SPLICE 방법에 기반한 잡음 환경에서의 음성 인식 성능 향상*

김종현(부산대), 송화전(부산대),
이종석(튜브미디어), 김형순(부산대)

<차 례>

<Abstract>

Performance Improvement of Speech Recognition Based on SPLICE in Noisy Environments

Jong Hyeon Kim, Hwa Jeon Song, Jong Seuk Lee, Hyung Soon Kim

The performance of speech recognition system is degraded by mismatch between training and test environments. Recently, Stereo-based Piecewise LInear Compensation for Environments (SPLICE) was introduced to overcome environmental mismatch using stereo data. In this paper, we propose several methods to improve the conventional SPLICE and evaluate them in the Aurora2 task. We generalize SPLICE to compensate for covariance matrix as well as mean vector in the feature space, and thereby yielding the error rate reduction of 48.93%. We also employ the weighted sum of correction vectors using posterior probabilities of all Gaussians, and the error rate reduction of 48.62% is achieved. With the combination of the above two methods, the error rate is reduced by 49.61% from the Aurora2 baseline system.

* Keywords : SPLICE, Robust speech recognition.

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

1. 서 론

음성은 인간의 가장 편리한 의사전달 수단이며, 차세대 사용자 인터페이스를 위한 핵심 요소 기술로 그 필요성이 더욱 증대되고 있다. 특히 각종 멀티미디어의 발달로 다양한 음성 관련 응용 제품들의 상용화가 이루어지고 있으나, 실제 환경에서 주변 잡음의 영향으로 음성 인식 시스템의 성능 저하가 문제가 되고 있다. 이러한 문제를 해결하기 위해서 잡음 환경 보상 기술은 필수적이며, Aurora project[9]를 비롯하여 선진 각국의 연구 기관이 다양한 방법들을 연구하고 있다.

잡음 환경을 극복하기 위한 방법들은 크게 음질 개선(speech enhancement)방법과 모델 적응(model adaptation)방법의 두 가지 접근방법으로 나눌 수 있다. 첫 번째 음질 개선 방법은 특징 벡터 영역에서 불일치를 줄이는 방법으로 잡음 음성 특징 벡터로부터 관찰된 왜곡을 추정하여 이를 제거하는 방법이다. 여기에는 Cepstrum Mean Subtraction(CMS)[1], Histogram Equalization(HE)[2][3], Vector Taylor Series(VTS)[4][5] 등이 있으며 이들 방법은 기존의 음성 인식 시스템의 구조 변화 없이 전처리 기법으로 구현이 가능하다는 장점이 있다. 두 번째 방법인 모델 적응은 미리 훈련되어 있는 인식 모델을 입력 잡음 음성을 잘 표현할 수 있도록 적응시키는 것이다. 여기에는 Maximum Likelihood Linear Regression(MLLR)[6]와 Parallel Model Combination(PMC)[7] 등이 있다. 모델 적응방법은 정적/비정적 잡음을 다양하게 다룰 수 있고, 부가 잡음뿐만 아니라 채널 왜곡도 동시에 처리할 수 있지만, 계산량과 메모리가 상대적으로 많이 소요되는 단점이 있다.

최근 깨끗한 음성과 잡음 음성이 동시에 녹음된 stereo 데이터와 잡음의 Gaussian Mixture Model(GMM)을 이용한 Stereo-based Piecewise Linear Compensation for Environments(SPLICE)[8] 방식이 제안되어 우수한 성능을 보여주고 있다. SPLICE는 프레임 기반의 특징 벡터 영역에서의 잡음 제거 방법으로서, 잡음으로 인해 발생하는 왜곡을 잡음 음성의 GMM을 이용하여 부분 선형적으로 모델링한다. 그리고 이런 왜곡을 stereo 데이터를 이용하여 훈련시킨다. SPLICE 방법은 잡음에 대한 어떠한 가정도 하지 않기 때문에 부가 잡음 뿐만 아니라 채널 왜곡도 동시에 보상해줄 수 있다.

본 논문에서는 평균 벡터만을 보상하는 기존의 방법과 다르게 공분산 행렬 또한 보상하는 generalized SPLICE 방법을 제안한다. 또한 최적 가우시안 믹스처(Gaussian mixture)에 대한 보상 벡터만을 적용하는 기존의 방법과 다르게, 모든 가우시안 믹스처를 고려한 사후확률의 가중합으로 보상 벡터를 적용하는 방법에 대해서도 제안한다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서 기존의 SPLICE 방법을 개괄적으로 살펴보고, 3장에서 본 논문에서 제안한 generalized SPLICE 방법과 보상 벡터 가중합 방법에 대해서 설명한다. 4장에서 실험에 사용된 데이터 베이스에

대해 언급한 후 기존 방법과 제안한 방법의 실험 결과를 기술하며, 5장에서 결론을 맺는다.

2. SPLICE 방법

본 장에서는 천천히 변하는 잡음뿐만 아니라 급격히 변하는 잡음과 채널 왜곡까지 동시에 보상해줄 수 있는 SPLICE(Stereo-based Piecewise LInear Compensation for Environments)에 대해 설명하고자 한다[8].

2.1 음성 모델과 왜곡

SPLICE 방식은 잡음이 섞이지 않은 깨끗한 음성 $\mathbf{x}$와 부가 잡음과 채널에 의해 왜곡된 음성 $\mathbf{y}$에 대해서 다음의 두 가지 가정을 전제로 한다.

첫 번째 가정은 잡음 음성의 켑스트럼 벡터 분포가 M개의 Gaussian mixture로 모델링 될 수 있다는 것이다.

$$p(\mathbf{y}) = \sum_{k=1}^{M} p(\mathbf{y}|k)p(k) \tag{1}$$

여기서 $p(\mathbf{y}|k) = N(\mathbf{y};\ \mu_k,\ \Sigma_k)$이고 $p(k)$, μ_k 및 Σ_k는 각각 k번째 Gaussian mixture의 사전 확률, 평균 벡터 그리고 공분산 행렬이다. 각각의 잡음 환경마다 이러한 Gaussian 모델을 개별적으로 훈련시킨다.

두 번째 가정은 잡음 음성 $\mathbf{y}$가 주어졌을 때 깨끗한 음성 $\mathbf{x}$의 평균 벡터는 잡음 음성의 평균 벡터와 선형 변환의 관계를 가진다는 것이다. 이때 선형 행렬을 단위 행렬로 가정하면 원 음성의 잡음 음성에 대한 조건부 확률 분포는 다음과 같은 형태로 표현될 수 있다.

$$p(\mathbf{x}|\mathbf{y}, k) = N(\mathbf{x};\ \mathbf{y} + \mathbf{b}_k,\ \Sigma_k) \tag{2}$$

여기서 $\mathbf{b}_k$와 Σ_k는 k번째 Gaussian에 의존하는 보상 벡터와 추정된 원 음성의 공분산 행렬이다

2.2 훈련 과정

잡음 음성의 특징 벡터의 분포 $p(\mathbf{y})$는 Gaussian mixture를 따른다고 가정하였으므로 EM알고리즘을 이용하여 $\boldsymbol{\mu}_k$와 $\boldsymbol{\Sigma}_k$를 추정할 수 있고 초기 파라메터는 VQ clustering 을 이용하여 구할 수 있다. 분포 $p(\mathbf{x}|\mathbf{y},k)$에 대한 보상 벡터 $\mathbf{b}_k$는 stereo 데이터가 주어진다면 MMSE에 의해서 다음과 같이 추정된다.

$$\overline{\mathbf{b}}_k = \frac{\sum_n p(k|\mathbf{y}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(k|\mathbf{y}_n)} \tag{3}$$

여기서

$$p(k|\mathbf{y}_n) = \frac{p(\mathbf{y}_n|k)p(k)}{\sum_n p(\mathbf{y}_n|k)p(k)} \tag{4}$$

이다.

2.3 보상 과정

2.1절의 두 가정은 SPLICE 방식에서 잡음 음성에 대한 원 음성의 Minimum Mean Squared Error(MMSE) 추정을 간단하게 해준다. 잡음 음성이 주어졌을 때 MMSE로 추정한 원 음성의 조건부 기댓값은

$$\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x}|\mathbf{y}] = \sum_k p(k|\mathbf{y}) E_k[\mathbf{x}|\mathbf{y},k] \tag{5}$$

와 같이 주어지며, 다음과 같이 정리된다.

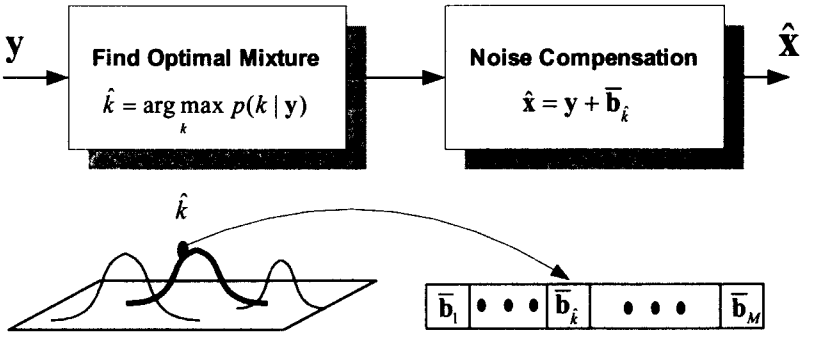
$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_k p(k|\mathbf{y})\ \overline{\mathbf{b}}_k \tag{6}$$

즉, 원 음성은 각각의 mixture에 대응되는 보상 벡터들의 가중 합으로 표현될 수 있다. 빠른 구현을 위해서 식 (6)의 $p(k|\mathbf{y})$는 식 (7)로 제한을 두면 식 (6)은 식 (8)로 간략화 될 수 있다.

$$\hat{p}(k|\mathbf{y}) = \begin{cases} 1 & \hat{k} = \arg\max_k p(k|\mathbf{y}) \\ 0 & otherwise \end{cases} \tag{7}$$

$$E_x[\ \mathbf{x}|\ \mathbf{y}, k] = \mathbf{y} + \overline{\mathbf{b}}_k \tag{8}$$

SPLICE는 <그림 1>과 같이 두 단계로 적용된다. 첫 단계에서 잡음 음성의 매 프레임마다 식 (7)에 의해 최적 mixture를 찾는다. 그리고 두 번째 단계에서 식 (8)과 같이 그 mixture에 대응하는 보상벡터를 잡음 음성의 특징 벡터에 더해주면 된다.



<그림 1> SPLICE 잡음 보상 과정

3. SPLICE 방법 기반의 성능 향상 방법

3.1 Generalized SPLICE

기존의 SPLICE 방법은 평균 벡터만을 보상하며, 공분산 행렬의 변화에 대해 보상하는 회전 행렬을 항등 행렬로 가정한다. 그 왜곡 모델은 다음과 같다[8].

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \overline{\mathbf{b}}_k \tag{9}$$

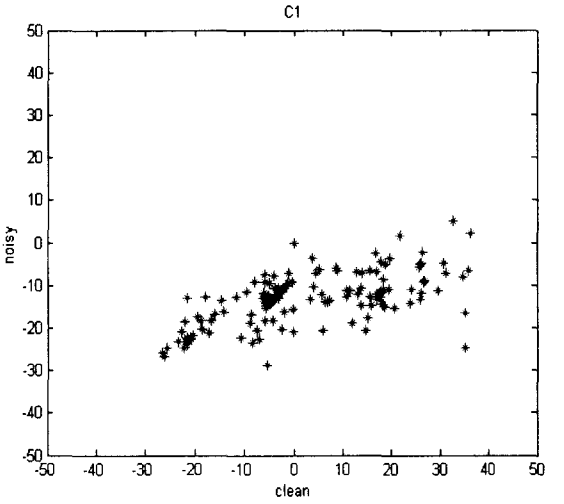
여기서 $\mathbf{y}$와 $\hat{\mathbf{x}}$는 각각 잡음 음성의 특징 벡터와 추정된 원 음성의 특징 벡터이며, $\overline{\mathbf{b}}$는 평균 벡터를 보상하기 위해 추정된 보상 벡터이다.

<그림 2>는 원 음성의 특징 벡터의 첫 번째 차원의 특정 Gaussian에서 주변 잡음 또는 채널 특성의 영향에 의해 잡음 음성의 특징 벡터로 변환된 형태를 보여주고 있다. SPLICE의 가정에 따르면 기울기 값이 1이 되어야 하지만 그보다 좀 더 작은 기울기 값을 관찰할 수 있다. 즉 기존의 SPLICE 방법의 회전 행렬에 대한 가정이 타당하지 않음을 알 수 있다.

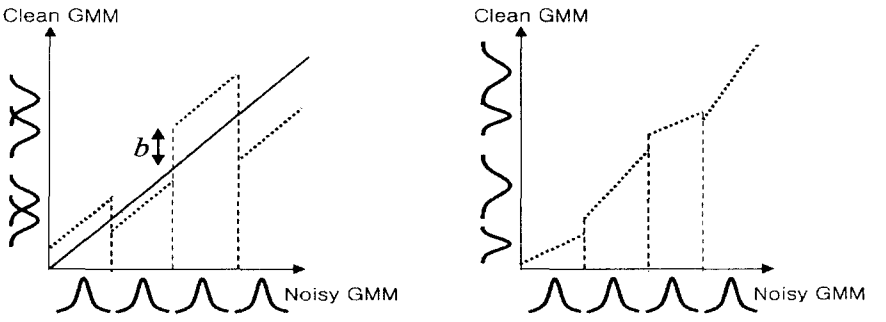
본 논문에서는 평균 벡터뿐만 아니라 공분산 행렬까지 보상할 수 있는 generalized SPLICE 방법을 제안하며, 식 (9)로부터 회전 행렬을 항등 행렬로 가정하지 않은 형태로 변형하면 다음 식과 같이 나타낼 수 있다.

$$\hat{\mathbf{x}}_{MMSE} = \overline{\mathbf{B}}_k \mathbf{y} + \overline{\mathbf{b}}_k \tag{10}$$

여기서 $\mathbf{y}$와 $\hat{\mathbf{x}}$는 식 (9)과 동일하며, $\overline{\mathbf{B}}$와 $\overline{\mathbf{b}}$는 각각 공분산 행렬을 보상하기 위해 추정된 회전 행렬과 평균 벡터를 보상하기 위해 추정된 보상 벡터이다. 여기에서 $\overline{\mathbf{B}}$가 항등 행렬이면 기존의 SPLICE와 동일함을 알 수 있으며, <그림 3>은 제안한 방법과 기존의 SPLICE 방법의 차이점을 보여준다.



<그림 2> 원 음성과 잡음 음성의 캡스트럼 첫 번째 차원(C1)의 상관관계



<그림 3> 기존의 SPLICE 방법(왼쪽)과 제안한 방법(오른쪽)의 차이

기존의 SPLICE 방법과 동일하게 generalized SPLICE도 훈련과정과 보상과정으로 나눌 수 있으며, 훈련 과정에서 분포 $p(\mathbf{x}|\mathbf{y},k)$에 대한 회전 행렬 $\overline{\mathbf{B}}$와 보상 벡터 $\overline{\mathbf{b}}$는 스테레오 데이터가 주어진다면, MMSE에 의해서 다음과 같이 추정될 수 있다

$$\overline{\mathbf{B}}_k = \mathbf{S}_{xy} / \mathbf{S}_{yy} \tag{11}$$

$$\overline{\mathbf{b}}_k = \tilde{\mathbf{x}} - \overline{\mathbf{B}}_k^T \tilde{\mathbf{y}} \tag{12}$$

여기서

$$\mathbf{S}_{xy} = \sum_n p(k|\mathbf{y}_n)\ \mathbf{x}_n^T \mathbf{y}_n - \sum_n p(k|\mathbf{y}_n)\ \tilde{\mathbf{x}}^T \tilde{\mathbf{y}} \tag{13}$$

$$\mathbf{S}_{yy} = \sum_n p(k|\mathbf{y}_n)\ \mathbf{y}_n^T \mathbf{y}_n - \sum_n p(k|\mathbf{y}_n)\ \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} \tag{14}$$

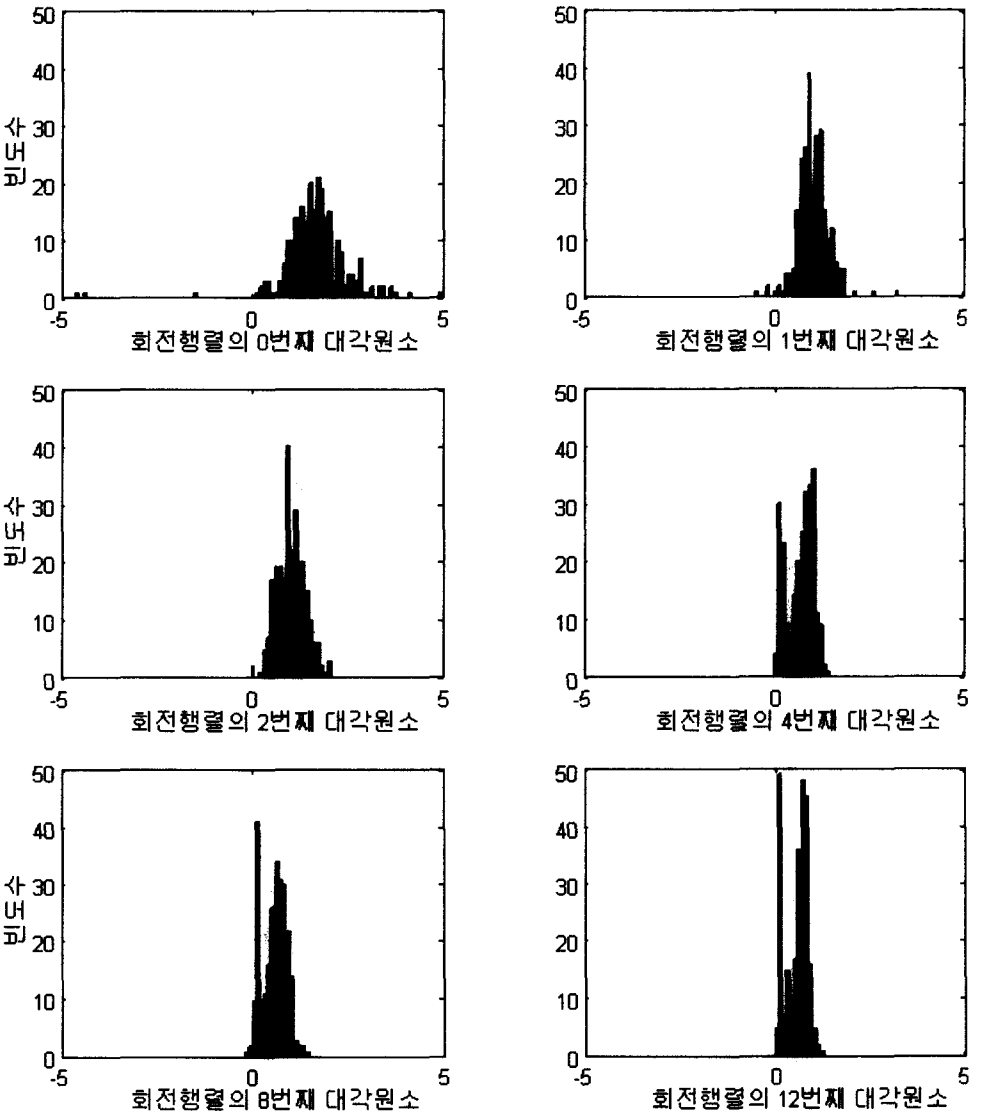
$$\tilde{\mathbf{x}} = \frac{\sum_n p(k|\mathbf{y}_n)\ \mathbf{x}_n}{\sum_n p(k|\mathbf{y}_n)} \quad \tilde{\mathbf{y}} = \frac{\sum_n p(k|\mathbf{y}_n)\ \mathbf{y}_n}{\sum_n p(k|\mathbf{y}_n)} \tag{15}$$

이다.

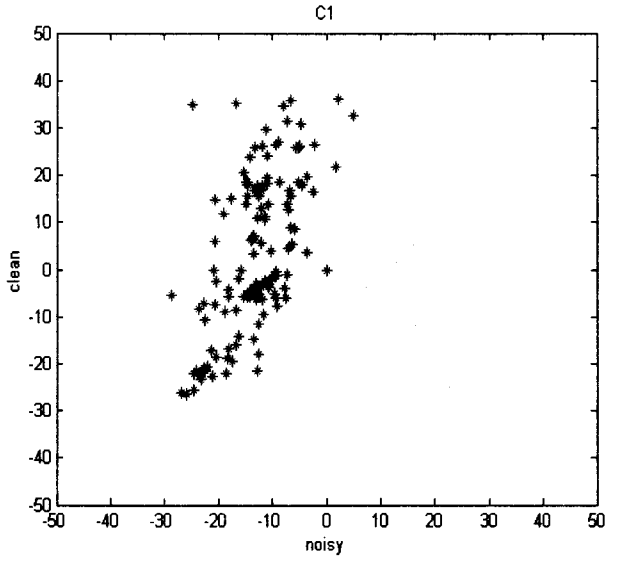
실험 결과 제안한 generalized SPLICE 방법을 모든 차원에 적용하였을 때 성능이 향상되지 않았다. <그림 4>는 추정된 회전 행렬의 각 대각 원소에 대한 히스토그램을 그려본 결과이다. 특징 벡터의 0,1,2번째 차원은 1 근처로 특징 벡터의 공분산 행렬이 크게 변하지 않은 형태로 추정이 잘 되었다고 할 수 있으나, 높은 차원에서 추정된 회전행렬은 0에 가까운 값, 즉 공분산 행렬이 크게 변하는 형태로 추정이 되는 경향이 커짐을 볼 수 있다. 특징 벡터의 높은 차원은 깨끗한 음성의 특징 벡터와 잡음 음성의 특징 벡터사이의 상관관계가 낮기 때문에 회전 행렬의 높은 차원의 대각원소가 잘 추정이 되지 않는다고 볼 수 있다. <그림 5>와 <그림 6>은 각각 특징 벡터의 첫 번째 차원과 12번째 차원에 대해 잡음 음성의 특징 벡터와 원 음성의 특징 벡터 사이의 상관관계를 나타낸 것이며, 높은 차원에서 두 특징 벡터 사이의 상관관계가 작은 것을 볼 수 있다.

본 논문에서는 위에서 언급한 관찰 결과를 바탕으로 회전 행렬의 보상은 특징 벡터의 0,1,2번째 차원에만 적용하고 나머지 특징 벡터의 차원들은 회전 행렬을 항등 행렬을 사용하는 방법을 제안한다.

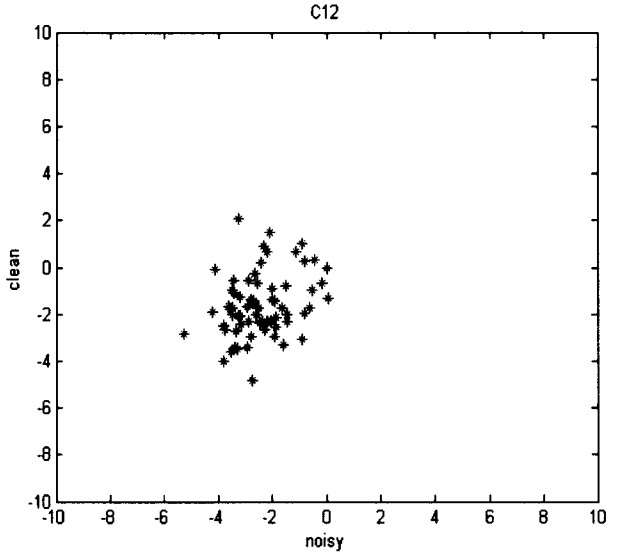
또한 원 음성에 잡음 음성이 더해질수록 분산이 줄어드는 경향을 보이며, 이러한 관찰 결과를 고려하여 <그림 7>과 같이 각 차원별로 분산을 보상하는 경우 항상 증가하는 방향으로 기울기를 추정하여 보상하는 방법을 제안한다



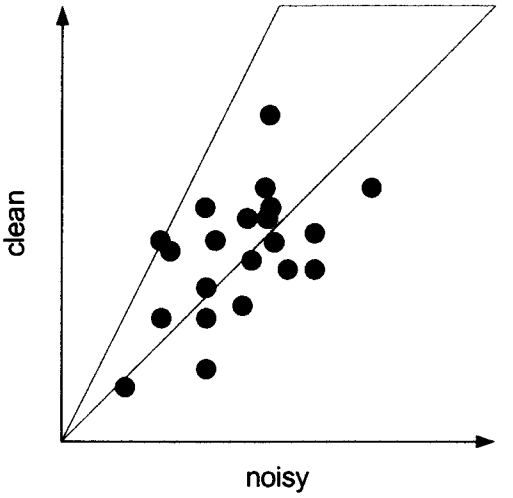
<그림 4> 회전 행렬의 각 대각원소의 히스토그램



<그림 5> 특징 벡터 1번째 차원에서 원 음성과 잡음 음성의 상관관계



<그림 6> 특징 벡터 12번째 차원에서 원 음성과 잡음 음성의 상관관계



<그림 7> 회전행렬의 특정 차원에 대한 제약

3.2 보상 벡터 가중합 방법

기존의 SPLICE 방법은 보상 과정의 시간을 단축하기 위해 보상 과정에서 최적의 Gaussian의 보상 벡터만을 이용하여 잡음 음성을 보상했다. 즉, 다음 식

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_k p(k|\mathbf{y})\ \bar{\mathbf{b}}_k \tag{16}$$

에서 사후확률을 다음과 같이 근사화 시켰다.

$$\hat{p}(k|\mathbf{y}) = \begin{cases} 1 & k = \arg\max_k p(k|\mathbf{y}) \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

하지만 식 (16)과 (17)의 근사식에서 최적 Gaussian을 찾기 위해서는 어차피 모든 Gaussian에 대해 사후확률을 구해야 하기 때문에 보상 과정에서 많은 시간이 단축되지는 않으며, 최적 Gaussian 하나만을 이용하기 때문에 또 다른 오차가 발생하게 된다.

본 논문에서는 식 (16)의 사후확률을 식 (17)과 같이 근사화 하지 않고 식 (16)과 같이 기존의 방법대로 보상 벡터를 모든 Gaussian에 대해 가중합 형태로 구하여 원 음성의 특징 벡터를 추정하는 방법을 제안한다.

4. 실험 및 결과

4.1 음성 데이터 베이스

Aurora 2 DB[9]는 미국 영어 화자에 의해서 발성된 한 자리에서 일곱 자리까지의 연속 숫자로 구성된 TI Digit DB에 실제 환경의 잡음이 더해진 것으로 깨끗한 음성과 잡음 음성 모두 채널을 통과한 데이터들이다.

Aurora 2 DB는 훈련 DB와 테스트 DB로 구성되어 있다. 테스트 DB는 세 개의 subset으로 구성되며 set A는 suburban train, babble, car 그리고 exhibition hall의 4가지 종류의 잡음으로 구성되어 있으며, set B는 restaurant, street, airport 그리고 train station으로 구성되어 있다. Set A와 set B는 모두 G.712의 채널 특성을 가진 필터를 통과한 것이다. 그리고 서로 다른 채널 환경에 대한 성능 평가를 위해서 MIRS(Modified Intermediate Reference System) 채널 특성을 가진 필터를 통과하고 suburban train과 street 잡음이 각각 더해진 set C가 있다. 각 subset들은 각 잡음 종류에 대해서 clean을 포함해서 20dB, 15dB, 10dB, 5dB, 0dB, -5dB의 SNR을 고려하여 더해지며, 각 잡음 종류와 레벨에 대해 52명의 남성과 여성화자에 의해 발성된 1001개의 연속 숫자로 이루어져있다.

훈련 DB는 clean-condition과 multi-condition으로 나누어져 있다. Clean-condition training은 잡음이 더해지지 않은 깨끗한 8440개의 발성으로 이루어졌다. Multi-condition training은 clean training DB를 각각 422발성으로 이루어진 20개의 subset으로 나누어서 clean과 20dB, 15dB, 10dB, 5dB의 5가지의 SNR에 대해서 set A에 사용되었던 4종류의 잡음을 각각 더한 것이다. 훈련 DB는 각각 55명의 남성 화자와 여성화자로 구성되어 있다.

Aurora 2 baseline 시스템은 특징 벡터 추출을 위해 WI007 front-end[10]를 사용한다. WI007 front-end는 12차 MFCC와 0차 켑스트럼 그리고 log energy를 추출할 수 있다. Aurora 2 baseline 시스템은 12차 MFCC와 log energy 에 각각의 delta와 delta-delta 파라메터를 포함하여 총 39차 MFCC를 사용하였다. 인식 모델은 [9]에 미리 정의되어 있는 HTK스크립트를 사용하였으며, 이는 각 숫자당 16개의 상태를 사용하며 각 상태당 3개의 Gaussian을 사용한다. 단 silence 모델은 3개의 상태를 사용하며 상태당 6개의 Gaussian을 사용한다.

4.2 제안한 방법의 실험 결과

기존의 SPLICE 방법이 평균 벡터만을 보상하는 것과는 달리 공분산 행렬까지 보상하는 generalized SPLICE 방법은 켑스트럼의 0,1,2번째 차원에만 분산을 보상하였고 나머지 차원은 기존의 SPLICE 방법과 동일한 방법을 적용하였다. 회전 행렬

의 각 대각원소의 추정치에 대한 제약은 분산이 늘어나는 방향인 1이상으로 제한하였으며 over-estimation이 되는 것을 막기 위해 2보다 큰 추정치에 대해서는 강제적으로 2를 대입하였다. 또한 SPLICE 전단계에 CMS를 적용하는 CMS SPLICE 방법[12]을 적용하였다. 실험 결과는 <표 1>에서 볼 수 있으며 baseline 실험 결과[9]에 비해 48.93%의 상대적 인식 향상율을 얻었으며, 이는 CMS SPLICE 방법의 상대적 인식 향상율인 47.60%에 비해 조금 개선된 결과이다.

<표 1> Generalized SPLICE 방법의 요약된 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.78	89.58	90.57	90.66
Clean Only	86.97	86.70	85.78	86.62
Average	89.38	88.14	88.17	88.64

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	32.56%	24.07%	41.85%	31.35%
Clean Only	66.30%	69.94%	58.00%	65.51%
Average	49.43%	47.00%	49.92%	48.93%

최적 mixture에 대한 보상 벡터 하나만을 사용하는 기존의 SPLICE 방법과는 다르게 모든 mixture에 대한 보상벡터를 사후확률의 가중합으로 보상을 하는 방법에 대한 실험 결과는 다음 표와 같으며, CMS SPLICE의 상대적 인식 향상율인 47.60%에 비해 48.62%로 조금 향상되었다.

<표 2> 보상 벡터의 가중합 방법의 요약된 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.67	89.65	90.33	90.59
Clean Only	86.72	86.82	85.73	86.56
Average	89.19	88.23	88.03	88.58

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	31.65%	24.59%	40.42%	30.89%
Clean Only	65.64%	70.22%	57.87%	65.35%
Average	48.64%	47.41%	49.14%	48.62%

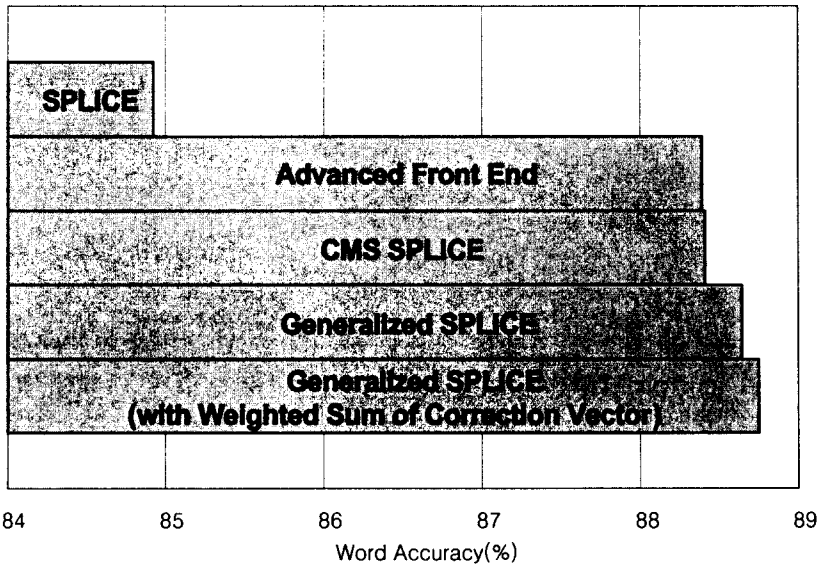
잡음 음성의 특징 벡터의 평균 벡터와 공분산 행렬 모두를 보상해주는 generalized SPLICE 방법과 보상 벡터의 사후 확률로 가중합 방법을 결합한 방법에 대한 실험 결과는 다음 표와 같으며, CMS SPLICE의 상대적 인식 향상율인 47.60%에 비해 49.61%로 향상되었다.

<표 3> Generalized SPLICE와 보상 벡터의 가중합을 결합한 방법의 요약된 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multi-condition	91.91	89.82	90.68	[illegible]
Clean only	86.94	86.80	85.82	[illegible]
Average	89.43	88.31	88.25	88.75

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multi-condition	33.62%	25.89%	42.57%	[illegible]
Clean only	66.22%	70.17%	58.11%	[illegible]
Average	49.92%	48.03%	50.34%	49.61%

<그림 8>은 기존의 방법과 제안한 방법의 실험 결과들을 비교한 그림이며, multi-condition 실험결과와 clean-condition 실험결과를 평균낸 결과이다. 제안한 Generalized SPLICE 방법이 Advanced Front End[11]와 CMS SPLICE 방법보다 성능이 더 좋음을 알 수 있다.



<그림 8> 기존의 방법과 제안한 방법의 성능 비교

5. 결 론

본 논문에서는 배경잡음과 채널 왜곡에 강인한 음성 인식을 위한 전처리 방법에 대해서 연구하였다. 이를 위해서 잡음 음성의 특징 벡터로부터 관찰된 왜곡을 추정하여 이를 제거하는 음질 개선 방법 중 최근 좋은 성능을 보이고 있는 SPLICE 방법을 검토하고 추가적인 성능 향상을 모색하였다.

SPLICE 방법은 잡음을 제거하기 위해 잡음 음성의 Gaussian mixture 모델과 깨끗한 음성과 동시에 녹음된 배경잡음이 섞인 잡음 음성인, stereo 데이터를 이용하여 잡음을 보상하는 방법이다. 이것은 정적인 잡음뿐만 아니라 비정적인 잡음도 제거해 줄 수 있으며, 채널 왜곡까지 동시에 보상해 줄 수 있다.

본 논문에서는 채널 왜곡 등에 의해서 발생할 수 있는 켑스트럼 영역에서의 바이어스 성분을 CMS를 적용하여 먼저 제거한 후에 SPLICE를 적용하는 CMS SPLICE 방법을 기반으로 하여 몇 가지 성능 향상 방법을 제안하였다.

기존의 SPLICE 방법은 평균 벡터만을 보상하며 공분산 행렬의 보상을 하지 않는다. 본 논문에서는 공분산 행렬까지 보상하기 위해 회전 행렬을 항등 행렬로 가정하지 않고, 직접 데이터로부터 계산하는 generalized SPLICE 방법을 제안하여 baseline 성능에 비해 48.93%의 상대적 인식 향상율을 얻었다.

또한 최적 mixture의 보상 벡터만을 이용하는 기존의 SPLICE 방법과는 달리 사후확률을 가중치로 모든 mixture의 보상 벡터를 더한 새로운 보상 벡터를 구하는 방법을 제안하였으며, 계산량의 큰 증가 없이 baseline 성능에 비래 48.62%의 상대적 인식 향상율을 얻었다.

그리고, 이들 두 가지 제안 방법들을 결합한 방법을 제안하였으며, CMS SPLICE의 상대적 인식 향상율 47.60%보다 향상된 49.61%의 성능을 나타내었다.

SPLICE를 이용한 방법이 우수한 성능을 보이지만, 한 프레임을 처리하기 위해 소요되는 많은 계산량으로 인한 실시간 처리의 어려움과 실제 다양한 잡음환경에 대한 stereo data의 필요성은 큰 제약이다. 앞으로 이런 제약을 완화하기 위한 방법들이 연구되어야 할 것으로 판단된다.

참고문헌

[1] B. Atal, "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification", *Journal of the Acoustical Society of America*, vol.5, no.6, pp.1304-1312, 1974.

[2] G. A. Saon and J. M. Huerta, "Improvements to the IBM Aurora 2 multi-condition system", *in Proc. of ICSLP*, pp.469-472, 2002.

[3] F. Hilger, S. Molau and H. Ney, "Quantile based histogram equalization for online applications", *in Proc. of ICSLP*, pp.237-240, 2002.

[4] P. J. Moreno, N. Raj and R. M. Stern, "A vector Traylor series approach for environment-independent speech recognition", *in Proc. of ICASSP*, pp.733-736, 1996.

[5] N. S. Kim, D. Y. Kim, B. G. Kong et al., "Application of VTS to environment compensation with noise statistics", *in Proc. of ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, Apr. 1997.

[6] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Computer, Speech and Language*, vol.9, pp.171-186, 1995.

[7] M. Gales and S. J. Young, "Parallel model combination on a noise corrupted resource management task", *in Proc. of ICSLP*, pp.255-258, 1994.

[8] J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora 2 database", *in Proc. of EUROSPEECH*, pp.217-220. Sep. 2001.

[9] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *ISCA ITRW AST2000* "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, Sep. 2000.

[10] ETSI standard document, "Speech Processing, Transmission and Quality aspects(STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", *ETSI ES* 201 108 v1.1.1 Feb. 2000.

[11] ETSI standard document, "Speech Processing, Transmission and Quality aspects(STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", *ETSI ES* 202 050 v1.1.1 Oct. 2002.

[12] 김두희, 송화전, 김형순, "음성학적인 정보를 포함한 SPLICE를 이용한 잡음환경에서의 음성 인식", *한국음향학회 하계학술발표대회 논문집* 제 21권 제 1호, pp.83-86, 2002.

접수일자 : 2005년 2월 10일
게재결정 : 2005년 3월 15일

▶ 김종현 (Jong Hyeon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: jhstudio@pusan.ac.kr

▶ 송화전 (Song Hwa Jeon)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: hwajeon@pusan.ac.kr

▶ 이종석 (Lee Jong Seok)
주소: 135-500 서울시 강남구 대치동 1024번지 나산빌딩 5층
소속: 튜브미디어
전화: 02) 3016-8500
E-mail: jslee@voiceware.co.kr

▶ 김형순 (Hyung Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr