강인한 음성인식을 위한 SPLICE 기반 잡음 보상의 성능향상*

Performance Improvement of SPLICE-based Noise Compensation for Robust Speech Recognition

김 형 순** · 김 두 희**
Hyung Soon Kim · Doo Hee Kim

ABSTRACT

One of major problems in speech recognition is performance degradation due to the mismatch between the training and test environments. Recently, Stereo-based Piecewise LInear Compensation for Environments (SPLICE), which is frame-based bias removal algorithm for cepstral enhancement using stereo training data and noisy speech model as a mixture of Gaussians, was proposed and showed good performance in noisy environments. In this paper, we propose several methods to improve the conventional SPLICE. First we apply Cepstral Mean Subtraction (CMS) as a preprocessor to SPLICE, instead of applying it as a postprocessor. Secondly, to compensate residual distortion after SPLICE processing; two-stage SPLICE is proposed. Thirdly we employ phonetic information for training SPLICE model. According to experiments on the Aurora 2 database, proposed method outperformed the conventional SPLICE and we achieved a 50% decrease in word error rate over the Aurora baseline system.

Keywords: Robust Speech Recognition, SPLICE, Noise Estimation, Gaussian Mixture Model, Residual Noise

1. 서 론

배경잡음 및 채널왜곡으로 인한 음성인식의 성능 하락은 음성인식 시스템의 실용화를 위한 큰 장애 요인이다. 이 문제의 해결을 위해서 적절한 잡음 환경 보상 기술의 도입은 필수적이며, 현재 Aurora 프로젝트를 비롯한 여러 연구를 통해 다양한 방법들이 제안되고 있다[1].

잡음 환경을 극복하기 위한 방법들은 특징벡터 영역에서의 보상 방법과 모델적응 방법

* 본 논문은 부산대학교 교내연구비 지원에 의해 수행된 것입니다.
** 부산대학교 전자공학과

의 두 가지 접근 방식으로 크게 나눌 수 있다. 첫 번째 방법은 특징벡터 영역에서 불일치를 줄이는 방법으로 잡음 음성의 특징벡터로부터 관찰된 왜곡을 추정하여 이를 제거하려는 방법이다. 여기에는 RASTA[2], 스펙트럼 차감법[3], 켑스트럼 평균 차감법(Cepstral Mean Subtraction, CMS)[4], 그리고 Vector Taylor Series(VTS)[5] 등이 있으며 이들은 기존의 음성인식 시스템의 구조 변경 없이 전처리 기법으로 구현이 가능하다는 장점이 있다. 두 번째 접근 방법인 모델 적응 방법은 미리 훈련되어 있는 인식모델을 입력 잡음 음성을 이용하여 입력 잡음 특성을 잘 나타낼 수 있도록 적응시키는 것이며, Parallel Model Combination(PMC)[6] 방식이 대표적인 예이다. 모델 적응방법은 정적/비정적 잡음을 다양하게 다룰 수 있고, 부가 잡음뿐만 아니라 채널왜곡도 동시에 처리할 수 있지만 일반적으로 계산량과 메모리가 많이 소요되는 단점이 있다.

최근 깨끗한 음성과 잡음 음성이 동시에 녹음된 스테레오 데이터와 잡음의 Gaussian Mixture Model (GMM)을 이용한 Stereo-based Piecewise LInear Compensation for Environments (SPLICE) 방식이 제안되어 좋은 성능을 보여주고 있다[7]. SPLICE는 프레임 기반의 켑스트럼 영역에서의 잡음 보상 방법으로서, 잡음으로 인해 발생되는 왜곡을 잡음 음성의 GMM과 스테레오 데이터를 이용하여 모델링한다. SPLICE 방법은 잡음에 대한 어떠한 가정도 하지 않기 때문에 부가 잡음뿐만 아니라 채널 왜곡도 동시에 보상해 줄 수 있다.

본 논문에서는 기존의 SPLICE 방식의 성능을 더욱 향상시키기 위해서, 잡음 음성 모델링과 보상벡터 훈련에 음성학적 정보를 추가하여 잡음을 보상하는 방법과 2단계 SPLICE를 이용하여 잔류왜곡을 효과적으로 보상하는 방법을 제안한다. 그리고 CMS를 전 단계에 사용하여 더욱 성능이 향상된 SPLICE 구현에 대해서도 보고한다.

본 논문의 구성은 다음과 같다. 2 장에서는 SPLICE를 이용한 잡음 보상 방법에 대해서 살펴보고, 3 장에서는 본 논문에서 제안한 SPLICE 기반의 잡음 보상의 성능향상 방법에 대해서 설명한다. 4 장에서는 실험 및 결과를 기술하고, 마지막으로 5 장에서 결론을 맺는다.

2. SPLICE 방법

2.1. 음성 모델과 왜곡

SPLICE 방식[11]은 잡음이 섞이지 않은 깨끗한 음성의 켑스트럼 벡터 $\mathbf{x}$ 와 부가잡음과 채널에 의해 왜곡된 음성의 켑스트럼 벡터 $\mathbf{y}$ 에 대해서 다음의 두 가지 가정을 전제로 한다.

첫 번째 가정은 잡음 음성의 켑스트럼 분포가 M 개의 Gaussian mixture로 모델링 될 수 있다는 것이다.

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

여기서 $p(\mathbf{y} \mid k) = N(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 이고 $p(k)$, $\boldsymbol{\mu}_k$ 및 $\boldsymbol{\Sigma}_k$ 는 각각 k 번째 Gaussian mixture의 사전확률, 평균벡터 그리고 공분산 행렬이다. 각각의 잡음 환경마다 이러한 Gaussian mixture 모델(GMM)을 개별적으로 훈련시킨다.

두 번째 가정은 잡음 음성이 주어졌을 때 깨끗한 음성의 평균벡터는 잡음 음성의 평균벡터와 선형 변환의 관계를 가진다는 것이다. 이때 선형행렬을 단위행렬로 가정하면 원음성의 잡음 음성에 대한 조건부 확률 분포는 다음과 같이 표현될 수 있다.

$$p(\mathbf{x} \mid \mathbf{y}, k) = N(\mathbf{x}; \mathbf{y} + \mathbf{r}_k, \boldsymbol{\Gamma}_k) \tag{2}$$

여기서 $\mathbf{r}_k$ 와 Γ_k 는 k 번째 mixture에 대응되는 보상벡터와 추정된 원음성의 공분산 행렬이다.

2.2. 켑스트럼 보상

앞의 두 가정은 SPLICE 방식에서 잡음 음성에 대한 원음성의 최소평균자승 오차 (*Minimum Mean Squared Error*, MMSE) 추정을 간단하게 해준다. 잡음 음성이 주어졌을 때 MMSE로 추정한 원음성의 조건부 기대값은 다음과 같이 주어지며,

$$\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x} \mid \mathbf{y}] = \sum_{k} p(k \mid \mathbf{y}) E_k[\mathbf{x} \mid \mathbf{y}, k] \tag{3}$$

식 (2)에 의해서 다음과 같이 정리된다.

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_{k} p(k \mid \mathbf{y}) \mathbf{r}_k \tag{4}$$

즉, 원음성은 각각의 mixture에 대응되는 보상 벡터들의 가중 합에 의해 표현될 수 있다. 빠른 구현을 위해서 식 (4)의 $p(k \mid \mathbf{y})$ 는 다음과 같이 간략화 할 수 있다.

$$\hat{p}(k \mid \mathbf{y}) \cong \begin{cases} 1 & k = \arg\max_{k} p(k \mid \mathbf{y}) \\ 0 & otherwise \end{cases} \tag{5}$$

일반적으로 SPLICE 방식에서는 정적(static) 파라메터 뿐만 아니라 delta 파라메터에 대해서도 효과적인 보상을 해 주기 위하여 low-pass 필터를 이용하여 매 프레임마다 구해진 보상벡터를 smoothing해 준다[11].

2.3. SPLICE 훈련

SPLICE를 통한 잡음 보상을 위해서, 각 잡음 환경에 대한 잡음 음성의 GMM과 GMM에 대응하는 보상벡터가 필요하다. 잡음 음성의 켑스트럼 벡터의 분포 $p(\mathbf{y})$는 벡터 양자화를 통한 군집화(clustering)를 이용하여 평균 벡터와 공분산 행렬을 초기화하고, EM 알고리즘을 이용하여 훈련할 수 있다. 그리고 스테레오 데이터가 주어진다면 분포 $p(x|\mathbf{y},k)$에 대한 보상벡터 $\mathbf{r}_k$는 최대유사도척도(maximum likelihood criterion)에 의해서 다음과 같이 추정된다.

$$\mathbf{r}_k = \frac{\sum_n p(k|\mathbf{y}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(k|\mathbf{y}_n)} \tag{6}$$

여기서

$$p(k|\mathbf{y}_n) = \frac{p(\mathbf{y}_n|k)p(k)}{\sum_k p(\mathbf{y}_n|k)p(k)} \tag{7}$$

이다.

SPLICE 방식은 다음과 같은 두 단계로 구현된다. 첫 단계에서 잡음 음성의 매 프레임마다 식 (5)에 의해 최적 mixture를 찾고, 두 번째 단계에서 그 mixture에 대응하는 보상벡터를 가져와서 보상벡터를 smoothing한 후 잡음 음성의 특징 벡터에 더해준다.

2.4. 환경 모델 선택

SPLICE 방식은 특정 잡음환경에 대한 스테레오 데이터를 이용하여 잡음환경에 의해 발생하는 왜곡을 보상벡터를 통해 표현한다. 이는 SPLICE가 특정 잡음환경에 최적화됨을 의미하며, 만약 SPLICE의 훈련 환경과 테스트 환경이 다르다면 성능은 저하된다. 이러한 문제점은 다양한 잡음 환경에 대해서 SPLICE 시스템을 훈련하고, 이들 중에 테스트 환경과 가장 일치하는 환경모델을 실시간으로 선택함으로써 해결할 수 있다[8]. 이 방법은 매 프레임마다 입력 음성 $\mathbf{y}_n$에 대한 각 환경 e의 유사도(likelihood), $p(\mathbf{y}_n|e)$를 추정하고, 이 값을 최대화하는 환경 $\hat{e}$을 선택한다.

$$\hat{e} = \arg\max_e p(\mathbf{y}_n|e) \tag{8}$$

식 (8)의 경우 매 프레임 단위로 잡음 환경을 선택하게 되므로 시변(time-varying) 환경에도 대응할 수 있다는 장점이 있으나, 추정오류로 인해 선택된 잡음 환경이 프레임별로 불필요하게 fluctuation하는 문제를 일으킬 수 있다. 따라서 실제로는 프레임별로 추정된 $p(\mathbf{y}_n|e)$ 값을 그대로 사용하는 대신, 시간에 대해 smoothing한 값을 이용하여 환경모델을 선택함으로써 환경모델 선택의 신뢰도를 높인다[8]. 본 논문에서도 이와 동일한

방법으로 환경선택을 하였다.

3. SPLICE 방식 기반의 인식성능 향상을 위한 제안된 방법들

3.1. CMS를 전처리에 사용한 SPLICE 방식

기존의 SPLICE 논문에서는 Cepstral Mean Subtraction (CMS)에 의한 채널보상 과정을 SPLICE 처리 후에 수행한다[11]. 채널왜곡 특성이 시불변(time-invariant)이라면, 이 왜곡은 켑스트럼 영역에서 음성특징벡터에 더해지는 상수 형태의 바이어스(bias) 벡터로 표현되며, 전체 음성의 평균벡터를 차감하는 CMS 과정을 통해 효과적으로 제거될 수 있다. 그러나 SPLICE 처리 과정에서 프레임별로 잡음보상을 하면, 이때의 보상벡터가 완벽한 것이 아니기 때문에 결과적으로 프레임 별로 채널왜곡 특성이 달라지는 것과 동일한 상황이 발생한다. 이 경우 CMS에 의한 채널보상 모델의 가정이 맞지 않으므로, SPLICE 이후에 CMS를 적용하면 채널왜곡 보상 성능이 떨어짐을 예상할 수 있다.

본 논문에서는 CMS를 SPLICE 처리 이후에 사용하는 대신, SPLICE의 전처리 과정으로 사용하는 방안을 제안하며, 이 경우 채널왜곡 보상은 아무런 문제없이 수행될 수 있다 그런데, CMS 과정 역시 채널성분만 제거하는 것이 아니라 발화음성의 평균벡터도 함께 제거하며, 발화길이가 충분히 길지 않을 경우 발화 내용에 따라 음성평균이 다르기 때문으로 인한 왜곡을 초래한다. 그러므로 CMS를 SPLICE의 전처리로 사용할 경우 이 왜곡이 SPLICE의 성능에 부정적인 영향을 끼칠 수 있다. 이 문제를 해결하기 위하여 본 논문에서는 훈련용 스테레오 데이터 모두에 대해 CMS를 전처리로 사용한 다음 SPLICE 모델을 훈련시킴으로써, SPLICE 보상벡터에 CMS의 영향이 미리 반영되도록 하였다.

3.2. 음성학적 정보를 이용한 SPLICE

본 논문에서는 기존의 SPLICE에 음성학적 정보를 추가하여 잡음 처리에 적용함으로써 성능을 향상시키고자 하였다. SPLICE에서는 모든 음향공간에 대해서 오차가 최소화되도록 보상벡터를 구하지만, 인식 어휘에 따라 제한된 음향 공간을 가지는 상황에서는 이 방법이 최적이라고 볼 수 없다. 그 대신 음향 공간이 비슷한 프레임들을 묶어서 그 클러스터 내에서 오차가 최소가 되게 보상벡터를 훈련한다면 보다 정확한 보상벡터를 구할 수 있을 것이다. 이러한 클러스터를 구성하기 위해 본 논문에서는 음소 정보를 이용하였다. 이러한 음소 종속적인 보상 벡터는 각 잡음 환경이 각 음소에 미치는 정보를 구체적으로 표현하기 때문에 인식 성능의 향상을 기대할 수 있다. 음소에 의해 분할된 음향 공간에 대해 잡음 음성은 다음 식과 같이 나타낼 수 있다.

$$p(\mathbf{y}) = \sum_{s}\sum_{k} p(\mathbf{y} \mid k, s)p(k \mid s)p(s) \tag{9}$$

여기서 $p(\mathbf{y}|k,s)=N(\mathbf{y};\boldsymbol{\mu}_{k,s},\boldsymbol{\Sigma}_{k,s})$이고, s와 k는 각각 음소 인덱스와 mixture 인덱스이다 잡음 음성에 대한 원음성의 조건부 확률분포는 다음과 같고,

$$p(\mathbf{y}|k,s)=N(\mathbf{y};\boldsymbol{\mu}_{k,s},\boldsymbol{\Sigma}_{k,s}) \tag{10}$$

원음성의 MMSE 추정 값은 다음과 같이 나타낼 수 있다.

$$\hat{\mathbf{x}}_{MMSE}=\mathbf{y}+\sum_{k}\sum_{s}p(k,s|\mathbf{y})\mathbf{r}_{k,s} \tag{11}$$

이 때 보상벡터는 다음과 같이 훈련된다.

$$\mathbf{r}_{k,s}=\frac{\sum_{n}p(k,s|\mathbf{y}_n)(\mathbf{x}_n-\mathbf{y}_n)}{\sum_{n}p(k,s|\mathbf{y}_n)} \tag{12}$$

여기서 훈련 시에는 Viterbi 디코딩에 의한 강제정렬(forced alignment)를 통해 각 특징 벡터의 음소 정보를 알 수 있으므로 $p(s|\mathbf{y}_n)=\delta(s_n-s)$가 된다. 따라서

$$\begin{aligned}p(k,s|\mathbf{y}_n)&=p(k|\mathbf{y}_n,s)p(s|\mathbf{y}_n)\\&=p(k|\mathbf{y}_n,s)\delta(s_n-s)\end{aligned} \tag{13}$$

이 되고, 다음과 같이 보상벡터를 구할 수 있다.

$$\mathbf{r}_{k,s}=\frac{\sum_{n}p(k|\mathbf{y}_n,s)\delta(s_{\mathbf{y}_n}-s)(\mathbf{x}_n-\mathbf{y}_n)}{\sum_{n}p(k,s|\mathbf{y}_n)\delta(s_{\mathbf{y}_n}-s)} \tag{14}$$

인식 시에는 각 입력 프레임에 대한 음소 정보를 알 수 없으며, 따라서 기존의 SPLICE와 동일하게 적용한다.

3.3. 잔류 잡음 보상을 위한 2 단계 SPLICE 방식

SPLICE와 같은 특징벡터 영역 기반의 잡음 보상 방법은 기존의 음성인식 시스템의 구조 변화 없이 전처리 기법으로 구현이 가능하다는 장점이 있다. 그러나 잡음 섞인 음성에서 잡음을 제거하더라도 보상된 특징벡터에 잔류 왜곡(residual distortion)이 남게 되는 문제점이 있으며, 특히 이러한 잔류 왜곡은 SNR이 낮은 부분에 많이 발생하여 훈련 환경과 인식과정에서의 새로운 불일치를 초래한다. 본 논문에서는 잔류 왜곡의 영향을 최소화 하기 위해 SPLICE를 2 단계로 적용함으로써 SNR이 낮은 부분에서의 왜곡을 효과적으로 보상하는 방법을 제안한다.

이 방법에서는 그림 1과 같이 1차 잡음을 제거한 후에 잔류왜곡에 대해 훈련된 SPLICE를 2 차로 적용하여 왜곡을 보상하게 된다. 1차로 적용하는 SPLICE는 기존의 방식과 동일하게 잡음 음성에 대해서 GMM과 스테레오 데이터를 이용하여 구한 보상벡터로 구성된다. 2 차로 적용하는 SPLICE는 1 차 SPLICE를 통해 보상된 데이터를 또 다른 잡음 음성으로 가정하고, 이 데이터에 대한 GMM과 보상벡터를 훈련하여 구성한다. 그림에서 보상벡터 smoothing은 2.2 절 끝 부분에서 언급된 내용과 동일하다.

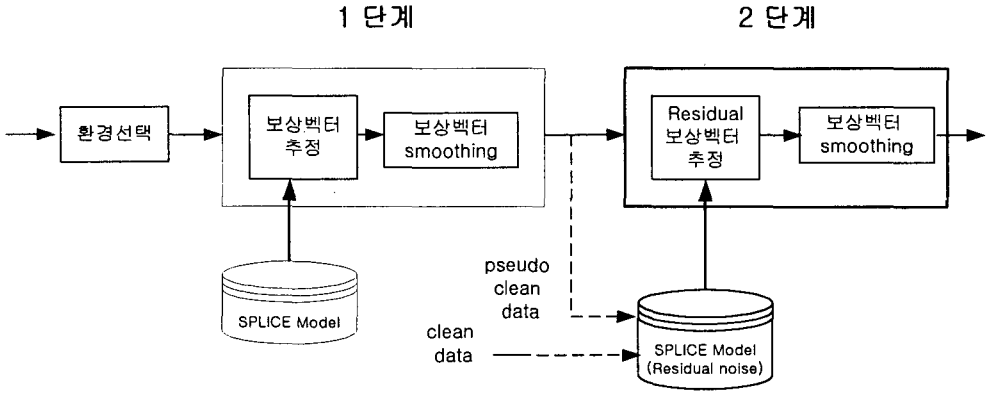

그림 1. 2 단계 SPLICE를 이용한 잔류 왜곡 보상

4. 실험 및 결과

4.1. 실험환경

Aurora 2 데이터베이스[1]는 한 자리에서 일곱 자리까지의 영어 연속 숫자로 구성된 TIDigit DB[9]에 실제 환경의 잡음을 인위적으로 더한 것이다. Aurora 2 DB는 훈련 데이터와 테스트 데이터로 구분되어 있으며 테스트 데이터는 채널특성은 동일하고 서로 다른 잡음이 더해진 두개의 subset(set A, set B)과 채널특성이 다른 subset으로 총 세 개의 subset으로 구성되어 있다. 각 subset들은 깨끗한 음성과 20 dB, 15 dB, 10 dB, 5 dB 0 dB, −5 dB의 SNR을 가지도록 잡음이 더해졌다. 훈련 모드는 clean 모드와 multi-condition 모드의 두 가지로 나누어져 있다. Clean 모드는 잡음이 더해지지 않은 깨끗한 발성으로만 HMM 모델을 훈련하는 것이고 multi-condition 모드는 clean 모드에서 사용한 훈련용 음성 데이터를 20 개의 subset으로 나누어서 clean과 20 dB, 15 dB, 10 dB, 5 dB의 5 가지의 SNR에 대해서 set A에 사용되었던 4 종류의 잡음을 각각 더한 데이터를 이용하여 HMM 모델을 훈련한 것이다.

Aurora baseline 시스템은 특징벡터 추출을 위해 WI007 front-end[10]를 사용하며 이것은 12 차 MFCC와 0 차 켑스트럼 그리고 로그 에너지를 추출해준다. Aurora

baseline 시스템은 12 차 MFCC와 로그 에너지 그리고 각각의 delta와 delta-delta 파라메터를 포함하여 총 39 차 특징벡터를 사용하였다. 인식 모델은 Aurora 평가를 위해 미리 정의되어 있는 HTK 스크립트를 사용하였으며, 이는 각 상태 당 3 개의 Gaussian 과 각 숫자 단어 당 16 개의 상태를 사용하고 묵음 모델은 3 개의 상태에 각 상태 당 6 개의 Gaussian을 사용하는 단어 모델이다[1]. 표 1은 Aurora 2 DB의 baseline 실험 결과이며 인식성능 평가는 0dB에서 20dB까지의 SNR에 대해서 수행된다.

표 1. Aurora WI007 전처리의 단어 인식률(%)

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	87.82	86.27	83.78	86.39
Clean Only	61.34	55.75	83.78	66.14
Average	74.58	71.01	83.78	76.27

4.2. 실험결과 및 검토

본 논문에서 사용한 파라메터는 13 차 멜-켑스트럼(MFCC)으로서 기존의 SPLICE [11]와 동일한 조건 하에서의 성능비교를 위해, WI007 front-end를 변형하여 크기(magnitude) 스펙트럼을 사용하는 대신 전력(power) 스펙트럼을 사용하여 추출하였으며 로그 에너지 대신에 0 차 켑스트럼을 사용하였다. 또한, 이러한 구성이 Aurora baseline 시스템의 구성보다 성능이 약간 우수함을 나타냈다. 잡음 음성은 256 개 Gaussian mixture로 모델링 하였으며, clean을 포함하여 multi-condition에서 사용되었던 잡음 종류와 SNR에 따라 총 17 개의 Gaussian mixture model이 구성되었다. 표 2는 Aurora 2 DB에 대해서 평가된 본 논문에서 구현한 SPLICE 실험 결과이다.

표 2. 본 논문에서 구현한 SPLICE의 요약된 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	90.11	86.39	83.63	87.82
Clean Only	84.95	82.68	77.39	82.53
Average	87.53	84.53	80.51	84.93

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	18.81%	0.86%	-0.92%	6.87%
Clean Only	61.06%	60.85%	33.21%	56.25%
Average	39.94%	30.85%	18.14%	31.56%

기존의 SPLICE 논문[11]의 결과와 비교해 본 연구에서 구현한 것이 성능이 조금 떨어짐을 알 수 있는데, 이는 잡음 음성에 대한 GMM 결과가 완전히 동일하지 않고 이로 인해 추정된 보상 벡터 값과 잡음환경 모델 선택의 차이 등의 요인으로 인식률의 차이가

발생한 것으로 추정된다.

다음은 본 논문에서 제안한 성능향상 방법에 대해서 실험한 결과이다. 표 3과 표 4는 보다 정확한 채널 보상을 위해 CMS를 SPLICE 전 단계에서 적용시킨 실험 결과이다. 전체 단어 인식률을 비교해 보았을 때, 기존의 SPLICE를 적용한 경우인 84.93%에서 88.41%로 성능이 향상되었음을 볼 수 있으며, 표 2와 비교해 볼 때 채널 환경이 동일한 set A나 set B보다 채널환경이 다른 set C에 대해서 인식 성능이 더욱 향상 되었음을 볼 수 있다. 이는 3.1 절에서 설명한 바와 같이 CMS를 SPLICE 전 단계에 사용하여 채널왜곡 보상 성능이 개선된 때문으로 풀이된다.

표 3. SPLICE를 전 단계에 사용한 SPLICE의 요약된 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.43	89.37	90.12	90.34
Clean Only	86.72	86.66	85.67	86.49
Average	89.07	88.02	87.89	88.41

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	29.67%	22.55%	39.11%	29.05%
Clean Only	65.64%	69.86%	57.67%	66.16%
Average	47.65%	46.21%	48.39%	47.60%

표 4. SPLICE를 전 단계에 사용한 SPLICE의 결과

Aurora 2 Multicondition Training - Results

	A					B					C					Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Improvement
Clean	98.74	98.97	98.78	98.98	98.87	98.74	98.97	98.78	98.98	98.87	98.77	98.91	98.84	98.86	22.45%
20dB	98.31	98.46	98.57	98.24	98.40	98.22	97.82	98.24	98.36	98.16	98.37	97.88	98.13	98.25	32.85%
15dB	97.36	97.85	98.15	97.66	97.76	97.39	97.31	97.14	97.32	97.29	97.30	97.04	97.17	97.45	30.26%
10dB	95.39	96.16	96.36	95.62	95.88	94.50	95.01	94.75	95.34	94.90	94.60	94.53	94.57	95.23	22.16%
5dB	90.64	89.27	92.42	90.84	90.79	87.29	87.85	88.28	87.60	87.76	90.54	88.33	89.44	89.31	25.47%
0dB	75.25	68.74	77.99	75.32	74.33	65.86	68.17	70.89	69.98	68.73	73.01	69.62	71.32	71.48	29.06%
-5dB	39.21	26.78	34.09	43.54	35.91	27.20	31.14	32.00	30.55	30.22	38.23	31.29	34.76	33.40	11.55%
Average	91.39	90.10	92.70	91.54	91.43	88.65	89.23	89.86	89.72	89.37	90.76	89.48	90.12	90.34	
	23.43%	17.81%	45.81%	29.29%	29.67%	22.32%	16.93%	17.99%	31.42%	22.55%	44.88%	32.95%	39.11%		29.05%

Aurora 2 Clean Training - Results

	A					B					C					Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Improvement
Clean	99.96	99.15	99.05	99.14	99.07	99.96	99.15	99.05	99.14	99.07	99.96	99.18	99.07	99.07	3.74%
20dB	97.82	98.37	98.72	98.21	98.28	98.34	97.73	98.33	98.61	98.25	98.04	97.58	97.81	98.18	61.68%
15dB	96.04	97.25	97.29	96.14	96.68	97.42	96.43	97.70	97.19	97.19	96.25	96.34	96.30	96.81	72.27%
10dB	92.35	94.17	94.12	92.63	93.32	93.95	92.41	94.69	94.01	93.77	92.26	92.11	92.19	93.27	78.09%
5dB	84.83	85.01	87.35	83.99	85.30	84.62	83.37	86.37	83.49	84.46	84.49	82.35	83.42	84.59	74.12%
0dB	60.21	56.65	61.68	61.49	60.01	60.33	56.08	62.18	60.04	59.66	59.81	57.44	58.63	59.59	51.02%
-5dB	24.04	18.62	20.34	27.09	22.52	23.64	19.47	25.05	21.85	22.50	22.84	19.86	21.35	22.28	14.95%
Average	86.25	86.29	87.83	86.49	86.72	86.93	85.20	87.85	86.67	86.66	86.17	85.16	85.67	86.49	
	54.94%	72.64%	69.12%	60.97%	65.64%	72.43%	61.55%	74.02%	69.95%	69.86%	59.13%	56.21%	57.67%		66.16%

다음으로 음성학적 정보를 이용한 SPLICE 방식에 대한 실험을 수행하였다. 이 실험에서는 먼저 잡음이 섞이지 않은 clean training 음성 데이터를 이용하여 monophone 모델을 만든 후, 이것을 이용하여 강제 정렬(forced alignment)을 통해 음소 레이블 정보를 구하였다. 영어 숫자 zero에서 nine까지 그리고 oh를 포함하여 11 개의 숫자를 구성하는 21 개의 음소에, 묵음을 포함하여 22 개의 음소를 사용하였다. 기존의 256 개의 mixture개수와 비슷한 환경에서 실험하기 위해 각 음소에 대해서 12 개의 mixture로 구성된 GMM을 사용하였다. 각 음소에 대한 VQ로 GMM을 초기화하고 EM 알고리즘을 통해 음소모델을 생성하였다. 이렇게 구성된 음향모델을 이용하여 각 음소에 대한 보상벡터를 구하였다.

표 5와 표 6은 CMS를 전처리로 사용하고 음성학적 정보를 이용한 SPLICE 모델을 적용하여 실험한 결과이다. 전체 단어 인식률이 CMS만 전처리로 적용한 SPLICE 결과인 88.41%에서 89.39%로 향상되었으며, Aurora 2 baseline 시스템에 대한 상대적인 인식 향상률은 47.60%에서 50.01%로 향상되었음을 볼 수 있다. 이는 음향공간을 일반적인 GMM으로 표현하는 것보다, 음성학적 정보를 고려하여 구성할 경우 SPLICE 보상을 통해 음성학적 변별력이 향상되기 때문으로 풀이된다.

본 논문에서는 SPLICE 처리 후에 CMS를 적용하는 기존의 SPLICE 방식에도 음성학적 정보를 이용한 SPLICE 모델을 적용해 보았다. 그 결과 Aurora 2 baseline 시스템에 대한 인식 성능 향상률이 31.56%에서 34.15%로 역시 개선되었으며, 이를 통해 음성학적 정보 사용의 일관적인 유용성을 확인할 수 있었다.

표 5. 음성학적 정보를 이용한 SPLICE의 요약된 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.41	89.17	90.47	90.39
Clean Only	89.02	88.20	87.85	88.46
Average	90.21	88.68	89.16	89.39

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	29.51%	21.12%	41.28%	28.93%
Clean Only	71.59%	73.32%	64.12%	71.09%
Average	50.55%	47.22%	52.70%	50.01%

표 6. 음성학적 정보를 이용한 SPLICE의 결과

Aurora 2 Multicondition Training - Results

	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M.	Street M.	Average	Overall	Improvement
Clean	98.93	99.03	98.96	98.95	98.97	98.93	99.03	98.96	98.95	98.97	98.96	99.06	99.01	98.98	30.30%
20 dB	98.10	98.46	98.54	97.75	98.21	97.88	97.82	98.24	98.40	98.09	98.31	97.97	98.14	98.15	29.02%
15 dB	97.48	97.52	98.00	97.22	97.56	97.08	97.34	96.90	96.88	97.05	97.30	96.70	97.00	97.24	24.60%
10 dB	95.09	95.74	96.84	94.91	95.65	93.77	94.68	94.90	95.16	94.63	94.87	94.65	94.76	95.06	19.69%
5 dB	90.70	89.33	93.41	90.87	91.08	85.51	87.39	87.98	88.65	87.38	90.57	87.36	88.97	89.18	24.88%
0 dB	75.68	66.69	80.41	75.47	74.56	62.36	69.17	71.55	71.74	68.71	76.24	70.77	73.51	72.01	30.06%
-5dB	43.69	28.51	40.47	45.54	39.55	25.45	36.06	36.74	35.45	33.43	43.81	36.94	40.38	37.27	16.63%
Average	91.41	89.55	93.44	91.24	91.41	87.32	89.28	89.91	90.17	89.17	91.46	89.49	90.47	90.33	
	23.60%	13.26%	51.32%	26.85%	29.51%	13.20%	17.30%	18.38%	34.40%	21.12%	49.02%	33.01%	41.28%		28.93%

Aurora 2 Clean Training - Results

	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M.	Street M.	Average	Overall	Improvement
Clean	98.96	99.15	99.08	99.17	99.09	98.96	99.15	99.08	99.17	99.09	98.96	99.12	99.04	99.08	4.48%
20 dB	97.64	98.52	98.60	97.16	97.98	97.88	97.76	98.27	98.64	98.14	97.91	97.73	97.82	98.01	57.57%
15 dB	96.65	97.34	97.85	95.65	96.87	97.27	96.28	97.23	97.38	97.04	96.53	95.92	96.23	96.81	72.74%
10 dB	92.75	95.34	95.50	91.89	93.87	93.77	92.50	94.69	95.06	94.01	92.82	91.87	92.35	93.62	78.99%
5 dB	87.38	87.91	90.87	87.01	88.29	84.99	84.49	87.62	87.60	86.18	87.35	84.73	86.04	87.00	78.18%
0 dB	69.14	64.39	70.92	67.85	68.08	63.28	63.45	68.95	66.80	65.62	69.39	64.27	66.83	66.84	59.89%
-5dB	34.57	26.81	28.87	36.35	31.65	29.20	28.87	32.48	29.56	30.03	34.66	29.05	31.86	31.04	24.58%
Average	88.71	88.70	90.75	87.91	89.02	87.44	86.90	89.35	89.10	88.20	88.80	86.90	87.85	88.46	
	63.01%	77.45%	76.52%	65.07%	71.59%	73.50%	65.95%	77.22%	75.42%	73.32%	66.90%	61.35%	64.12%		71.09%

다음으로 잔류왜곡 보상을 위한 2 단계 SPLICE 적용 실험을 수행하였다. 잔류 왜곡을 모델링하기 위한 SPLICE에서는 각 잡음 환경에 대해서 256 개의 mixture를 사용하여 적용하였으며, 이 방식에서는 2 단계의 SPLICE를 적용한 뒤에 CMS를 적용하였다. 실험 결과가 표 7과 표 8에 정리되어 있다. 표에서 보는 바와 같이 기존의 SPLICE와 비교할 때, 전체 인식률은 84.93%에서 87.92%로 향상되었으며, 인식 향상률은 31.56%에서 45.33%로 높아졌다.

표 7. 잔류왜곡 보상을 위한 2 단계 SPLICE의 요약된 결과

Absolute performance

Training Mode	Set A	Set B	Set C	Overall
Multicondition	90.81	88.94	90.09	89.92
Clean Only	85.83	86.33	85.26	85.92
Average	88.32	87.64	87.67	87.92

Performance relative to Mel-cepstrum

Training Mode	Set A	Set B	Set C	Overall
Multicondition	24.58%	19.45%	38.92%	27.62%
Clean Only	63.36%	69.11%	56.45%	63.14%
Average	43.97%	44.28%	47.69%	45.33%

표 8. 잔류왜곡 보상을 위한 2 단계 SPLICE의 결과

Aurora 2 Multicondition Training - Results

	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Improvemen
Clean	98.83	98.88	98.72	98.95	98.85	98.83	98.88	98.72	98.95	98.85	98.93	98.91	98.92	98.86	22.52%
20 dB	98.19	98.43	98.48	98.09	98.30	97.97	97.94	98.12	98.46	98.12	97.97	97.76	97.87	98.14	28.92%
15 dB	97.08	97.70	97.91	97.28	97.49	97.48	97.43	97.17	97.04	97.28	97.39	96.98	97.19	97.35	26.90%
10 dB	94.78	95.89	95.91	95.22	95.45	94.41	94.92	94.48	95.00	94.70	94.65	95.10	94.88	95.04	18.45%
5 dB	90.48	88.57	92.13	89.26	90.11	86.25	87.18	87.41	87.23	87.02	90.11	87.18	88.65	88.58	20.40%
0 dB	73.96	67.17	76.35	73.31	72.70	63.00	68.89	70.30	68.13	67.58	73.93	69.83	71.88	70.49	26.46%
-5dB	36.78	25.70	34.21	41.35	34.51	27.54	31.08	32.00	29.90	30.13	37.80	31.29	34.55	32.77	10.70%
Average	90.90	89.55	92.16	90.63	90.81	87.82	89.27	89.50	89.17	88.94	90.81	89.37	90.09	89.92	
	19.09%	13.29%	41.79%	21.74%	24.58%	16.63%	17.23%	15.00%	27.77%	19.45%	45.17%	32.25%	38.92%		25.93%

Aurora 2 Clean Training - Results

	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Improvemen
Clean	98.93	99.15	99.05	99.17	99.09	98.93	99.15	99.05	99.17	99.08	98.96	99.15	99.06	99.07	3.64%
20 dB	97.79	98.31	98.54	97.38	98.01	98.22	97.49	98.45	98.49	98.16	97.70	97.70	97.70	98.01	57.47%
15 dB	95.76	97.13	97.20	95.43	96.38	97.64	96.64	97.38	97.22	97.22	96.13	96.25	96.19	96.68	70.79%
10 dB	91.46	93.71	94.18	91.98	92.83	94.29	92.93	94.63	93.43	93.82	91.83	92.38	92.11	93.08	77.34%
5 dB	84.07	83.01	85.42	82.57	83.77	84.00	82.36	85.77	83.92	84.02	83.67	82.35	83.01	83.72	72.70%
0 dB	58.83	54.72	58.69	60.51	58.19	58.00	56.47	61.65	57.61	58.43	58.83	55.71	57.27	58.10	49.24%
-5dB	22.51	18.11	20.85	26.44	21.98	22.60	19.14	26.10	22.49	22.58	23.18	20.13	21.66	22.16	14.81%
Average	85.58	85.38	86.81	85.57	85.83	86.43	85.18	87.58	86.13	86.33	85.63	84.88	85.26	85.92	
	52.75%	70.82%	66.51%	58.31%	63.36%	71.37%	61.49%	73.42%	68.75%	69.11%	57.54%	55.37%	56.45%		64.74%

그림 2는 2 단계 SPLICE를 통해 잡음을 보상한 결과를 SNR에 따라 살펴 본 것이다. 예상과 같이 SNR이 낮은 10 dB, 5 dB, 0 dB와 −5 dB에서 성능이 많이 향상되었으며, 불일치가 큰 clean-condition에서 이 방법이 더욱 효과적임을 알 수 있다.

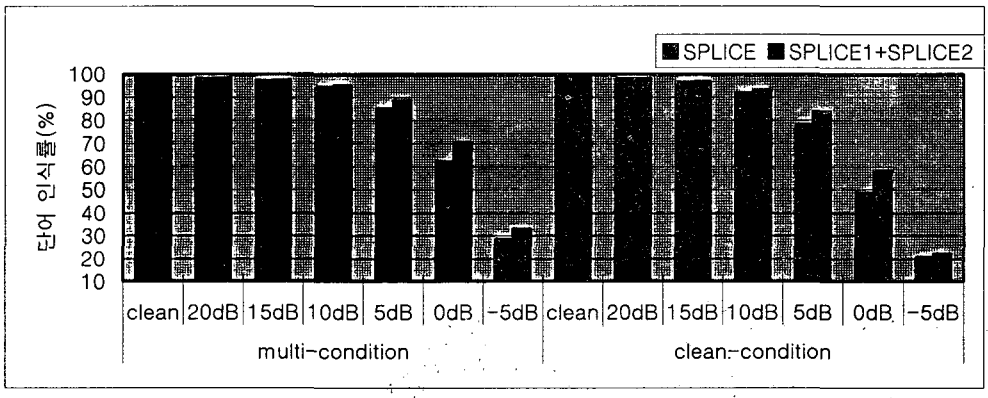

그림 2. SNR에 따른 2 단계 SPLICE를 적용 결과 비교 (단어 인식률 (%))

표 9는 지금까지 수행한 실험 결과들을 정리 요약한 것이다. 표로부터 음성학적 정보를 사용한 SPLICE 방식이 성능 면에서 가장 우수함을 확인할 수 있으며, 그 다음으로 CMS를 사전에 적용한 SPLICE 방식, 2 단계 SPLICE 방식, 기존의 SPLICE 방식, 그리

고 baseline의 순서였다. 2 단계 SPLICE 방식의 경우 기존의 SPLICE 방식보다는 성능이 우수하지만 CMS를 사전에 적용한 SPLICE 방식에 비해서는 미세하기는 하지만 조금 뒤떨어지는 성능에 머물렀다. 이는 2 단계 SPLICE 방식 적용 시, 기존의 SPLICE 방식과 마찬가지로 CMS를 SPLICE 이후에 적용했음을 고려하더라도 사실상 기대에는 미치지 못한 결과이며, 이에 대한 추가분석 및 성능개선 방안을 강구할 필요가 있다고 판단된다.

표 9. SNR에 대한 각 알고리즘의 성능 비교(단어 인식률 (%))

	Clean Only Training							Multicondition Training						
	clean	20dB	15dB	10dB	5dB	0dB	−5dB	Clean	20dB	15dB	10dB	5dB	0dB	−5dB
Baseline	99.03	94.07	85.04	65.52	38.61	17.09	8.53	98.52	97.35	96.29	93.79	85.52	59.00	24.50
기존의 SPLICE	99.07	97.88	96.38	91.70	78.15	48.53	20.67	98.86	98.03	97.13	94.31	85.09	62.06	28.47
CMS 적용 후 SPLICE	99.07	98.18	96.81	93.27	84.59	59.59	22.28	98.86	98.25	97.45	95.23	89.31	71.48	33.40
Phonetic SPLICE	99.08	98.01	96.81	93.62	87.00	66.84	31.04	98.98	98.15	97.24	95.06	89.18	72.01	37.27
2단계 SPLICE	99.07	98.01	96.68	93.08	83.72	58.10	22.16	98.86	98.14	97.35	95.04	88.58	70.49	32.77

6. 결 론

본 논문에서는 잡음 환경 음성인식을 위한 전처리 방법 중 효과적이라고 알려진 SPLICE 방식을 기반으로 하여 몇 가지 성능향상 방법을 제시하였다. 먼저 채널 왜곡 등에 의해서 발생할 수 있는 켑스트럼 영역에서의 바이어스 성분을 CMS를 적용하여 사전에 제거한 후에 SPLICE를 적용하는 방법과, 특징벡터 영역 기반의 잡음 보상 과정에서 발생하는 잔류 왜곡을 보상해 주기 위해서 2 단계 SPLICE를 적용하는 방법을 제안하였다. 이들은 기존의 SPLICE가 Aurora 2 baseline 시스템[12]에 대해서 31.56%의 단어 인식 향상률을 나타낸 것에 반해, 각각 단어 인식 향상률을 47.60%와 45.33%로 향상시켰다. 그리고 음성학적 정보를 이용하여 잡음 음성 모델과 보상 벡터를 훈련하고 CMS를 전처리로 사용함으로써 최고 50.01%의 인식 향상률을 얻을 수 있었다.

기존의 SPLICE 연구에서는 잡음에 대해서 정규화된 켑스트럼을 이용하여 SPLICE 모델을 구성함으로써 SPLICE 훈련과정에서 나타나지 않은 잡음 환경에 대해 강인성을 향상시키는 잡음평균정규화(noise mean normalization, NMN)라는 부가과정을 통해 인식 성능을 추가적으로 향상시켰다[11]. 앞으로 본 논문에서 제안한 방식과 NMN을 통합하여 성능을 더 개선시키는 방안 등에 대해 연구가 계속될 예정이다.

참 고 문 헌

[1] Hirsch, H. G. & D. Pearce. 2000. "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.

[2] Hermansky, H. & N. Morgan. 1994. "RASTA processing of speech." *IEEE Trans. on Speech and Audio Processing*, 2(4), 578-589.

[3] Boll, S. 1979. "Suppression of acoustic noise in speech using spectral subtraction." *IEEE Trans. ASSP*, 27(2), 113-120.

[4] Atal, B. 1974. "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification." *Journal of the Acoustical Society of America*, 5(6), 1304-1312.

[5] Moreno, P. J., B. Raj & R. M. Stern. 1996. "A vector Traylor series approach for environment-independent speech recognition." *Proc. ICASSP*, 733-736.

[6] Gales, M. & S. J. Young. 1993. "HMM recognition in noise using parallel model combination." *Proc. EUROSPEECH*, 837-840.

[7] Deng, L., A. Acero, M. Plumpe & X. Haung. 2000. "Large vocabulary continuous speech recognition under adverse conditions." *Proc. ICSLP*, 806-809.

[8] Droppo, J., A. Acero & L. Deng. 2001. "Efficient on-line acoustic environment estimation for FCDCN in a continuous speech recognition system." *Proc. ICASSP*, 209-212.

[9] Leonard, R. G. 1984. "A database for speaker independent digit recognition." *Proc. ICASSP*, 3, 42-45.

[10] ETSI standard document. 2001. "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; front-end feature extraction algorithm; compressing algorithm." ETSI ES 201 108 v1.1.1.

[11] Droppo, J., L. Deng & A. Acero. "Evaluation of the SPLICE algorithm on the Aurora 2 database." *Proc. Eurospeech*, 217-220.

접수일자: 2003. 7. 28.
게재결정: 2003. 9. 8.

▲ 김형순
부산시 금정구 장전동 산 30번지 (우: 609-735)
부산대학교 전자공학과
Tel: +82-51-510-2452
Fax: +82-51-515-5190
E-mail: kimhs@pusan.ac.kr

▲ 김두희
부산시 금정구 장전동 산 30번지 (우: 609-735)
부산대학교 전자공학과 음성통신실험실
Tel: +82-51-510-1704 Fax: +82-51-515-5190
E-mail: gurmy@pusan.ac.kr