# 잡음마스킹을 이용한 환경보상기법

김 영 준[1,2], 김 남 수[2], 이 윤 근[1]
한국전자통신연구원[1]
서울대학교[2]

# Feature Compensation with Model-based Estimation for Noise Masking

Young Joon Kim[1,2], Nam Soo Kim[2], Yun Gun Lee[1]
Electronics and Telecommunications Research Institute(ETRI)[1]
School of Electrical Engeenring, Seoul National University[2],
E-mail : kjun@etri.re.kr

## Abstract

본 논문에서는 음성의 모델을 이용하여 확률적인 기반으로 잡음의 마스킹 정도를 측정하는 방법에 대해서 제시한다. 잡음의 마스킹 정도를 측정하는 기준으로서 '잡음 마스킹 확률'을 구하는 방법에 대해서 설명하고 이의 특성에 대해서 알아본다. 그리고 잡음에 대한 '잡음 마스킹 확률'을 이용하여 잡음 환경에서의 음성인식 특징벡터의 성능 향상에 대해 적용해 보았다. 제안된 방법은 ETSI 에서 음성인식 표준실험으로 제시한 Aurora2 데이터베이스 상에서 실험해 보았다. 그 결과 기존의 알고리즘에 비해 16.58%의 성능 향상을 이루어 낼 수 있었다.

## I. 서론

주변의 소음, 마이크의 특성, 통신 환경의 채널 왜곡 및 사용된 단말기의 부화화기 영향등 다양한 요소들이 음성인식의 성능을 저해하는 요인으로 작용한다. 특히, 환경 잡음의 경우 자동차나 지하철 소음, 거리에서 발생하는 다양한 잡음 그리고 주위 사람에 의한 잡음 등의 다양한 형태로 발생하는데, 이러한 잡음은 예측이 어려울 뿐만 아니라 음성의 왜곡을 가져와 심각한 성능 저하를 초래하게 된다. 더욱이 이러한 잡음들은 비정상 특성을 가지고 있어 확률적으로 모델링이 어렵고 그 특성이 시간에 따라 변하기 때문에 이를 보상하기가 매우 어렵다.

그래서 실제 생활에 적용되기 위한 음성인식 시스템에서는 학습 환경과 음성인식환경의 불일치를 극복하는 강인성이 중요한 문제로 대두된다. 이러한 강인성을 얻는 방법은 일반적으로 두 가지 영역으로 분류된다. 첫 번째 방법은 음성인식에 사용되는 acoustic model을 환경이나 채널의 보상기법에 관한 연구는 크게 Feature domain 기법과 Model-domain 방법으로 분류된다. Model-domain 접근방법의 목적은 잡음환경의 테스트 음성에서의 통계치와 일치하도록 미리 훈련된 reference HMM의 파라미터들을 변경하는 것이다. 이에 반해 Feature-domain 접근방법은 인식과정 전에 전처리단에서 잡음환경에 강인한 특징추출 파라미터나 채널잡음에 의한 영향을 보상해 주는 방법이다.

이러한 Feature-domain 기법에 대해서도 다양한 방법들이 연구되었다. 그중에서 가장 효과적인 방법중의 하나인 음성과 잡음의 결합을 조각적 선형화 모델을 통한 근사하여 잡음환경에서의 음성의 특성을 복원해 내는 방법이다. 이러한 기법은 잡음이 비교적 적은 높은 SNR에서는 효과적이지만 잡음의 정도가 심해져 낮은 SNR로 갈수록 많은 왜곡과 에러를 발생시키게 된다. 이러한 문제를 해결하기위하여 저자는 이전에 soft-decision IMM(SDIMM) 방법을 제시하였다 [3]. 이 방법에서는 구체적인 음성 모델을 이용하여 음성구간에서의 예측이 정확하게 되는 IMM(Interacting Multiple Model) 방법과 음성부재구간에서 안정적인 예측을 하는 SS(Spectral Subtraction) 방법을 음성부재 확률인 SAP(Speech Absent Probability)를 이용하여 결합하는 형태를 가지고 있다.

하지만 이러한 결합의 기준이 되는 SAP가 구체적인

음성모델을 이용하는 것이 아니라 오염된 입력 신호를 이용해 예측하는 것이기 때문에 잡음 특성이 비정상적이거나 잡음의 정도가 심해질수록 강인한 예측이 어렵게 된다. 게다가 음성과 잡음의 결합에서 발생하는 Masking의 효과에 의해서 정확한 음성과 잡음의 구분은 더욱 어렵게 된다

본 논문에서는 먼저 잡음에 의한 마스킹의 문제를 분석하고 그 이후에 구체적인 음성 모델을 이용하여 잡음의 마스킹 정도를 확률적으로 구하는 방법에 대해 제시한다. 이러한 잡음의 마스킹 정도를 '잡음 마스킹 확률(NMP: Noise Masking Probability)' 라 하기로 한다. 이 잡음 마스킹 확률은 매 frame 어느 band에서 잡음에 의해 마스킹 된 정도가 어느 정도인지 예측이 가능하게 한다. 이렇게 구해진 잡음 마스킹 확률을 이용하여 음성구간과 음성부재 구간에서의 다른 알고리즘 결합을 통하여 음성인식 성능 향상 기법을 제시하도록 한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 IMM을 이용한 환경보상기법, 3장에서는 마스킹에 의한 왜곡을 분석하고, 4장에서는 마스킹 확률을 이용한 환경보상기법을 제시한다. 5장에서는 HTK를 이용하여 제안한 보상기법에 따른 인식실험 및 결과를 검토한 후, 6장에서 결론을 맺는다.

## II. IMM 알고리즘

$z=[z_1,z_2,\cdots z_D]^t$ 를 D차 로그스펙트럼 벡터라고 하면 로그스펙트럼에서 음성과 잡음은 다음과 같은 결합 관계를 가지게 된다.

$$z_d = x_d + \log[1+\exp(n_d - x_d)] \quad \text{for } d=1,2,\cdots,D \tag{1}$$

여기에서 $x=[x_1,x_2,\cdots,x_d]^t$와 $n=[n_1,n_2,\cdots,n_d]^t$는 각각 음성과 잡음의 로그스펙트럼 벡터를 의미한다. IMM방법에서는 음성의 분포는 다음과같이 GMM(Gaussian Mixture Model)에 의해 모델링된다.

$$p(x)=\sum_{k=1}^{M} p(k)N(x;\mu_k,\Sigma_k) \tag{2}$$

여기에서 M은 총 믹스쳐수, p(k),$\mu_k$,$\Sigma_k$ 는 각각 사전 확률과 음성의 평균, 분산을 의미한다. 잡음의 분포는 믹스쳐 하나의 가우시안으로 다음과 같이 모델링한다.

$$p(n)=N(n;\mu_n,\Sigma_n)$$

먼저 식(1)의 비선형 함수를 통계적 선형화(SLA: Statistical Linear Approximation) 방법을 통하여 아래와 같이 선형화 한다 [2].

$$z=A_k x + B_k n + C_k \tag{3}$$

변화하는 환경 잡음을 예측하기 위해서 잡음의 모델은 아래와 같이 시간에따라 변화한다고 가정한다.

$$n_{t+1}=n_t+w_t \tag{4}$$

여기에서 $w_t$는 평균 0, 분산이 Q인 가우시안 프라세스라고 가정한다. 식 (4)와 (5)는 선형 공간 모델을 형성하게 되어 잡음의 파라미터 $\lambda_n=\{\mu_n,\Sigma_n\}$에 대해 kalman filter update가 가능하게 된다. 하지만 여러개의 믹스쳐가 존재하기 때문에 각 믹스쳐의 영향을 확률적으로 반영해주는 mixing의 과정이 추가되게 되어 IMM이 아래와 같은 네 가지 단계로 완성되게 된다 [1].

Step 1) Mixing step :

$$\mu_n^0(t-1|k)=E[n_{t-1}|k_t=k,Z_{t-1}] = \sum_{j=1}^{M}\gamma_k(t-1)\hat{\mu}_n(t-1|j) \tag{5}$$

$$\Sigma_n^0(t-1|k)=Cov[n_{t-1}|k_t=k,Z_{t-1}] = \sum_{j=1}^{M}\gamma_k(t-1)[\widehat{\Sigma_n}(t-1|j)+ (\hat{\mu}(t-1|j)-\widehat{\mu_n^0}(t-1|j)) (\hat{\mu}(t-1|j)-\widehat{\mu_n^0}(t-1|j))^t] \tag{6}$$

여기에서

$$\hat{\mu}_n(t-1|j)=E[n_{t-1}|k_{t-1}=j,Z_{t-1}] \tag{7}$$

$$\hat{\Sigma}_n(t-1|j)=Cov[n_{t-1}|k_{t-1},Z_{t-1}] \tag{8}$$

$$\gamma_j(t-1)=p(k_{t-1}=j|Z_{t-1}) \tag{9}$$

Step 2) Kalman Step

- Time update

$$\mu_n^p(t|j)=\hat{\mu}_n^0(t-1) \tag{10}$$

$$\Sigma_n^p(t|j)=\hat{\Sigma}_n^0(t-1)+Q \tag{11}$$

- Innovation and its covariance

$$e(t|j)=z_t-A_j\mu_j-B_j\mu_n^p(t|j)-C_j \tag{12}$$

$$R_e(t|j)=B_j\Sigma_n^p(t|j)B_j^t+A_j\Sigma_jA_j^t \tag{13}$$

- Kalman Gain

$$K_f(t|j)=\Sigma_N^p(t|j)B_j^tR_e^{-1}(t|j) \tag{14}$$

$$K_f^*(t|j)=\alpha K_f(t|j) \tag{15}$$

- Measurement update

$$\widehat{\mu_n}(t|j)=\mu_n^p(t|j)+K_f^*(t|j)e(t|j) \tag{16}$$

$$\widehat{\Sigma_n}(t|j)=\Sigma_n^p(t|j)-K_f^*(t|j)B_j\Sigma_n^p(t|j) \tag{17}$$

Step 3) Probability Calculation Step

$$\gamma_j(t) = \frac{p(z_t|k_t = j, Z_{t-1})p(k_t = j)}{p(z_t|Z_{t-1})} \tag{18}$$

Step 4) Output Generation Step

$$\widehat{\mu_n}(t) = \sum_{j=1}^{M} \gamma_j(t)\hat{\mu}_n(t|j) \tag{19}$$

$$\widehat{\Sigma_n}(t) = \sum_{j=1}^{M} \gamma_j(t)\,[\widehat{\Sigma_n}(t|j) + (\widehat{\mu_n}(t|j) - \widehat{\mu_n}(t))(\widehat{\mu_n}(t|j) - \widehat{\mu_n}(t))^t] \tag{20}$$

여기에서

$$\hat{\mu}_n(t) = E[n_t|k_t = j, Z_t] \tag{21}$$

$$\hat{\Sigma}_n(t) = Cov[n_t|k_t = j, Z_t] \tag{22}$$

$$\gamma_j(t) = p(k_t = j|Z_t) \tag{23}$$

## III. 음성 모델과 잡음의 마스킹 효과

음성 모델이 잡음의 영향에 따라 어떻게 변하는지 살펴보기 위하여 잡음의 크기는 고정시키고 음성의 크기만 변화시켰을때 잡음과 음성의 결합함수가 어떻게 변하는지 fig.1에서 살펴본다. fig.1 에서와 같이 음성의 크기가 잡음의 크기보다 많이 작을 때에는 전체 결합함수는 잡음의 크기만을 따라가게 된다. 이렇게 잡음의 크기가 큰 경우에는 음성에 의한 정보가 관측 데이터에서 거의 사라져서 이를 예측하기가 무척 어렵다.

음성의 사전 모델을 GMM 등과 같은 형태로 가지고 있는 환경보상 기법들에서는 대부분 선형화의 기법을 이용하여 수학적으로 결합가능 하도록 근사화 한다. 하지만 여기에서 마스킹의 효과를 고려하지 않은 근사화는 자칫 더 잘못된 결합식을 유도하게 되어 더 많은 편향(bias)와 에러를 야기하게 된다. 더욱이, GMM의 컴포넌트 대부분이 잡음에 의해 마스킹이 된 경우에의 음성의 예측은 잘못된 경우가 많아 오히려 음성인식이나 음성향상의 성능에 악영향을 미치게 된다.

이러한 문제점을 해결하기 위하여 우리는 '잡음에 의해 마스킹된 믹스쳐 클러스러'를 구분하여 다루는 방법을 제시한다. 잡음에 의해 마스킹된 클러스터들은 음성의 정보가 마스킹되어 음성의 변화에 민감하지 않다는 사실을 이용하여 마스킹 된 클러스터들을 다음과 같은 방법으로 구분해낸다. 즉, 오염된 컴포넌트들의 깨끗한 음성에 대한 도함수가 너무 작게 되면 깨끗한 음성에 의한 영향이 없다고 판단하는 것이다.

$\mu_k = [\mu_{k,1}, \mu_{k,2}, \cdots, \mu_{k,D}]^t$를 깨끗한 음성 GMM의 k번째 컴포넌트라고 가정하고 IMM에 의해 예측된 잡음이 $\mu_n = [\hat{\mu}_{n,1}, \hat{\mu}_{n,2} \cdots \hat{\mu}_{n,D}]$라고 하면 모든 GMM

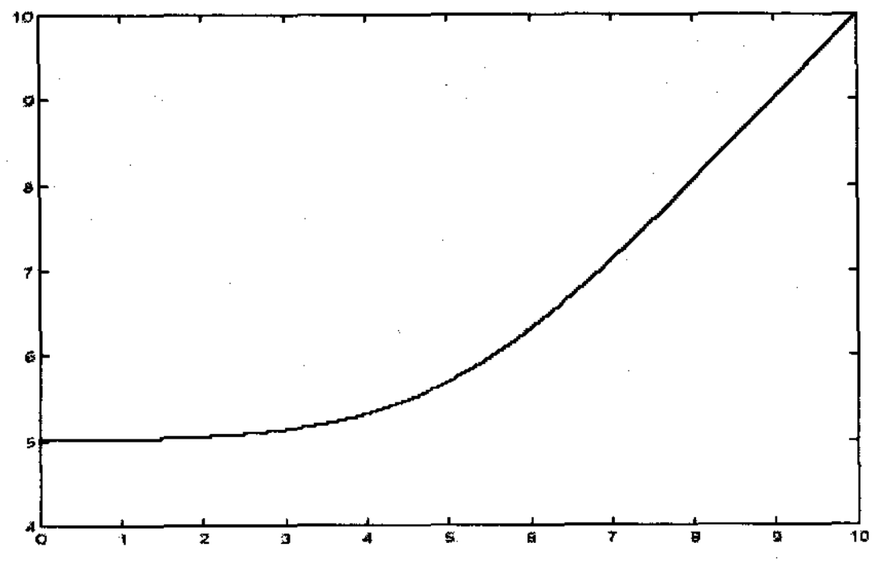


Figure 1: Plot of function $z = x + \log[1 + \exp(n - x)]$. $n$= 5.0 and $x$ ranges from 0 to 10

컴포넌트들은 다음과 같이 잡음에 의해 마스킹 된 집합 $M_{m,d}$과 그렇지 않은 집합 $M_{o,d}$ 로 나눌 수 있다.

즉, $$\frac{\partial z_d}{\partial x_d} = \frac{1}{1 + \exp(\hat{\mu}_{n,d} - \mu_{k,d})} < \eta \tag{24}$$

위와같은 조건에서 이 믹스쳐 컴포넌트가 마스킹 되었다고 할 수 있다. 여기에서 $\eta$는 아주 작은 양수값을 의미한다.

## IV. 마스킹을 이용한 환경 보상

이 장에서는 잡음마스킹 확률(NMP)의 의미와 계산 방법에 대해서 설명하고 그것을 이용해 환경보상하는 방법을 제시한다. NMP는 IMM에 의한 예측값이 얼마나 믿을 만한 것인가에 대한 정도를 말한다. NMP는 각 차수별로 다음과 같이 구할 수 있다.

$$\begin{aligned} NMP_d(z) &= p(k \in M_{m,d}|z) \\ &= \frac{p(z, k \in M_{m,d})}{p(z)} \\ &= \frac{p(z, k \in M_{m,d})}{p(z, k \in M_{m,d}) + p(z, k \in M_{o,d})} \end{aligned} \tag{25}$$

윗 식에서 GMM의 모든 믹스쳐 컴포넌트들이 매 프레임별로 식(5)의 구분기준에 의해 $M_{m,d}$와 $M_{o,d}$로 구분한 각 값들은 다음과 같이 구할 수 있다.

$$p(z, k \in M_{m,d}) = \sum_{k \in M_{m,d}} p(k)p(z|k, \hat{\lambda}_n) \tag{26}$$

$$p(z, k \in M_{o,d}) = \sum_{k \in M_{o,d}} p(k)p(z|k, \hat{\lambda}_n) \tag{27}$$

여기서 $\hat{\lambda}_n$은 IMM에서 예측된 잡음이고 p(k)와 $p(z|k, \hat{\lambda}_n)$는 사전확률과 likelihood를 의미한다.

이렇게 예측된 NMP는 IMM 알고리즘의 신뢰 정도를 반영하고 있기 때문에 그 신뢰도가 떨어지는 경우에는 잡음 마스킹이 상대적으로 덜한 Wiener 알고리

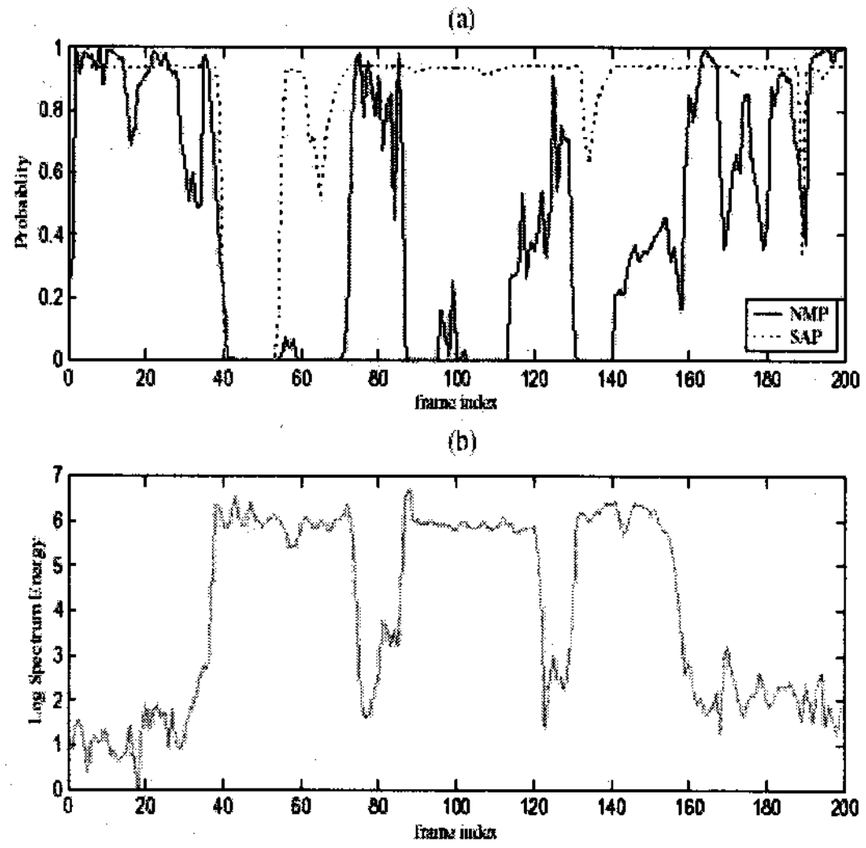


Figure 2: Comparison of NMP and SAP in car noise condition at 10dB SNR. (a) NMP and SAP estimated from a noisy speech signal (b) Corresponding clean speech log spectral energy

즘을 이용하여 아래와 같이 보상함으로써 음성인식 성능을 향상시킬 수 있다.

$$\hat{x_d} = NMP_d(z)\,\widehat{x_d^{wiener}} + (1 - NMP_d(z))\,\widehat{x_d^{imm}} \quad (28)$$

## Ⅳ. 인식실험 및 결과

제안된 알고리즘은 Aurora2 Database를 사용하여 실험하였다. 이 database는 8 kHz의 TI DIGITS로 구성되어 있고 ETSI 음성인식 전처리 표준을 위해 만들어진 데이터이다. A 집합은 suburban train, babble, car, exhibition hall의 네 가지의 잡음이 각각 20, 15, 10, 5, 0 ,-5 dB로 부가되어 있다. B 집합은 다른 네가지의 잡음인 restaurant, street, airport, train station으로 구성되어 있고 SNR별 구성은 A 집합과 같다. C 집합은 A집합에서 두가지 잡음(subway와 street)에 실제 존재하는 채널 영향을 주어 만들어진 데이터로 구성되어 있다 [5] .

베이스라인 음성인식기는 각 digit별로 6개의 state를 사용하였고, 3개의 state로 구성된 하나의 silence 모델과 1 state short pause 모델을 사용하였다. 훈련과 테스트 모두 HTK 를 사용하여 ETSI 전처리 표준에 따랐다 [4]. 제안한 알고리즘은 log spectrum에서 적용하여 DCT를 이용해 cepstrum 전환하였다.

비슷한 효과를 나타내는 SAP와 비교하기 위하여 다음과 같이 SAP를 이용한 환경 보상 방법과 비교하였다.

$$\hat{x_d} = SAP_d(z)\,\widehat{x_d^{wiener}} + (1 - SAP_d(z))\,\widehat{x_d^{imm}} \quad (29)$$

표1은 베이스라인 결과와 두 결과를 비교 실험한 결과를 나타내었다. 여기에서 IMM+NMP와 IMM+SAP는

| | set A | set B | set C | Average |
|---|---|---|---|---|
| Baseline | 61.34 | 55.75 | 66.14 | 60.06 |
| IMM only | 80.69 | 81.35 | 76.23 | 80.06 |
| IMM+SAP | 80.94 | 81.96 | 77.15 | 80.59 |
| IMM+NMP | 83.77 | 83.90 | 78.80 | 82.83 |

표 1. Word accuracies to IMM averaged over the SNR ranges 0-20 dB

각각 NMP와 SAP를 기반으로 환경보상을 한 결과를 나타낸다. 표1과 같이 모든 SNR과 모든 잡음환경에서 IMM+NMP가 가장 좋은 성능을 나타내고 있다. 이는 음성모델을 이용하여 잡음의 마스킹을 예측하는 것이 오염된 음성입력을 통하여 음성의 존재 유무를 판단하는 것보다 효과적이기 때문이라고 할 수 있다. IMM+NMP 방법은 IMM 알고리즘보다 16.58%의 성능 향상을 나타내고 있다.

## Ⅴ. 결론

본 논문에서는 음성모델을 이용한 잡음마스킹 확률을 구하는 방법과 이를 이용한 전처리방법에 대해서 제시하였다. 실험을 통하여 NMP이 효과적임을 증명하였다.

## 참고문헌

[1] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," Speech Comm. vol.37, pp231-248, July 2002.

[2] N. S. Kim, "Statistical linear approximation for environmental compensation," IEEE Signal Process. Lett., vol.5, no.1, pp.8-10, Jan. 1998.

[3] N. S. Kim, Y. J. Kim and H. W. Kim, "Feature compensation based on soft decision," IEEE Signal Process. Lett., vol.11, no.3, pp.378-381

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK book -version3.0, July 2000.

[5] AU/225/00 "Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation", Nokia, Jan. 2000.