잡음 차폐를 이용한 온라인 모델 보상

정규준°, 조훈영, 오영환
한국과학기술원 전자전산학과 전산학전공

On-line model compensation using noise masking effect for robust speech recognition

Gue-Jun Jung, Hoon-Young Cho, Yung-Hwan Oh
Department of Electrical Engineering and Computer Science
Division of Computer Science
Korea Advanced Institute of Science and Technology
E-mail : {sylph, hycho, yhoh}@speech.kaist.ac.kr

Abstract

In this paper we apply PMC (parallel model combination) to speech recognition system online. As a representative of model based noise compensation techniques, PMC compensates environmental mismatch by combining pretrained clean speech models and real-time estimated noise information. This is very effective approach for compensating extreme environmental mismatch but is inadequate to use in on-line system for heavy computational cost. To reduce the computational cost and to apply PMC online, we use a noise masking effect - the energy in a frequency band is dominated either by clean speech energy or by noise energy - in the process of model compensation. Experiments on artificially produced noisy speech data confirm that the proposed technique is fast and effective for the on-line model compensation.

Ⅰ. 서론

서버기반 음성 인식 기술은 현재 실용화 단계에 근접하여 다양한 분야에의 활용이 기대되고 있으나 학습 환경과 사용 환경의 불일치에 의해 발생하는 음성 인식 시스템의 성능 저하 문제로 인해 실용화에 어려움을 겪고 있다. 환경 불일치의 원인은 크게 음성 신호의 스펙트럼 영역에서 가산적인 배경 잡음과 켑스트럼 영역에서 가산적인 채널 왜곡으로 구분할 수 있다. 이러한 환경 불일치 요인을 제거하기 위해 잡음이 포함된 관측 자료에서 잡음을 제거하는 방법에서부터 학습된 음성 모델을 사용 환경에 적응시키는 방법까지 다양한 연구가 진행되어 왔다 [1].

모델 보상은 극심한 잡음 환경에서도 높은 인식률을 유지하는 방법으로 로그 필터뱅크 에너지 특징 파라미터를 보상하는 음성 잡음 분해법(speech and noise decomposition), MFCC (Mel Frequency Cepstral Coefficient) 특징 파라미터를 보상하는 PMC, 인공적으로 생성한 자료를 이용하여 재학습하는 data-driven PMC 등이 제안되었다 [2][3]. 그러나 모델 보상은 정확한 잡음 정보와 과도한 연산을 요구하기 때문에 음성인식기에서 온라인으로 적용하기에는 부적합하였다.

본 논문에서는 모델 보상 방법을 온라인에 적용한 기존 방법을 살펴보고, 기존 방법에서 과다한 보상 시간으로 인해 제외한 공분산 보상을 효과적으로 수행할 수 있는 방법을 제안한다. 2장과 3장에서는 PMC와 이를 이용한 온라인 모델 보상 방법에 대해 살펴보고, 4장에서는 제안한 온라인 모델 보상 방법에 대해 설명한다. 5장에서는 오프라인 PMC, 기존 온라인 PMC 및 제안한 온라인 PMC 방법을 인식기에 적용한 결과를 비교한 후, 6장에서 결론을 맺는다.

Ⅲ. Parallel Model Combination

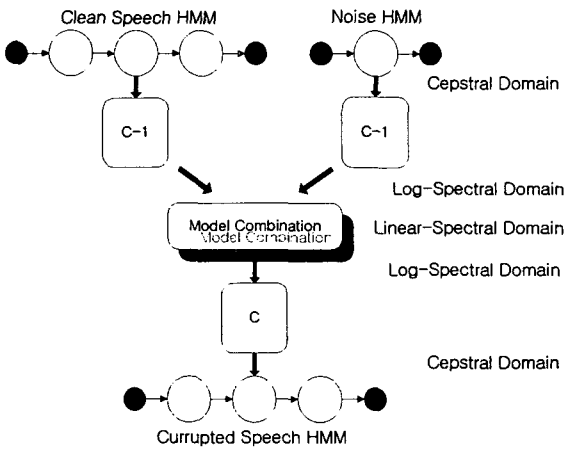

그림 1. Basic PMC process

PMC는 MFCC 특징파라미터로 학습된 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 잡음의 종류가 바뀔 때마다 새로이 자료를 수집하는 것은 효과적이지 못하다. 만약 음성 모델이 학습 자료의 통계적 특성을 잘 표현하고 있다면 그림 1과 같은 모델 파라미터 보상으로 동일한 효과를 기대할 수 있다.

O, S, N 이 각각 관측자료, 깨끗한 음성, 잡음의 크기를 의미한다고 할 때, 깨끗한 음성과 잡음은 스펙트럼 영역에서 불일치 함수에 근거하여 혼합되며, 불일치 함수를 $O^l(\tau) = \log(\exp(S^l(\tau)) + \exp(N^l(\tau)))$ 와 같이 표현할 때, 잡음음성의 모델 파라미터는 다음 식(1)과 같이 추정된다.

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

식에서 위첨자 l은 로그 스펙트럼 영역, τ는 프레임 번호, i와 j는 벡터의 요소번호를 나타내며 $\hat{\mu}$과 $\hat{\Sigma}$은 각각 잡음이 추가된 모델의 평균과 공분산을 의미한다.

Ⅲ. 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 [4]. 본 장에서는 기존 온라인 모델 보상 방식에서 사용된 잡음추정법과 모델 보상에 관해 간략히 기술한다.

3.1 가중 평균을 이용한 배경 잡음 추정

일반적으로 배경 잡음은 입력 신호에서 음성과 비음성 구간을 구분한 뒤 비음성 구간에서 추정된다. 그러나 배경 잡음이 첨가될 경우 음성과 비음성 구간의 구분에 신뢰성이 떨어지게 된다.

이러한 문제를 극복하고 실시간으로 배경 잡음을 추정하기 위해 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,

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

식에서 $\sqrt{\hat{N}(t_i,f)}$과 $\sqrt{X(t_i,f)}$은 각각 t_i번째 프레임의 부대역 주파수 f 에서 추정된 잡음과 입력 신호의 크기를 의미하고, α, β는 가중치 및 음성 구간 여부를 판별하는 임계치를 의미한다.

3.2. 로그가산 근사법을 이용한 온라인 모델 보상

PMC는 공분산의 영역 변환에 많은 연산을 필요로 한다. 온라인에서 모델 보상이 이루어질 경우 보상 시간이 제한되므로, 공분산은 평균벡터에 비해 인식성능에 영향을 적게 미친다는 가정을 적용하여 보상 과정에서 공분산을 제외시킨다. 이 가정을 적용하여 식(1)을 식(3)와 같이 간략화 할 수 있으며 이를 로그가산(Log-Add) 근사법이라 한다 [3].

$$\hat{\mu}_i^l = \log(\exp(\mu_i^l) + \exp(\tilde{\mu}_i^l)) \quad (3)$$

식(3)에서 μ_i^l과 $\tilde{\mu}_i^l$은 각각 깨끗한 음성과 잡음의 로그 스펙트럼 영역에서 모델의 평균을 의미한다. 이 방법은 보상 속도는 빠르지만, 보상의 정확도는 떨어져 인식성능 향상이 오프라인 PMC에 비해 낮다.

Ⅳ.잡음차폐를 이용한 온라인 모델보상

기존 온라인 모델 보상 방법에서는 잡음 추정과 모델 보상 과정을 음성 모델의 평균에만 적용하여 보상 속도를 줄였으나, 성능 향상 측면에서는 많은 제약을 갖게 되었다. 본 장에서는 이러한 제약을 극복하기 위해 잡음 차폐 가정을 이용하여 모델의 공분산까지도

적은 연산으로 효율적으로 보상하는 방법을 제안한다.

4.1. MFCC histogram 잡음 추정

공분산 보상을 위해서는 잡음의 공분산 정보를 실시간으로 추정하여야 한다. 이를 위해 가중 평균을 이용한 배경 잡음 추정을 그림 2와 같이 확장한다. 제안한 방법에서는 현재 프레임이 음성으로 판별될 때까지 과거 잡음으로 추정된 프레임들의 MFCC 정보를 이용한다 [6].

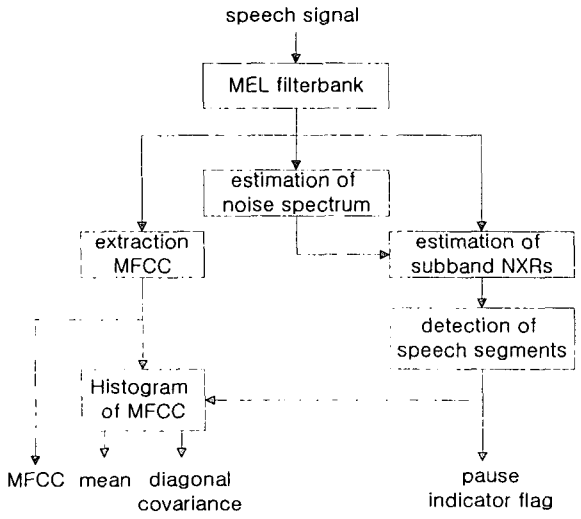

그림 2. MFCC Histogram 잡음 추정 과정

4.2. 로그최대(Log-Max) 근사법

기존 시스템에서 보상시간 문제로 인해 모델 보상단계에서 제외된 공분산 보상식은 식(4)와 같다.

$$\widehat{\Sigma}_{ij}^{y} = E[\log(S_i+N_i)\log(S_j+N_j)] - E[\log(S_i+N_i)]E[\log(S_j+N_j)] \quad (4)$$

식(4)를 직접 계산할 수 있는 방법은 존재하지 않기 때문에 근사법을 이용하여 공분산을 추정한다. 본 논문에서는 입력 신호에 음성과 잡음이 동시에 존재할 경우 신호는 에너지가 우세한 쪽을 따르는 경향을 보인다는 잡음 차폐 가정을 이용한다. 잡음 차폐 가정은 식(5)와 같은 근사식으로 표현할 수 있다.

$$\log(S_i+N_i) \approx \max\{\log S_i, \log N_i\} \quad (5)$$

이 잡음 차폐 가정을 식(4)에 적용하면 식(6)과 같이 간략하게 공분산을 보상할 수 있다 [7].

$$\begin{aligned}
&S_i > N_i, S_j > N_j \text{ 일때} \\
&\quad \widehat{\sigma}_{ij}^{y} = E[\log S_i \log S_j] - E[\log S_i]E[\log S_j] = \Sigma_{ij}^{l} \\
&S_i < N_i, S_j < N_j \text{ 일때} \\
&\quad \widehat{\sigma}_{ij}^{y} = E[\log N_i \log N_j] - E[\log N_i]E[\log N_j] = \widehat{\Sigma}_{ij}^{n} \\
&S_i > N_i, S_j < N_j \text{ 일때} \\
&\quad \widehat{\sigma}_{ij}^{y} = E[\log S_i \log N_j] - E[\log S_i]E[\log N_j] = 0 \\
&S_i < N_i, S_j > N_j \text{ 일때} \\
&\quad \widehat{\sigma}_{ij}^{y} = E[\log N_i \log S_j] - E[\log N_i]E[\log S_j] = 0
\end{aligned} \quad (6)$$

제안한 방법은 선형 스펙트럼 영역에서 공분산을 보상하는 오프라인 PMC와 달리 로그 스펙트럼 영역에서 직접 공분산을 보상하므로 공분산의 파라미터의 영역 변환에 필요한 시간을 줄일 수 있다.

4.3. 로그보간(Log-Interpolation) 근사법

다음과 같이 Δ 함수를 정의하자. 식에서 p는 특징벡터의 크기를 의미한다.

$$\Delta_i = \begin{cases} 0, & \text{if}(S_i/N_i) < 1 \\ 1, & \text{if}(S_i/N_i) \geq 1 \end{cases} \quad i = 1,2,\cdots,p \quad (7)$$

이렇게 정의한 Δ 함수를 이용하면 식(6)은 다음과 같이 표현할 수 있다.

$$\widehat{\sigma}_{ij}^{y} = \Delta_i\Delta_j\Sigma_{ij}^{l} + (1-\Delta_i)(1-\Delta_j)\,\widehat{\Sigma}_{ij}^{n} \quad (8)$$

식(8)을 살펴보면 로그최대 근사법은 깨끗한 음성 에너지와 배경 잡음 에너지의 우세 여부를 가중치로 이용하여 깨끗한 음성과 잡음의 공분산을 가산한다고 볼 수 있다. 즉, 로그최대 근사법의 경우 0 또는 1의 이산(discrete) 가중치를 사용한다고 할 수 있다.

이산 가중치를 의미를 가지는 연속(continuous) 가중치로 대체할 경우 음성과 잡음의 상대적 크기가 반영된 공분산을 구할 수 있다. 제안한 로그보간 보간법은 식(7)에서 정의한 Δ 함수를 식(9)와 같이 음성과 잡음의 에너지 비율 Θ 로 변형한다.

$$\Theta_i = \frac{S_i}{S_i+N_i} \quad (i = 1,2,\cdots,p) \quad (9)$$

이 값을 식(8)에 대입하면 식(10)과 같이 음성과 잡음의 에너지 비를 반영하여 공분산을 보상할 수 있다.

$$\hat{\sigma}_{ij} = \Theta_i \Theta_j \Sigma'_{ij} + (1-\Theta_i)(1-\Theta_j)\,\hat{\Sigma}_{ij} \qquad (10)$$

V. 실험 및 결과

실험을 위해서 TIDIGITS의 고립 숫자음을 사용하였으며, 자료는 "1"에서 "9", "zero", "oh" 총 11개의 영어 숫자음으로 구성되어 있고, 가산된 잡음 신호는 NoiseX 92에 있는 자료 중 백색 잡음을 사용하였다[8][9]. 자료는 8Khz로 다운샘플링하였으며, 학습 자료로는 남여 112명의 깨끗한 음성을 이용하였고, 평가 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 4개의 가우스(Gaussian) 혼합 밀도 함수를 가진다. 특징 벡터로는 에너지를 포함한 13차 MFCC, 13차 차분 및 가속 파라미터를 사용하였다.

제안한 방법의 유효성을 평가하기 위해 PMC, 기존 온라인 모델 보상에 적용된 로그가산 근사법, 제안한 로그최대 근사법 및 로그보간 근사법을 비교 실험하였다. 그림 3은 실험에 사용한 모델 보상 방법들의 단어 인식률을 보이며 표 1은 각 모델 보상 방법으로 모델 보상을 하는데 필요한 시간을 PMC를 기준으로 나타내고 있다.

로그가산 근사법의 경우 공분산을 보상하지 않기 때문에 보상 시간은 가장 빠르지만 배경 잡음의 크기가 커짐에 따라 급격한 성능저하를 보였다. 이와 반대로 PMC의 경우 인식 성능은 보다 완만하게 저하되었으나, 많은 보상 시간을 필요로 하였다. 제안한 로그최대 근사법 및 로그 보간 근사법의 경우 인식 성능은 PMC와 거의 비슷하면서도 보상시간은 PMC에 비해 60% 정도였다.

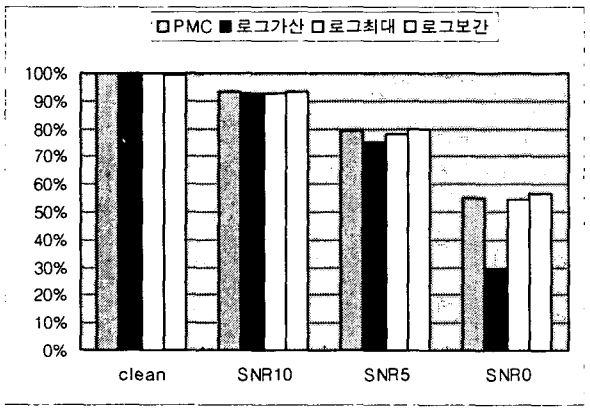

그림 3. 단어 인식률 (%)

표 1. 보상에 걸린 시간 (PMC = 100)

	PMC	로그가산	로그최대	로그가산
PMC기준	100	40	60	60

VI. 결 론

본 논문에서는 PMC를 기반으로한 온라인 모델 보상에서 음성 모델을 효과적으로 보상하기 위해 잡음 차폐 가정을 이용한 로그최대 근사법 및 로그보간 근사법을 제안하였다. 제안한 방법은 기존 온라인 모델 보상법에서 보상 속도 문제로 제외한 공분산을 비교적 적은 시간에 효과적으로 보상하였으며, 기존 온라인 보상 방법에 비해 낮은 SNR에서도 높은 인식률을 얻을 수 있었다. 따라서 제안한 방법은 온라인 모델 보상에 유용한 방법임을 알 수 있었다.

향후에는 잡음과 동시에 채널 왜곡을 효과적으로 추정하고 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 recognition," 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] N.W.D Evans, J.S. Mason, "Noise estimation without explicit speech, non-speech detection: a comparison of mean, modal and median based approaches," Proc. Eurospeech, pp 893-896, 2001.

[6] 정규준, 조훈영, 오영환, "빠른 공분산 보상을 이용한 온라인 HMM 적응," 한국정보과학회 학술발표대회 논문집, 제 28권 2호, pp. 34-36, 2001.

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

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

[9] 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.