼합 모델 기반의 특징 보상 기술," 한국음향학회지, 22(7), pp.603-611, Oct., 2003.
7. P. J. Moreno, B. Raj, and R. M. Stern, "Data-driven Environmental Compensation for Speech Recognition: A Unified Approach," Speech Communication, 24(4), pp.267-285, July 1998.
8. P. J. Moreno, Speech Recognition in Noisy Environments, PhD Thesis, Carnegie Mellon University, 1996.
9. S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
10. H. G. Hirsch & D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000, Sep. 2000.
11. ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," ETSI ES 201 108 v1.1.2 (2000-04), Feb. 2000.
12. R. Martin, "Spectral Subtraction Based on Minimum Statistics," EUSIPCO-94, pp.1182-1185, Sep. 1994.
13. http://www.sitec.or.kr

저자 약력

• 김 우 일 (Wooil Kim)
1996년 2월: 고려대학교 전자공학과 (공학사)
1998년 8월: 고려대학교 전자공학과 (공학석사)
2003년 8월: 고려대학교 전자공학과 (공학박사)
2003년 9월~2004년 8월: 고려대학교 BK21 정보기술사업단 박사후 연구원
2004년 8월~현재: 미국 카네기멜론 대학교 박사후 연구원
주관심분야: 신호처리, 음성인식, 잡음처리

• 고 한 석 (Hanseok Ko)
1982년 5월: 미국 카네기 멜론 대학교 전기공학 (공학사)
1986년 5월: 미국 메릴랜드 대학교 시스템공학 (공학석사)
1988년 5월: 미국 존스 홉킨스 대학교 전기공학 (공학석사)
1992년 5월: 미국 카톨릭 대학교 전기공학 (공학박사)
1995년 3월현재: 고려대학교 전자컴퓨터공학과 교수
주관심분야: 영상 및 음성 신호처리, 패턴 인식, 데이터 융합