빠른 공분산 보상을 이용한 온라인 HMM 적응

정규준[0] 조훈영 오영환
한국과학기술원 전자전산학과 전산학전공
{sylph, hycho, yhoh}@bulsai.kaist.ac.kr

On-line HMM adaptation using fast covariance compensation for robust speech recognition

Gue-Jun Jung[0], Hoon-Young Cho, Yung-Hwan Oh
Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology

요 약

본 논문에서는 모델 기반의 잡음 보상 방법인 PMC (parallel model combination)를 온라인상에서 적용하는 방법에 관해 논한다. PMC는 파라미터 보상시 미리 계산된 잡음 모델을 필요로 하며 파라미터 보상에 많은 연산을 요구하므로 온라인으로 모델 파라미터를 보상하기가 어렵다. 본 논문에서는 이러한 문제를 해결하기 위해 기존에 제안된 온라인 모델 보상 방법을 살펴보고, 기존 방법에서 보상 시간 문제로 제외한 PMC의 공분산 보상을 비교적 적은 연산량으로 수행하여 인식성능을 더욱 향상시켰다.

고립 숫자음 인식시스템에 백색 잡음을 SNR 0,5,10 dB로 가산한 평가 자료로 실험한 결과, 제안한 방식은 PMC를 적용한 경우에 비해 모델 적응 시간은 적게 걸리면서도 기존의 온라인 모델 보상 방법에 비해 평균 10%의 인식률 향상을 보였다.

1. 서 론

음성 인식 기술은 실용화되어 다양한 분야에 활용이 기대되고 있으나 현재까지는 학습 환경과 사용 환경의 불일치에 의해 발생하는 음성 인식 시스템의 성능 저하 때문에 많은 제약을 받고 있다. 이러한 환경 불일치의 원인은 크게 음성 신호의 스펙트럼 영역에서 가산적인 배경 잡음과 음성의 켑스트럼 영역에서 가산적인 채널 왜곡으로 구분할 수 있으며, 환경 불일치 요인을 줄이기 위해 잡음이 포함된 관측 자료에서 첨가된 잡음을 제거하는 연구들을 시작으로 최근에는 음성 인식 모델이 학습된 환경을 사용 환경과 비슷하게 보상하는 방법들이 활발히 연구되고 있다 [1].

모델을 보상하는 방법은 로그 필터 뱅크 에너지 특징 파라미터를 보상하는 음성과 잡음 분해법(speech and noise decomposition)을 기초로 MFCC 특징 파라미터를 보상하는 PMC로 발전하였고, PMC의 복잡한 보상 과정을 단순화하기 위해 인공적으로 생성한 자료를 이용하는 data-driven PMC가 제안되었기도 하였으며, 현재 온라인 상에 적용하는 연구가 진행되고 있다 [2][3][4]. 이러한 모델 보상 방법은 극심한 잡음 환경에서도 높은 인식률을 보이지만, 정확한 잡음 정보와 많은 연산을 필요로 한다는 문제점을 가지고 있으며 이러한 문제를 해결하기 위한 연구들이 진행되고 있다 [3][4].

본 논문에서는 모델 보상 방법을 온라인에 적용한 기존 방법을 살펴보고, 기존 방법에서 보상 시간 문제로 제외한 공분산 보상을 효과적으로 수행하여 낮은 SNR에서도 높은 인식 성능을 유지하는 방법을 제안한다.

2, 3장에서는 PMC와 기존 온라인 모델 보상 방법에 대해 살펴보고, 4장에서는 제안한 온라인 모델 보상 방법에 대해 설명한다. 5장에서는 오프라인 PMC, 기존 온라인 보상법 및 제안한 방법을 각각 인식기에 적용한 결과를 비교한 후, 6장에서 결론을 맺는다.

2. Parallel Model Combination

PMC는 깨끗한 음성 모델과 잡음 모델을 조합하여 인식 모델을 사용 환경에 적응시키는 방법이다 [3]. 잡음이 혼합된 음성을 가장 잘 인식할 수 있는 방법은 동일한 잡음 환경에서 자료를 수집하고 인식기를 재학습시키는 것이지만 이 방법은 실용적이지 못하다. 만약 음성 모델이 학습 자료의 통계적 특성을 잘 가지고 있다면 그림 1과 같은 모델 파라미터 보상으로 동일한 효과를 얻을 수 있다.

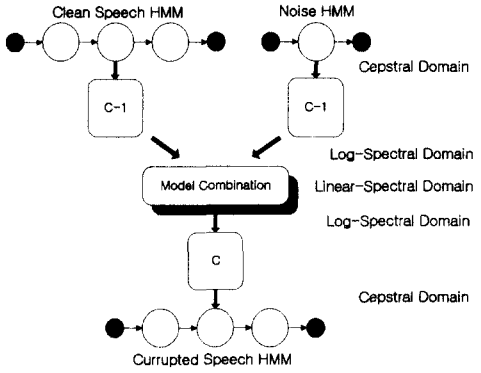

그림 1. Basic PMC process

깨끗한 음성과 배경 잡음은 스펙트럼 영역에서 불일치 함수에 근거하여 혼합되며, 이 불일치 함수를 근거로 모델 파라미터는 식 (1)과 같이 보상된다. $O^l(\tau) = \log(\exp(S^l(\tau)) + \exp(N^l(\tau)))$ 일 때,

$$\begin{aligned} \widehat{\mu}_i^l &= E[O_i^l] \\ \widehat{\Sigma}_{ij}^l &= E[O_i^l O_j^l] - \widehat{\mu}_i^l \widehat{\mu}_j^l \end{aligned} \tag{1}$$

식에서 O, S, N은 각각 관측 자료, 깨끗한 음성, 잡음의 크기를 윗첨자 l은 로그 스펙트럼 영역을, τ는 프레임 인덱스를 나타내며 i, j는

백터 요소 인덱스, $\hat{\mu}$과 $\hat{\Sigma}$은 각각 보상된 모델의 평균과 공분산을 의미한다.

3. 온라인 모델 보상

PMC는 음성 인식 모델을 재학습시키는 방법에 비해 간단하면서도 효과적으로 학습 환경과 사용 환경 사이의 차이를 줄여준다. 그러나 이를 온라인에 적용하기 위해서는 보다 빠른 모델 보상 방법과 실시간에 잡음을 추정하는 방법이 필요하다.

3.1. 실시간 잡음 추정

기존에 제안된 실시간 잡음 추정법은 선형 스펙트럼 영역에서 이루어지며, MEL 주파수 단위로 대역을 분할하여 잡음 정보를 추정한다. 현재 프레임이 음성구간으로 판별될 때까지 현재 프레임과 과거에 추정된 잡음 정보를 바탕으로 식 (2)와 같이 잡음을 추정한다 [5].

$\sqrt{X(t_i,f)} < \beta\sqrt{\hat{N}(t_{i-1},f)}$ 일 때,

$$\sqrt{\hat{N}(t_i,f)} = \alpha\sqrt{\hat{N}(t_{i-1},f)} + (1-\alpha)\sqrt{X(t_i,f)} \qquad (2)$$

식에서 $\sqrt{\hat{N}(t_i,f)}$과 $\sqrt{X(t_i,f)}$은 각각 시간 t_i일때 부대역 f에서 추정된 잡음과 입력 음성의 크기를 의미하고, α, β는 정보의 가중치 및 정보 갱신 유무를 판별하는 문턱값을 의미하며, 이 값들은 실험적으로 정해진다. 잡음 정보 갱신과 동시에 프레임의 음성 구간 식별은 식 (3)을 기준으로 이루어진다 [4].

$$NX(f) = \sqrt{\hat{N}(t_i,f)}/\sqrt{X(t_i,f)}$$

$$NX_{ref}(f) = \frac{NX(f) - NM_{\min}(f)}{NX_{\max}(f) - NX_{\min}(f)} \qquad (3)$$

식에서 $NX(f)$는 추정된 잡음과 신호의 비를 나타내며 $NX_{\max}$, $NX_{\min}$는 모든 이전 프레임에서 $NX(f)$값이 가장 큰 값과 작은 값을 의미한다. $NX_{ref}(f)$값은 0과 1사이의 값을 가지며 잡음에 가까울수록 1에 가까운 값을 가지게 된다.

$NX_{ref}(f)$값을 기준으로 한 부대역에서 적어도 연속한 세프레임에서 「$\gamma - NX_{\min}(f)$」 이하의 $NX_{ref}(f)$값이 발생될 때 현재 프레임을 음성으로 간주한다. 이때, γ값은 실험적으로 정해진다.

3.2. Log-Add 근사법을 적용한 온라인 모델 보상

PMC를 이용하여 모델 보상을 수행할 경우 공분산의 영역 변환에 많은 시간을 필요로 한다. 온라인 모델 보상의 경우 보상을 위한 시간이 한정되므로, 공분산의 영향이 적다는 가정을 적용하여 보상 과정에서 공분산을 제외시킨다. 이 가정을 적용하여 식 (1)을 식 (4)와 같이 간략화 할 수 있으며 이를 Log-Add 근사법이라 한다 [3].

$$\hat{\mu}_i^l = \log(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)) \qquad (4)$$

식 (4)에서 μ_i^l과 $\tilde{\mu}_i^l$은 각각 깨끗한 음성과 잡음의 로그 스펙트럼 영역에서의 평균을 의미한다. 이 방법은 보상 시간이 최소화 된다는 장점은 있지만 모델 보상이 부정확하기 때문에 PMC에 비해 성능은 떨어진다.

4. 공분산 보상을 적용한 온라인 모델 보상

기존 온라인 모델 보상 방법에서는 잡음 추정과 모델 보상 과정을 평균 보상에만 적용하여 보상 속도 측면에서는 많은 개선을 보이지만 성능 향상 측면에서 제약을 받게 된다. 이를 해결하기 위해 보상 속도 문제로 제외된 공분산을 빠르고 효과적으로 보상하는 방법을 제안한다.

4.1. MFCC histogram 잡음 추정

공분산 보상을 위해서는 잡음의 공분산 정보가 필요하며 이를 실시간으로 얻기위해 제안한 잡음 추정 방법은 그림 2와 같다. 이 방법은 기존의 잡음 추정 방법을 확장한 것으로 기존 방법에서는 잡음 정보로 실시간으로 갱신된 값을 사용하지만 제안한 방법에서는 갱신된 값 대신 과거 N개 프레임의 MFCC 값을 이용한다.

각 프레임별로 식 (2), (3)을 계산하여 현재 프레임의 음성 유무를 식별한 후, 음성일 경우 저장된 MFCC값을 이용하여 각 부대역별 평균과 분산을 구하고 이를 잡음 정보로 이용한다.

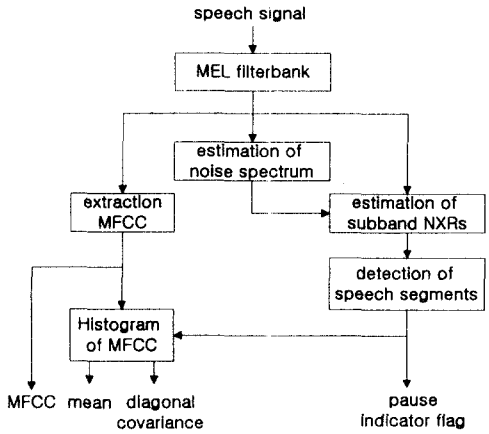

그림 2 제안한 실시간 잡음 추정 과정

4.2. Log-Max 근사법을 적용한 온라인 모델 보상

기존 시스템에서 보상 시간 문제 때문에 모델 보상에서 제외된 공분산 보상식은 식 (5)와 같다.

$$\hat{\Sigma}_{ij}^l = E[\log(S_i+N_i)\log(S_j+N_j)] - E[\log(S_i+N_i)]E[\log(S_j+N_j)] \qquad (5)$$

입력 음성에서 각 부대역의 에너지를 살펴보면 깨끗한 음성에너지와 잡음 에너지 중 우세한 쪽을 따르는 경향을 보이며 식 (6)과 같이 근사된다.

$$\log(S_i+N_i) \approx \max\{\log S_i, \log N_i\} \qquad (6)$$

이 근사식을 식 (5)에 적용하면 식 (7)과 같이 간략하게 공분산을 보상할 수 있다 [6].

$$
\begin{aligned}
S_i > N_i, S_j > N_j &: \widehat{\sigma_{ij}^l} = E[\log S_i \log S_j] - E[\log S_i]E[\log S_j] = \Sigma_{ij}^l \\
S_i < N_i, S_j < N_j &: \widehat{\sigma_{ij}^n} = E[\log N_i \log N_j] - E[\log N_i]E[\log N_j] = \widehat{\Sigma_{ij}^n} \\
S_i > N_i, S_j < N_j &: \widehat{\sigma_{ij}^l} = E[\log S_i \log N_j] - E[\log S_i]E[\log N_j] = 0 \\
S_i < N_i, S_j > N_j &: \widehat{\sigma_{ij}^l} = E[\log N_i \log S_j] - E[\log N_i]E[\log S_j] = 0
\end{aligned} \quad (7)
$$

이 방법은 로그 스펙트럼 영역에서 직접 공분산을 보상하므로 공분산의 영역 변환에 필요한 시간을 줄일 수 있다. 제안한 방법에서는 식 (7)에서 계산된 공분산을 PMC의 평균 보상에 다음과 같이 적용한다.

$$
\begin{aligned}
\text{영역변환} : \mu_i &= \exp(\mu_i^l + 0.5\Sigma_{ii}^l) \\
\widetilde{\mu_i} &= \exp(\widetilde{\mu_i^l} + 0.5\widetilde{\Sigma_{ii}^n}) \\
\text{보상} : \widehat{\mu_i} &= \mu_i + \widetilde{\mu_i} \\
\text{영역변환} : \widehat{\mu_i^l} &= \exp(\widehat{\mu_i}) - 0.5\log\left(\frac{\exp(\widehat{\sigma_{ii}^l})}{\widehat{\mu_i}^2} + 1\right)
\end{aligned}
$$

이 과정에서 Log-Max 근사법을 적용한 공분산이 PMC 평균 보상에 사용되어 잡음의 통계치를 잘 반영한 평균을 얻을 수 있다.

5. 실험 및 결과

실험에서는 TIDIGITS의 고립 숫자음을 사용하였으며, 자료는 "1"에서 "9", "zero", "oh" 총 11개의 영어 숫자음으로 구성되어 있고, 더해진 잡음 신호는 NoiseX 92에 있는 자료 중 백색 잡음을 사용하였다 [7][8]. 자료는 8kHz로 다운샘플링하였으며, 학습 자료로는 남여 112명의 깨끗한 음성을 이용하였고, 평자 자료로는 학습에 포함되지 않은 남여 각각 56명, 57명에 대한 2486개의 단어 발성을 이용하였다. 평가 자료는 SNR 0, 5, 10dB 3단계로 잡음을 혼합하였다. 인식 모델은 HTK 3.0을 이용하여 생성하였으며, 10개의 상태로 구성된 단어 모델이다. 각 상태는 1개의 가우스(Gaussian) 분포를 가진다. 특징 벡터로는 에너지를 포함한 13차 MFCC, 13차 차분 및 가속 파라미터를 사용하였다.

PMC와 기존의 온라인 모델 보상 방법인 Log-Add 근사법, 제안한 Log-Max 근사법을 각각 인식기에 적용하여 인식률을 측정하였으며, 결과는 그림 3과 같았다. 실험 결과를 살펴보면 제안한 방법은 기존의 온라인 모델 보상법에 비해 SNR 0 일때 15%, SNR 5일 때 8% 가량 성능을 향상시켜 효과적으로 공분산을 보상하고 있음을 확인할 수 있었다.

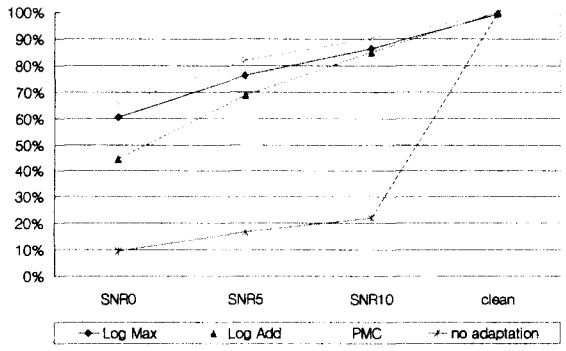

그림 3 단어 인식률 (%)

두 번째 실험에서는 온라인 보상에 중요한 요소인 보상 시간을 측정하였다. Pentium III-500 CPU, 128M RAM, windows 2000 환경에서 측정하였으며 측정결과는 표 1과 같았다.

표 1 보상에 걸린 시간(sec/ update)

	PMC	LogAdd	LogMax
보상시간 (s)	0.1	0.02	0.05

측정된 값은 11개 숫자음 전체에 대한 110개의 HMM상태를 보상할 때 필요한 시간을 나타낸다. 표 1의 결과에서 제안한 방법은 모델 보상에 적은 시간이 소요되어 온라인 보상에 적절함을 알 수 있었다.

6. 결 론

본 논문에서는 PMC를 기반으로 한 온라인 모델 보상에서 HMM의 공분산을 효과적으로 보상하기 위해 잡음 추정 방법과 Log-Max 근사 방식을 적용하였다. 이 방법에 의해 기존 온라인 모델 보상법에서 보상 속도 문제로 제외한 공분산을 적은 시간에 효과적으로 보상하였으며, 기존 방법에 비해 낮은 SNR에서 평균 10%의 인식률 향상을 얻었다. 실험 결과로 볼 때, 제안한 방법은 온라인 모델 보상에 유용한 방법임을 알 수 있었다.

향후에는 채널 왜곡을 실시간에 효과적으로 보상할 수 있는 방법에 관한 연구가 필요하다.

참 고 문 헌

[1] Y. Gong, "Speech recognition in noisy environments: A survey," Speech Communication, vol. 16, pp. 261-291, 1995

[2] A. P. Varga, R. K. Moore, "Hidden Markov model decomposition of speech and noise," Proc. ICASSP, pp 845-848, 1990

[3] M. J. F. Gales, S. Young, "Model based techniques for noise robust speech rocognition," Dissertation at the University of Cambridge, 1995

[4] H. G. Hirsch, "HMM adaptation for applications in telecommunication," Speech Communication, vol 34, pp 127-139, 2001

[5] H. G. Hirsch, C. Ehrlicher, "Noise estimation techniques for robust speech recognition," Proc. ICASSP, pp 153-156, 1995

[6] Ivandro Sanches, "Noise-Compensated Hidden Markov Models," IEEE Transaction on Speech and Audio Processing, vol 8, pp 533-540, 2000

[7] R.G. Leonard, "A database for speaker-independent digit recognition," Proc. ICASSP, pp 42.11, 1984

[8] A. P. Varga, H. J. M Steenken, M. Tomlinson, D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Technical report, DRA Speech Rearch Uint, 1992