Eigen-Environment 잡음 보상 방법을 이용한 강인한 음성인식*

송화전(부산대학교), 김형순(부산대학교)

<차 례>

<Abstract>

Robust Speech Recognition using Noise Compensation Method Based on Eigen-Environment

Hwa Jeon Song, Hyung Soon Kim

In this paper, a new noise compensation method based on the eigenvoice framework in feature space is proposed to reduce the mismatch between training and testing environments. The difference between clean and noisy environments is represented by the linear combination of K eigenvectors that represent the variation among environments. In the proposed method, the performance improvement of speech recognition systems is largely affected by how to construct the noisy models and the bias vector set. In this paper, two methods, the one based on MAP adaptation method and the other using stereo DB, are proposed to construct the noisy models. In experiments using Aurora 2 DB, we obtained 44.86% relative improvement with eigen-environment method in comparison with baseline system. Especially, in clean condition training mode, our proposed method yielded 66.74% relative improvement, which is better performance than several methods previously proposed in Aurora project.

* Keywords: Noise compensation, Environment adaptation in feature space

* 이 논문은 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행됨.

1.서 론

음성 인식기의 성능은 지난 몇 십 년간 많은 연구자 및 개발자들의 부단한 노력으로 계속해서 향상되고 있다. 하지만, 훈련 및 인식 환경이 달라지는 경우에는 근본적인 심각한 성능 저하를 나타내고 있다. 다양한 환경에서도 강인한 성능을 보이는 음성 인식기의 개발을 위해 인식 성능 저하를 막는 다양한 방법들이 시도되었으나, 여전히 해결해야 할 과제로 남아있다. 강인한 음성인식기의 목표는 가능한 훈련과 인식 환경 사이의 불일치를 발생시키는 요인들의 제거하여 두 환경 사이를 가깝게 함에 의해 인식성능의 향상을 얻는 것이다.

다양한 부가잡음과 채널 왜곡은 발성 환경을 변화시키는 주된 요인이다. 이러한 불일치를 제거하기 위해 다양한 방법들이 제안되었고, cepstral mean subtraction(CMN), spectral subtraction(SS)[1]등이 대표적인 방법이다. 또한 stochastic matching 방법도 개발되었으며, 이 방법이 CMN, SS와 다른 점은 모델의 각 상태별로 보상벡터를 추정하여 보상한다는 점이다.

최근 잡음 환경에서의 전 세계 개발자들이 동일한 DB와 동일한 음성인식기를 사용하여 DSR(Distributed Speech Recognition)의 성능을 서로 비교하는 Aurora project가 진행되고 있다[2]. 이 프로젝트에서 모토로라에서 제안한 방식이 Advanced Front-end[3] 방법으로 현재 표준안으로 채택되었으며, 계속해서 잡음에 강인한 feature 추출을 통해 좀 더 높은 성능을 얻고자 많은 기관이 참여하고 있다. 또한 advanced front-end외에 stereo DB 기반의 SPLICE(Stereo-based Piecewise LInear Compensation for Environments)라는 방법[4][5]도 상당히 좋은 성능을 보였다.

본 논문에서는 feature space에서 불일치를 감소시키기 위해 화자적응에서 사용되는 eigenvoice 방법[6]에 기반한 잡음 보상 방법을 개발하였다. Eigenvoice에서 사용한 방식과 유사하게 각각의 잡음 환경에 대해 깨끗한 환경과의 차이를 K개의 eigenvector의 선형 결합(linear combination)으로 표현하도록 하였다. 이 방식에서 잡음 모델(noisy model)을 구성하는 것과 바이어스(bias) 모델을 정의하는 것이 성능 향상에 큰 영향을 미친다. 본 논문에서는 잡음 모델을 구성하기 위해 MAP(Maximum A Posteriori) 적응 방식과 stereo DB를 이용하는 두 가지 방식을 제안하였다.

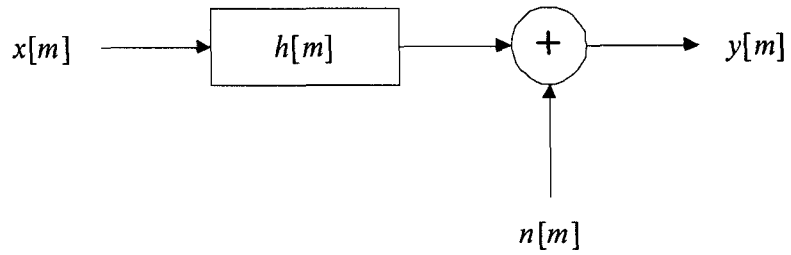
본 논문의 구성은 다음과 같다. 2 장에서는 음성신호의 잡음 왜곡 과정과 feature space에서 잡음을 제거하기 위한 기술로 Aurora 2 프로젝트에서 제안된 대표적인 방법인 SPLICE 에 대해 기술한다. 3장에서는 본 논문에서 제안한 feature space에서 잡음을 제거하기 위해 eigenvoice 방법을 응용한 eigen-environment 방법에 대해 기술한다. 4장에서 성능비교를 위한 task domain과 데이터베이스에 대해 기술한 후, Aurora 2 프로젝트에서 제안된 기존의 잡음 제거 기술들과 본 논문에

서 제안한 방법인 eigen-environment 방법과 성능을 비교한다. 마지막으로 5장에서 결론을 맺는다.

2. Stereo DB를 이용한 기존의 잡음제거 알고리즘

2.1. 음성신호 잡음 왜곡 과정 개요

음성인식에서 훈련과 인식시의 환경의 불일치는 성능을 저하시키는 중요한 요소 중의 하나이다.



<그림 1> 음성신호에 대한 잡음 왜곡 과정 모델

<그림 1>은 선형 필터 환경의 간단한 모델을 나타낸다. 잡음이 섞이지 않은 원음성 $x[m]$이 임펄스 응답 $h[m]$을 가지는 선형 채널을 거치고, 여기에 부가잡음 $n[m]$이 더해져서 왜곡된 음성 $y[m]$을 생성시킨다. $x[m]$, $h[m]$, $n[m]$, $y[m]$의 스펙트럼을 각각 $X(\omega)$, $Y(\omega)$, $N(\omega)$ 및 $Y(\omega)$라 하면, 이들은 다음과 같이 나타낼 수 있다.

$$Y(\omega) = H(\omega) \cdot X(\omega) + N(\omega) \tag{1}$$

식 (1)은 켑스트럼 영역에서 다음과 같이 표현될 수 있다.

$$\mathbf{y} = \mathbf{x} + \mathbf{g}(\mathbf{x}, \mathbf{h}, \mathbf{n}) \tag{2}$$

$$\mathbf{g}(\mathbf{x}, \mathbf{h}, \mathbf{n}) = \mathbf{h} + \mathbf{C} \ln(\mathbf{I} + \exp[\mathbf{C}^T(\mathbf{n} - \mathbf{h} - \mathbf{x})]) \tag{3}$$

여기서 $\mathbf{y}$ 는 관측 켑스트럼 벡터, $\mathbf{h}$ 는 채널 필터의 켑스트럼 벡터, $\mathbf{n}$ 은 부가잡음의 켑스트럼 벡터, 그리고 $\mathbf{x}$ 는 입력음성의 켑스트럼 벡터이고, $\mathbf{g}(\cdot)$는 켑스트럼 영역에서의 잡음에 대한 비선형(non-linear) 함수이다. 그리고, C는 DCT 행렬을 뜻한다. 따라서, 식 (2)에서 미지의 비선형 바이어스 항목으로 나타나는 $\mathbf{g}(\cdot)$를 제거해 줌으로써 훈련과 인식환경의 차이를 보상해 인식성능의 향상을 기대할 수 있다.

$\mathbf{g}(\cdot)$를 제거하기 위해 다양한 잡음 제거 알고리즘이 있으며, 본 논문에서는 제안한 잡음 제거 방법과 비교하기 위해 stereo DB 기반의 알고리즘인 SPLICE 방법에 대한 기술한다.

2.2 SPLICE 방법

2.2.1. 음성 모델과 왜곡

SPLICE 방법은 다음과 같은 두 가지 가정을 전제로 한다. 첫 번째 가정은 각각의 잡음음성의 특징 벡터 분포는 다음과 같은 M개의 가우시안 믹스쳐(Gaussian mixture)로 모델링 될 수 있다는 것이다.

$$p(\mathbf{y}) = \sum_{m=1}^{M} p(\mathbf{y} \mid m) p(m) \tag{4}$$

$$p(\mathbf{y} \mid m) = N(\mathbf{y} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{5}$$

여기서, $p(m)$, $\boldsymbol{\mu}_m$ 및 $\boldsymbol{\Sigma}_m$는 각각 m번째 가우시안 믹스쳐의 사전 확률, 평균 벡터 그리고 공분산 행렬이다. 그리고 각각의 잡음환경은 GMM(Gaussian Mixture Model)로 훈련된다.

두 번째 가정은 잡음음성 $\mathbf{y}$ 와 특정 믹스쳐가 주어졌을 때 원음성 $\mathbf{x}$ 의 조건부 확률 분포는 가우시안이라는 것이다.

$$p(\mathbf{x} \mid \mathbf{y}, m) = N(\mathbf{x}; \mathbf{y} + \mathbf{r}_m, \boldsymbol{\Gamma}_m) \tag{6}$$

여기서 $\mathbf{r}_m$ 과 $\boldsymbol{\Gamma}_m$는 깨끗한 원음성과 잡음이 섞인 음성 두 가지가 동시에 녹음된 스테레오 데이터를 사용하여 구한 보상 벡터 및 공분산 행렬이다.

즉, 식(6)은 식(2)에서 비선형 함수를 M개의 가우시안 믹스쳐를 이용하여

piecewise linear 함수로 근사화함을 의미한다.

2.2.2. SPLICE 훈련

각각의 잡음에 대해 잡음음성의 분포 $p(\mathbf{y})$도 또한 가우시안 믹스쳐로 모델링 될 수 있으며, 각각의 믹스쳐에서 분포 $p(\mathbf{x}\mid\mathbf{y},m)$에 대한 보상 벡터 $\mathbf{r}_m$는 스테레오 데이터가 주어진다면, maximum likelihood criterion에 의해서 다음과 같이 추정할 수 있다.

$$\mathbf{r}_m = \frac{\sum_n p(m\mid\mathbf{y}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(m\mid\mathbf{y}_n)} \tag{7}$$

여기서

$$p(m\mid\mathbf{y}_n) = \frac{p(\mathbf{y}_n\mid m)p(m)}{\sum_m p(\mathbf{y}_n\mid k)p(m)} \tag{8}$$

이다.

2.2.3. 특징 벡터 보상

2.2.1절에서의 두 가지 가정은 SPLICE 방법에서 잡음음성이 주어졌을 때 원음성의 Minimum Mean Square Estimation(MMSE)을 간단하게 해준다. 잡음음성이 주어졌을 때 구한 원음성의 MMSE는 다음과 같이 정리된다.

$$\hat{\mathbf{x}}_{MMSE} = E_x[\mathbf{x}\mid\mathbf{y},m] = \mathbf{y} + \sum_m p(m\mid\mathbf{y})\mathbf{r}_m \tag{9}$$

즉, 원음성은 각각의 믹스쳐에 관계된 보상 벡터들의 가중 합에 의해 표현될 수 있다. 빠른 구현을 위해서 식 (6)의 $p(m\mid\mathbf{y})$는 다음과 같이 간략화 할 수 있다.

$$\hat{p}(m\mid\mathbf{y}) = \begin{cases} 1 & m = \arg\max_m p(m\mid\mathbf{y}) \\ 0 & otherwise \end{cases} \tag{10}$$

SPLICE 잡음 보상은 다음의 두 단계로 적용된다. 첫 단계에서 잡음음성의 매

프레임마다 식 (10)에 의해 최적 믹스쳐를 찾는다. 다음 단계로 식 (9)을 사용하여 그 믹스쳐에 대응하는 보상 벡터를 잡음음성의 특징 벡터에 더해준다.

3. 환경 공간에서의 잡음 보상

3.1 Eigen-environment를 이용한 잡음보상방법

화자 적응 방법 중 MAP나 MLLR의 경우는 acoustic space에서 적응이 이루어지지만, eigenvoice의 경우에는 speaker space에서 적응이 이루어지는 차이점이 있다. 그리고, 대부분 잡음보상 방법들도 feature space에서 수행된다. 본 논문에서는 잡음보상을 feature space에서 수행하지 않고 eigenvoice와 같이 잡음을 나타내는 다른 space에서 잡음을 보상하는 방법을 고려하였다.

<그림 2>에 본 논문에서 제안한 eigen-environment에 방법에 대해 개략적으로 나타내었다. 이는 eigenvoice 방법과 유사하다.

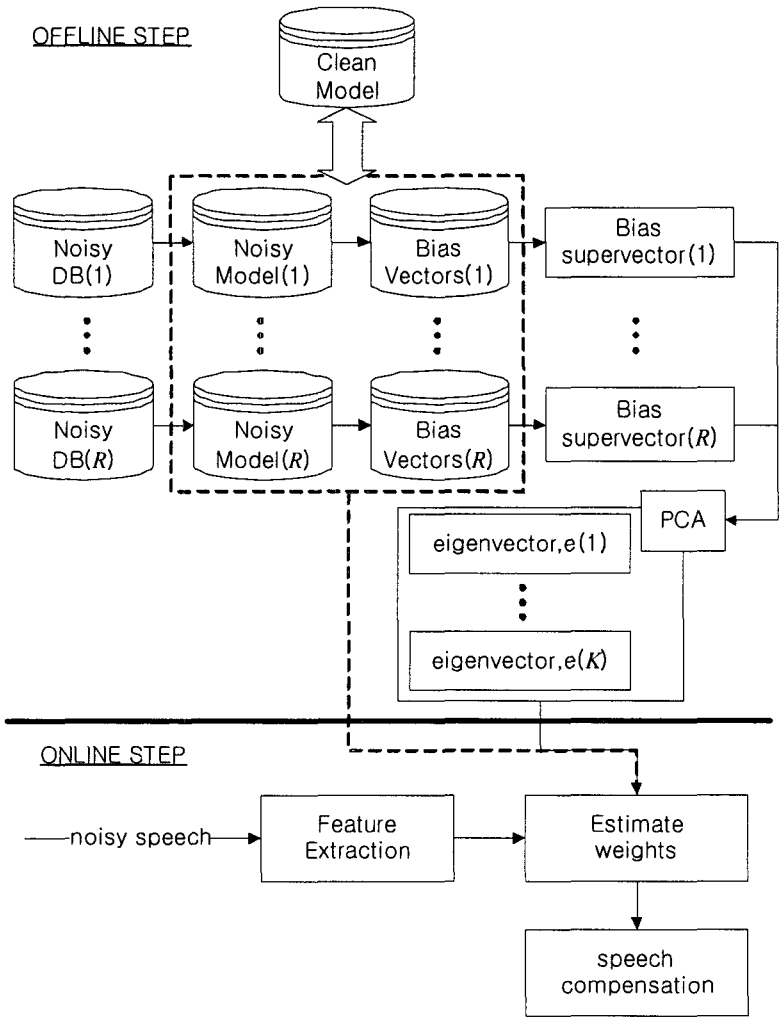
먼저 다양한 잡음 환경간의 변화를 알려주는 사전분포를 알기 위해 오프라인으로 R개의 잡음 환경에 대해서 M 개의 가우시안 믹스쳐를 구하고, 동시에 각각의 믹스쳐별로 깨끗한 모델과의 차이를 나타내는 M 개의 보상벡터 set도 구성한 후, M 개의 보상벡터들을 연결하여 supervector로 만든다. Supervector를 구성할 때 주의할 점은 supervector들의 각 차원간에는 유사한 성분을 나타내도록 하는 것이 중요하다. supervector의 차원 L는 GMM 믹스쳐 개수(M)에 특징벡터의 차원(D)을 곱한 값이 된다.

다음은 차원 축소를 위해서 PCA를 적용한다. 각각의 차수가 L인 R개의 supervector에 PCA를 적용하면 R개의 eigenvector가 생성된다. 이때 생성된 eigenvector를 eigen-environment라고 본 논문에서는 이름을 결정하였다. D차원의 벡터 x에 대해서, 만약 PCA로부터 생성된 K개의 주성분들이 R개의 분포의 대부분의 변이를 설명할 수 있다면, 다음과 같이 새로운 K개의 eigenvector의 가중합으로 x를 표현할 수 있을 것이다.

$$\mathbf{x} \cong \mathbf{e}(0) + \sum_{k=1}^{K} w(k)\mathbf{e}(k) \tag{11}$$

구성된 eigenvector를 사용하여 실제 잡음 음성이 인식기로 들어오는 경우 online으로 보상벡터를 추정하여 잡음 성분을 보상하게 된다. 이 때 offline에서 구성한 잡음 모델과 바이어스 벡터 set도 가중치 추정시 사용할 수도 있다. 즉, SPLICE와 같이 사후 확률(posteriori probability)를 구할 때 사용할 수 있으며, 또한

환경 선택 시에도 사용할 수 있다.



<그림 2> Eigen-environment 의 개략도

3.2 Eigen-environment에서 weight 추정 방법

Eigen-environment의 가중치를 추정하는 식을 유도하기 전에 잡음 보상을 위한 기본적인 수식을 도입한다. 먼저, 잡음 음성은 다음과 같이 M개의 가우시안 믹스쳐로 모델링 된다고 가정한다.

$$p(\mathbf{y}) = \sum_{m=1}^{M} p(\mathbf{y} \mid m)p(m) = \sum_{m=1}^{M} N(\mathbf{y}, \boldsymbol{\mu}_y^m, \boldsymbol{\Sigma}_y^m)p(m) \tag{12}$$

여기서, $p(m)$, $\boldsymbol{\mu}_y^m$ 및 $\boldsymbol{\Sigma}_y^m$는 각각 m번째 가우시안 믹스쳐의 사전확률, 평균벡터 그리고 공분산 행렬이다. 만약 x와 y가 믹스쳐 m 내에서의 joint Gaussian이라면, $p(\mathbf{x} \mid \mathbf{y}, m)$는 다음의 평균을 가지는 가우시안 분포이다.

$$\begin{aligned} E[\mathbf{x} \mid \mathbf{y}, m] &= \boldsymbol{\mu}_x^m + \boldsymbol{\Sigma}_{xy}^m (\boldsymbol{\Sigma}_y^m)^{-1} (\mathbf{y} - \boldsymbol{\mu}_y^m) \\ &= \boldsymbol{\Sigma}_{xy}^m (\boldsymbol{\Sigma}_y^m)^{-1} \mathbf{y} + (\boldsymbol{\mu}_x^m - \boldsymbol{\Sigma}_{xy}^m (\boldsymbol{\Sigma}_y^m)^{-1} \boldsymbol{\mu}_y^m) = \mathbf{C}_m \mathbf{y} + \mathbf{r}_m \end{aligned} \tag{13}$$

여기서, 믹스쳐 m에서 $\mathbf{C}_m = \boldsymbol{\Sigma}_{xy}^m (\boldsymbol{\Sigma}_y^m)^{-1}$는 회전 행렬이고, $\mathbf{r}_m = \boldsymbol{\mu}_x^m - \boldsymbol{\Sigma}_{xy}^m (\boldsymbol{\Sigma}_y^m)^{-1} \boldsymbol{\mu}_y^m$은 보상벡터이다. 그리고, 잡음음성 y에 대해 MMSE (Minimum Mean Square Error)를 사용하여 원음성 x를 추정하면, 다음 식과 같이 나타난다.

$$\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x} \mid \mathbf{y}] = \sum_m p(m \mid \mathbf{y}) E[\mathbf{x} \mid \mathbf{y}, m] \tag{14}$$

식 (13)를 식(14)에 대입하고 식(13)에서 회전 행렬 $\mathbf{C}_m$을 단위행렬로 가정하면 다음과 같이 추정된다.

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_m p(m \mid \mathbf{y}) r_m \tag{15}$$

또한, 식 (13)에서 보상벡터는 다음과 같이 간단하게 표현된다.

$$\mathbf{r}_m = \boldsymbol{\mu}_x^m - \boldsymbol{\mu}_y^m \tag{16}$$

따라서, 잡음 환경에서 원음성의 추정은 식 (15)에서 보상벡터와 잡음 음성에 대한 각 믹스쳐의 사후확률만 찾는다면 원음성을 추정할 수 있을 것이다.

GMM 모델 기반의 대부분의 보상방법[7]-[9]들은 식(16)에서 $\mathbf{r}_m$를 추정하는 방법이 다르다. 이러한 방법들은 $\mathbf{r}_m$를 online 또는 offline으로 추정하는가에 따라 분류할 수 있고, 또한 offline의 경우 stereo DB를 사용하는가 아니면 non-stereo DB를 사용하는가에 따라 분류할 수 있다.

CDCN[6], SPLICE, stereo-based RATZ[8] 등은 offline에서 stereo DB를 사용하여 미리 $\mathbf{r}_m$를 구하여 보상벡터를 각 mixture 마다 미리 가지고 있다. Blind RATZ[8]는 offline에서 non-stereo DB를 사용하여 보상벡터를 구해놓는다. 이 경우에서는 $\boldsymbol{\mu}_y^m$만을 추정하면 된다. VTS[9]의 경우는 online으로 $\boldsymbol{\mu}_y^m$를 추정함으로써 원음성을 추정하는데 사용한다.

본 논문에서 제안한 eigen-environment 방식에서는 식(16)에서 $\mathbf{r}_m$를 online으로 식(11)과 같이 eigenvector들의 가중합으로 추정하도록 하였다.

$$\mathbf{r}_m = \mathbf{e}_m(0) + \sum_{k=1}^{K} w(k)\mathbf{e}_m(k) \tag{17}$$

식(17)의 가중치를 구하기 위해서는 EM 알고리즘을 사용한다. 먼저 Q 함수를 다음과 같이 정의한다.

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2} P(O \mid \lambda) \sum_m \sum_t p(m \mid \mathbf{y}_t) f(\mathbf{y}_t, m) \tag{18}$$

여기서, $\mathbf{y}_t$는 잡음 음성의 시간 t에서의 관측 벡터이다.

$$f(\mathbf{y}, m) = -d \log(2\pi) - \log \mid \boldsymbol{\Sigma}_y^m \mid + h(\mathbf{y}, m) \tag{19}$$

$$h(\mathbf{y}, m) = (\mathbf{y} - \boldsymbol{\mu}_y^m)^T (\boldsymbol{\Sigma}_y^m)^{-1} (\mathbf{y} - \boldsymbol{\mu}_y^m) \tag{20}$$

식(16)과 (17)의 관계를 이용하여 잡음 음성에 대한 평균을 다음과 같이 가정한다.

$$\boldsymbol{\mu}_y^m = \boldsymbol{\mu}_x^m + \mathbf{e}_m(0) + \sum_{k=1}^{K} w(k)\mathbf{e}_m(k) \tag{21}$$

식(21)을 (20)에 대입하여 각각의 가중치에 대해 Q(.) 함수를 미분하면 eigenvoice에서 가중치 추정시 사용한 MLED[6]와 동일한 수식을 얻을 수 있으며, 방정식을 풀면 가중치를 추정할 수 있다. 추정된 원음성은 다음과 같이 나타낼 수 있다.

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_m p(m|\mathbf{y})\hat{\mathbf{r}}_m = \mathbf{y} + \sum_m p(m|\mathbf{y})\left(\mathbf{e}_m(0) + \sum_{k=1}^{K} \hat{w}(k)\mathbf{e}_m(k)\right) \tag{22}$$

여기서, $\hat{\mathbf{r}}_m$, $\hat{w}(k)$ 는 추정된 보상벡터와 추정한 k번째 eigen-environment의 가중치를 뜻한다.

따라서, 본 논문에서 제안한 eigen-environment 잡음 보상 방법은 SPLICE 또는 RATZ방식과는 달리 보상벡터가 정해져 있는 것이 아니라 VTS와 같이 입력되는 noisy speech에 따라 추정되는 가중치가 달라짐으로 인해 고정된 보상벡터를 사용하는 것보다 세세하게 잡음을 보상할 수가 있다.

3.3 잡음 모델 생성 방법

제안된 방식에서 가장 중요한 부분은 잡음 모델을 구성하는 방법과 그리고, eigen-environment를 얻기 위해 바이어스 supervector를 얻는 방법이다. 본 논문에서는 MAP 적응방법을 이용한 것과 stereo DB를 이용하는 두 가지 방법을 사용하였다. MAP 방식은 clean DB에 대한 GMM에 대해 잡음 음성에 대한 GMM을 MAP 방식을 통해 구성하는 것이다. 이 방법은 stereo DB가 필요하지 않는 장점을 가진다. 그리고, 바이어스 벡터 set은 식(16)과 같이 평균의 차이를 이용하여 각각의 믹스쳐마다 구성할 수 있다.

두 번째로 stereo DB를 이용하는 방식을 <그림 3>에 나타내었다.

깨끗한 DB로부터 깨끗한 GMM을 구성한 후, 깨끗한 DB를 깨끗한 GMM에 대해 alignment시켜 다음과 같이 프레임마다 믹스쳐 인덱스(mixture index)를 얻는다.

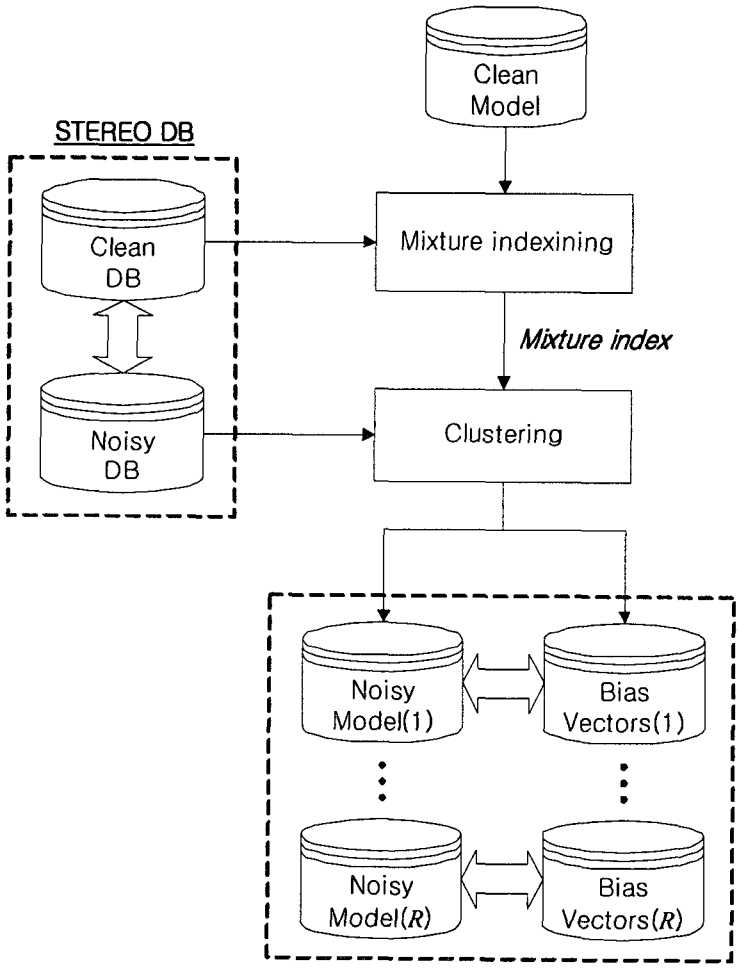
$$\hat{m}_t = \arg\max_m P(\mathbf{x}_t \mid M, \mathbf{\Lambda}_{\mathbf{x}}) \tag{23}$$

M개의 믹스쳐 중에 가장 높은 확률을 보이는 믹스쳐가 선택된다. 또한 잡음 DB는 깨끗한 DB와 동시에 stereo로 녹음되었으므로 잡음 음성의 각각의 프레임은 깨끗한 음성에서 구해진 믹스쳐 인덱스를 그대로 사용하면 된다. <그림 3>에서 clustering 모듈에서는 잡음 GMM과 바이어스 벡터 set을 구성한다. (24)와 같이 잡음 GMM을 구성하기 위해 각각의 잡음 환경 별로 깨끗한 음성과 잡음 음성의 각 프레임에 대해 믹스쳐 인덱스가 동일한 프레임들을 모아서 그 평균을 구한다.

$$\boldsymbol{\mu}_m^{\mathbf{y}} = E[\mathbf{y}_t \mid \hat{m}_t = m] \tag{24}$$

여기서, $E[\cdot]$는 기대값을 취하는 것을 의미한다. 그리고, (25)와 같이 각 믹스쳐별 바이어스 벡터는 잡음 음성과 원음성의 각 프레임간의 차이의 평균을 이용하여 구한다.

$$\mathbf{r}_m = E[\mathbf{x}_t - \mathbf{y}_t \mid \hat{m}_t = m] \tag{25}$$

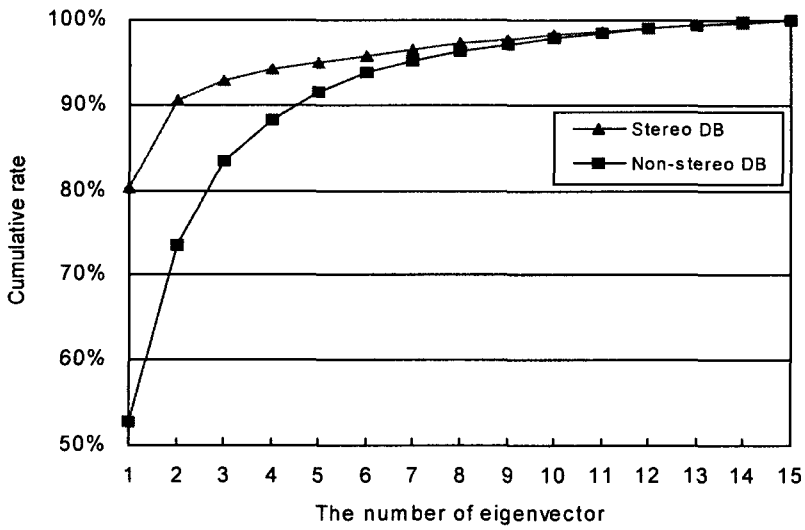


<그림 3> Stereo DB를 이용한 잡음 모델 및 바이어스 보상 벡터 구성에 대한 개략도

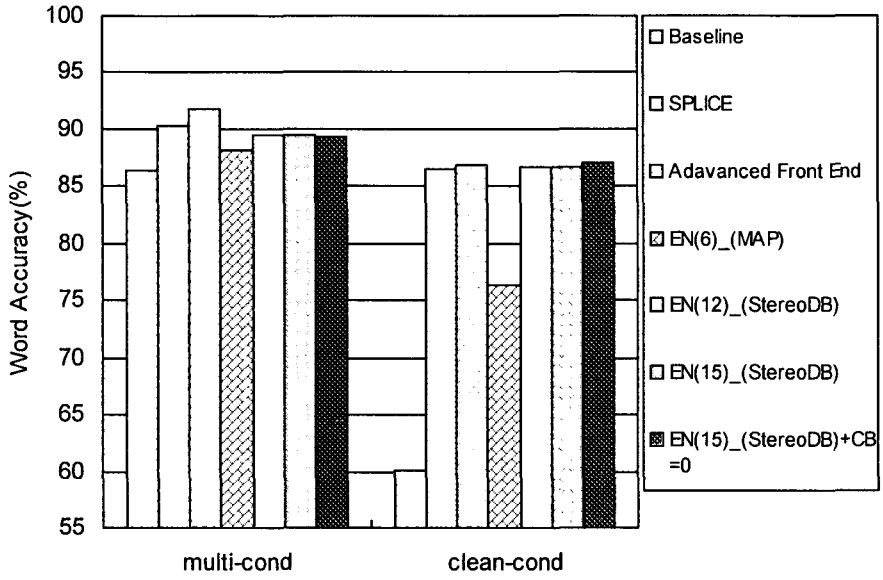
4. 실험 및 결과

제안된 방법의 성능 평가를 위해서 Aurora2 데이터베이스[2]가 사용되었다. Aurora2 데이터베이스는 1자리에서 7자리까지의 영어 연결숫자로 구성된 TI Digit에 다양한 잡음을 인공적으로 부가한 것이다. 인식 모델은 두 가지 방법으로 훈련이 되는데, 8440개의 깨끗한 발성을 사용하여 모델을 훈련하는 clean-condition과 동일한 발성을 20개의 잡음환경으로 나누어 각각의 환경별로 422개의 잡음 음성을 사용하여 훈련하는 multi-condition이 있다. 20개의 잡음환경은 4가지의 잡음종류(subway, babble, car, exhibition)와 각각 5 가지 잡음 레벨(clean, 20dB, 15dB, 10dB, 5dB)로 구성되어 있다. 테스트 데이터는 세 가지의 subset으로 구성되어 있는데, 훈련에 이용한 4가지 잡음 종류를 포함한 Set A와 훈련에 이용되지 않은 새로운 4가지 잡음 종류를 포함한 Set B, 그리고 훈련과 다른 채널 특성을 가지고 Set A와 Set B에 나타난 2가지 잡음을 포함한 Set C의 총 10종류 잡음으로 -5dB에서 clean까지의 7가지 잡음 레벨로 구성된다. 성능 평가는 각 잡음의 종류에 대해서 20dB에서 0dB까지의 잡음 레벨에 대해 수행된다.

먼저 eigen-environment가 environment space를 어느 정도 표현하는 지에 대한 정보를 얻고자 PCA를 한 후 eigenvector 수에 따른 eigenvalue값들의 누적분포를 <그림 4>에 나타내었다. <그림 4>에서 나타난 것과 같이 stereo DB의 경우에는 5개 정도를 사용하여 바이어스 변화의 95%정도를 나타냄을 알 수 있으며, non-stereo DB의 경우에는 7개 정도임을 알 수 있다. 본 논문에서는 eigenvector의 수를 변경시키면서 잡음 보상을 수행하였다.



<그림 4> Eigenvaule 수에 따른 누적율



<그림 5> 기존의 방법들에 대한 제안된 방법과의 성능 비교

<그림 5>에 제안한 방식과 기존의 방식에 대한 성능 비교를 실시하였다. 여기서 사용한 SPLICE 방법은 본 연구실에서 구현한 것[10]이므로 [5]에서 보인 실험결과와는 약간의 차이가 있다. Eigen-environment에서도 SPLICE와 동일하게 offline 시 구성된 잡음 모델을 사용하여 환경 선택에 이용하였다. 그리고, 바이어스 벡터 smoothing도 동일하게 적용하였다. <그림 5>에서 'EN(x)_(MAP/StereoDB)'는 x개의 eigen-environment를 사용한 경우를 뜻하고, 바이어스 벡터를 구성할 때 MAP을 기반으로 하였는지 또는 stereo DB를 사용하였는지를 나타낸다. 또한 'CB=0'는 환경 선택시 깨끗한 환경이 선택된 경우는 바이어스 값을 0이 되도록 제한을 둔 것을 의미한다.

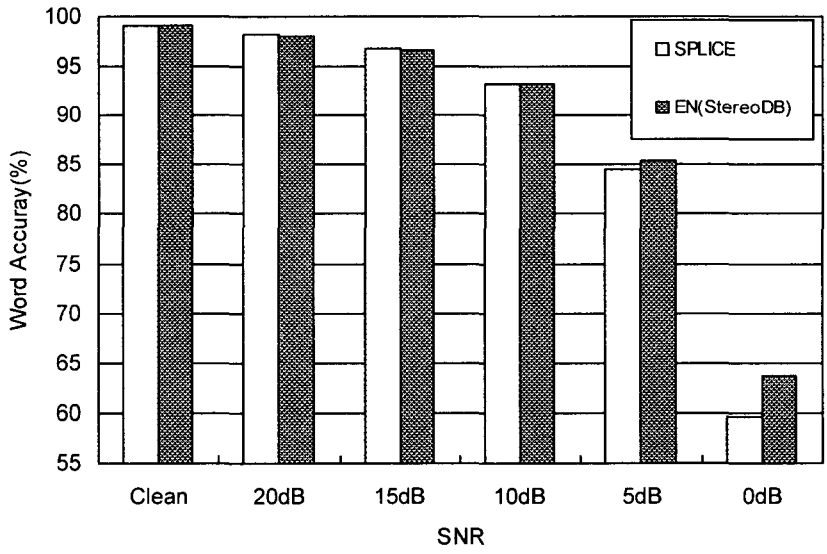
Clean condition인 경우에는 SPLICE나 advanced front-end보다 더 좋은 성능을 보인다. 그리고, non-stereo data를 이용한 MAP 방식으로 잡음 모델과 바이어스 벡터를 구성하는 경우 baseline에 비해 성능 향상을 있지만, stereo DB를 사용한 경우에 비해서는 낮은 인식 향상율을 보인다. 만약 MAP 방식의 경우에서도 사후 확률을 잘 추정할 수 있다면 좋은 성능을 낼 수 있다. 즉, 잡음 모델을 잘 만드는 것이 성능에 많은 영향을 미치는 것을 알 수 있다.

각각의 eigenvector 갯수에 따른 인식성능을 <표 1>에 나타내었다.

<그림 6>에 제안된 방식이 낮은 SNR에서 SPLICE보다 성능이 더 높은 것을 볼 수 있다. 즉, 잡음환경에 더욱더 강인함을 알 수 있다.

<표 1> Eigen-environment 수에 따른 제안된 방법의 성능 향상율

Number of Eigenvoice	Multi-condition		Clean-condition		Average Improvement (%)
	Word Accuracy (%)	Relative Improvement (%)	Word Accuracy (%)	Relative Improvement (%)	
Baseline	86.39	-	60.06	-	-
1	89.39	22.01	81.42	53.47	37.74
3	89.34	21.70	81.90	54.68	38.19
6	89.60	23.56	86.62	66.49	45.03
10	89.51	22.93	86.63	66.53	44.73
12	89.50	22.89	86.70	66.69	44.79
15	89.52	22.99	86.72	66.74	44.86



<그림 6> Clean-condition 훈련조건에서 SNR에 따른 SPLICE 방식과 제안된 방법의 성능비교

5. 결 론

본 논문에서는 feature space에서 훈련 및 인식 환경 사이의 불일치를 감소시키기 위해 eigenvoice 방법에 기반한 새로운 잡음 보상 방법을 개발하였다. Eigenvoice에서 사용한 방식과 유사하게 각각의 잡음 환경에 대해 깨끗한 환경과의 차이를 K개의 eigenvector의 선형 결합으로 표현하도록 하였다. 이 방식에서 잡음 모델을 구성하는 것과 바이어스 모델을 정의하는 것이 성능 향상에 큰 영향을 미친다. 본 논문에서는 잡음 모델을 구성하기 위해 MAP 적응 방식과 stereo DB를 이용하는 두 가지 방식을 제안하였다.

Aurora 2 DB에 대해 실험을 수행한 결과, 제안한 방법인 eigen-environment 잡음 보상방법이 baseline 시스템에 비해 44.86% 의 인식성능 향상률을 얻었다. 특히, clean condition 훈련 환경에서는 제안한 방법이 Aurora 2 프로젝트에서 기존의 제안된 방법들보다 높은 성능을 보였다.

참 고 문 헌

[1] A. Sanker and C. H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition", *IEEE Trans. Speech and Audio Processing*, vol.4, no.3, pp.190-202, May 1996.

[2] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *ISCA ITRW ASR2000 Automatic speech recognition: Challenges for the next millennium*, Paris, Sep. 2000.

[3] D. Macho, L. Mauuary, et al., "Evaluation of a noise-robust DSR front-end on Aurora databases", *Proc. ICSLP*, Denver, pp.17-20. Sep. 2002.

[4] L. Deng, A. Acero et al., "Large vocabulary continuous speech recognition under adverse conditions", *Proc. ICSLP*, Beijing, vol.3, pp.806-809, Oct. 2000.

[5] J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database (web update)", *Proc. Eurospeech*, Aalborg, pp.217-220, Sep. 2001.

[6] R. Kuhn, P. Nguyen et al., "Eigenvoices for speaker adaptation", *Proc. ICSLP*, vol.5, pp.1771-1774, Nov. 1998.

[7] A. Acero, "Acoustical and Environmental Robustness in automatic speech recognition", Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Sep. 1990.

[8] P. Moreno, B. Raj and R.M. Stern, "A unified approach to robust speech recognition", *Proc. Eurospeech*, Madrid, Spain, pp.480-484, Sep. 1995.

[9] P. J. Moreno, B. Raj and R. M. Stern, "A vector Taylor series approach for

environment-independent speech recognition", *Proc. ICASSP*, vol.1, pp.733-736, 1996.
[10] 김두희, 송화전, 김형순, "음성학적인 정보를 포함한 SPLICE를 이용한 잡음환경에서의 음성 인식", *한국음향학회 하계학술발표대회 논문집*, 제 21권 제 1 호, pp.83-86, 2002년 7월.

접수일자 : 2004년 11월 19일
게재결정 : 2004년 12월 10일

▶ 송화전(Hwa Jeon Song)
주소: 609-735 부산광역시 금정구 장전동 산30번지 부산대학교
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-1704
E-mail: hwajeon@pusan.ac.kr

▶ 김형순(Hyung Soon Kim)
주소: 3609-735 부산광역시 금정구 장전동 산30번지 부산대학교
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr