음성학적인 정보를 포함한 SPLICE를 이용한 잡음환경에서의 음성인식

김두희, 김형순
부산대학교 전자공학과

Speech Recognition in Noise Environments Using SPLICE with Phonetic Information

Doo Hee Kim, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
E-mail : {gurmy, kimhs}@pusan.ac.kr

요 약

훈련과정과 인식과정에서의 주변환경 잡음과 채널 특성 등의 불일치는 음성인식 성능을 급격히 저하시킨다. 이러한 불일치를 보상하기 위해서 켑스트럼 영역에서의 다양한 전처리 방법이 시도되고 있으며 최근에는 stereo 데이터와 잡음 음성의 Gaussian Mixture Model (GMM)을 이용해 보상벡터를 구하는 SPLICE 방법이 좋은 결과를 보이고 있다[1]. 기존의 SPLICE가 전체 발성에 대해서 음향학적인 정보만으로 Gaussian 모델을 구하는 반면 본 논문에서는 발성에 해당하는 음소정보를 고려하여 전체 음향 공간을 각 음소에 대해 나누어서 모델링하고 각 음소에 대한 Gaussian 모델과 그 음소에 해당하는 음성데이터만을 이용하여 음소별 보상벡터가 훈련되도록 하였다. 이 경우 보상벡터는 잡음이 각 음소에 미치는 영향을 보다 자세히 나타내게 된다. Aurora 2 데이터베이스를 이용한 실험결과, 제안된 방법이 기존의 SPLICE 방법에 비해 성능향상을 보였다.

1. 서 론

음성인식에서 훈련환경과 테스트 환경이 다르면 인식 성능은 급격히 저하된다. 이러한 불일치의 원인으로는 서로 다른 채널특성, 화자의 차이, 그리고 주변잡음의 영향 등을 들 수 있다. 이러한 문제의 해결책으로 잡음환경에 강인한 음성인식을 위한 다양한 접근 방법이 제안되어 왔다. 이러한 접근 방식은 특징벡터 영역에서의 음질 개선(speech enhancement)방법과 모델 적응(model adaptation)방법의 두 가지로 크게 나눌 수 있다.

첫번째 음질개선 방법은 잡음 음성 특징벡터로부터 관찰된 왜곡을 추정하여 이를 제거하려는 방법이다. 여기에는 RASTA, 스펙트럼 차감법 그리고 Cepstrum Mean Normalization(CMN) 방법 등이 있으며 이들 방법은 기존의 음성인식 시스템의 구조 변화 없이 전처리 기법으로 구현이 가능하다는 장점이 있다. 두 번째 방법인 모델 적응방법에는 Vector Taylor Series(VTS)와 Parallel Model Combination(PMC) 방법 등이 있다. 모델 적응방법은 정적/비정적 잡음을 다양하게 다룰 수 있고, 부가잡음 뿐만 아니라 채널왜곡도 동시에 처리할 수 있지만, 계산량이 많이 소요되는 문제점이 있다.

최근 음질개선 방법 중 하나인 Stereo-based Piecewise LInear Compensation for Environments (SPLICE) 방식이 제안되어 우수한 성능을 보여주고 있다[1]. 본 논문에서는 SPLICE 에서의 잡음 음성 모델링과 보상벡터 훈련에 음성학적 정보를 추가하여 잡음을 보상하는 방법을 제안하고 실험을 통해 기존의 SPLICE 에 비해 성능이 향상됨을 확인하였다.

본 논문의 구성은 다음과 같다. 2 절에서 기존의 SPLICE 방식에 대해 살펴보고, 3 절에서는 본 논문에서 제안한 음성학적인 정보를 포함한 SPLICE 방식에 대해 설명한다. 4 절에서는 실험 환경 및 결과에 대해서 언급하고, 마지막으로 5 절에서 결론을 맺는다.

2. SPLICE 방식

2.1 음성 모델과 왜곡

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

첫번째 가정은 잡음 음성의 켑스트럼 벡터 분포가 M

개의 Gaussian mixture 로 모델링된다는 것이다.

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

여기서

$$p(\mathbf{y} \mid k) = N(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

이고 $p(k)$, $\boldsymbol{\mu}_k$ 및 $\boldsymbol{\Sigma}_k$ 는 각각 k 번째 Gaussian mixture 의 사전 확률, 평균벡터 그리고 공분산 행렬이다. 이러한 Gaussian mixture 모델은 각각의 잡음환경에 대해서 훈련된다.

두 번째 가정은 잡음 음성 $\mathbf{y}$ 이 주어졌을 때 원음성 $\mathbf{x}$ 의 평균벡터는 잡음 음성의 평균벡터와 선형 변환의 관계를 가진다는 것이다. 이때 선형 행렬을 단위행렬로 가정하면 원음성의 잡음음성에 대한 조건부 확률 분포는 다음과 같은 형태로 표현될 수 있다.

$$p(\mathbf{x} \mid \mathbf{y}, k) = N(\mathbf{x}; \mathbf{y} + \mathbf{r}_k, \boldsymbol{\Gamma}_k) \tag{3}$$

여기서 $\mathbf{r}_k$ 와 $\boldsymbol{\Gamma}_k$ 는 각 mixture k 에 의존하는 보상벡터와 추정된 원음성의 공분산 행렬이다.

2.2 켑스트럼 보상

앞의 두 가정은 SPLICE 방식에서 잡음음성에 대한 원음성의 Minimum Mean Squared Error(MMSE) 추정을 간단하게 해준다. 잡음 음성이 주어졌을 때 MMSE 로 추정한 원음성의 조건부 기대값은

$$\hat{\mathbf{x}}_{MMSE} = E[\mathbf{x} \mid \mathbf{y}] = \sum_k p(k \mid \mathbf{y}) E_{\mathbf{x}}[\mathbf{x} \mid \mathbf{y}, k] \tag{4}$$

와 같이 주어지며, 식(2)에 의해서 다음과 같이 정리된다.

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_k p(k \mid \mathbf{y}) \mathbf{r}_k \tag{5}$$

즉, 원음성은 각각의 mixture 에 관계된 보상벡터들의 가중 합에 의해 표현될 수 있다. 빠른 구현을 위해서 식(4)의 $p(k \mid \mathbf{y})$ 는 다음과 같이 간략화 할 수 있다. 이는 MMSE 추정 방식이 MAP 추정방식으로 접근함을 보여준다.

$$\hat{p}(k \mid \mathbf{y}) \cong \begin{cases} 1 & k = \arg\max_k p(k \mid \mathbf{y}) \\ 0 & otherwise \end{cases} \tag{6}$$

$$E_{\mathbf{x}}[\mathbf{x} \mid \mathbf{y}, k] = \mathbf{y} + \mathbf{r}_k \tag{7}$$

2.3 SPLICE 훈련

잡음 음성의 켑스트럼 벡터의 분포 $p(\mathbf{y})$ 는 Gaussian mixture 를 따른다고 가정하였으므로 EM 알고리즘을 이용하여 $\boldsymbol{\mu}_k$ 와 $\boldsymbol{\Sigma}_k$ 를 훈련할 수 있고 초기 파라메터는 VQ clustering 을 이용하여 구할 수 있다. 그리고 분포 $p(\mathbf{x} \mid \mathbf{y}, k)$ 에 대한 보상벡터 $\mathbf{r}_k$ 는 stereo 데이터가 주어진다면, maximum likelihood criterion 에 의해서 다음과 같이 추정된다.

$$\mathbf{r}_k = \frac{\sum_n p(k \mid \mathbf{y}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(k \mid \mathbf{y}_n)} \tag{8}$$

여기서

$$p(k \mid \mathbf{y}_n) = \frac{p(\mathbf{y}_n \mid k) p(k)}{\sum_n p(\mathbf{y}_n \mid k) p(k)} \tag{9}$$

이다.

SPLICE 는 다음의 두 단계로 적용된다. 첫 단계에서 잡음 음성의 매 프레임에 대해서 식(5)에 의해 최적 mixture 를 찾는다. 그리고 다음 단계에서 그 mixture 에 대응하는 보상벡터를 잡음 음성의 특징 벡터에 더해주면 된다.

2.4 환경 모델 선택

SPLICE 방식은 특정 잡음환경에 대한 stereo 데이터를 이용하여 이러한 잡음환경에 의해 발생하는 왜곡을 보상벡터를 통해 직접적으로 표현한다. 이는 SPLICE 가 특정 잡음 환경에 최적화 되어있음을 의미하며, 만약 SPLICE 의 훈련 환경과 테스트 환경이 다르다면 성능은 저하된다. 이러한 문제점은 다양한 잡음환경에 대해서 SPLICE 시스템을 훈련하고, 이들 중에 테스트 환경과 가장 일치하는 환경 모델을 실시간으로 선택함으로써 해결할 수 있다[2]. 이 방법은 매 프레임마다 입력 음성 $\mathbf{y}_n$ 에 대한 각 환경 e 의 likelihood $p(\mathbf{y}_n \mid e)$ 를 추정하고, 이 값을 시간에 대해 smoothing 한 값을 이용하여 환경모델을 선택하는 것이며, 본 논문에서도 동일한 방법으로 환경모델을 선택하였다.

3. 음성학적 정보를 이용한 SPLICE 방식

본 논문에서는 기존의 SPLICE 에 음성학적 정보를 추가하여 잡음 처리에 이용하였다. 실제 모든 음향공간에 대해서 에러가 최소화되도록 구해진 보상벡터는 실제 각 인식 단어에 따라 제한된 음향 공간을 가지는 현실에서 정확하다고 볼 수 없다. 실제 각 어휘에 대해서 음향 공간이 비슷한 것을 묶어서 그 cluster 내에서 에러가 최소가 되게 보상벡터를 훈련하게 된다면, 보다 정확한 보상 벡터를 구할 수 있을 것이다. 본 논문에서는 이러한 cluster 를 구성하기 위해 음소정보를 이용하였다. 이러한 음소 종속적인 보상벡터는 각 잡음 환경이 각 음소에 미치는 정보를 구체적 표현하기 때문에 인식 성능의 향상을 가대할 수 있다.

음소에 의해 분할된 음향 공간에 대해 잡음 음성은 다음과 같이 나타낼 수 있다.

$$p(\mathbf{y}) = \sum_s \sum_k p(\mathbf{y} \mid k,s) p(k \mid s) p(s) \quad (11)$$

여기서

$$p(\mathbf{y} \mid k,s) = N(\mathbf{y}; \boldsymbol{\mu}_{k,s}, \boldsymbol{\Sigma}_{k,s}) \quad (12)$$

이고, s 는 음소 index 이다.

잡음 음성에 대한 원음성의 조건부 확률 분포는

$$p(\mathbf{x} \mid \mathbf{y},k,s) = N(x; y + \mathbf{r}_{k,s}, \boldsymbol{\Gamma}_{k,s}) \quad (13)$$

와 같고, 원음성의 MMSE 추정 값은 다음과 같이 나타낼 수 있다.

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_k \sum_s p(k,s \mid \mathbf{y}) \mathbf{r}_{k,s} \quad (14)$$

이때 보상벡터 $\mathbf{r}_{k,s}$ 는 다음과 같이 훈련된다.

$$\mathbf{r}_{k,s} = \frac{\sum_n p(k,s \mid \mathbf{y}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(k,s \mid \mathbf{y}_n)} \quad (15)$$

여기서 훈련 시에는 각 특징벡터의 음소 정보를 알 수 있으므로 $p(s \mid y_n) = \delta(s_{y_n} - s)$ 가 된다. 따라서

$$\begin{aligned} p(k,s \mid \mathbf{y}_n) &= p(k \mid \mathbf{y}_n, s) p(s \mid \mathbf{y}_n) \\ &= p(k \mid \mathbf{y}_n, s) \delta(s_{y_n} - s) \end{aligned} \quad (16)$$

이 되고, 다음과 같이 보상벡터를 구할 수 있다.

$$\mathbf{r}_{k,s} = \frac{\sum_n p(k \mid \mathbf{y}_n, s) \delta(s_{y_n} - s)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(k \mid \mathbf{y}_n, s) \delta(s_{y_n} - s)} \quad (17)$$

테스트 시에는 각 입력 프레임에 대한 음소정보를 알 수 없으므로 기존의 SPLICE 와 동일하게 적용한다.

4. 실험 및 결과

4.1 Baseline 시스템과 DB

제안된 방식의 평가를 위해 Aurora 2 데이터베이스[3]가 사용되었다. Aurora 2 데이터베이스는 1 자리에서 7 자리까지의 영어 연결숫자로 구성된 TI Digit 에 다양한 잡음을 인공적으로 부가한 것이다. 음향 모델은 2 가지 모드로 훈련되는데, 8440 개의 clean utterance 로 훈련된 clean-condition 과 동일한 발성을 20 개의 잡음환경에 나누어 각 422 개의 noisy utterance 로 구성된 잡음 발성으로 훈련된 multi-condition 이 있다. 20 개의 잡음환경은 4 가지의 잡음종류(subway, babble, car, exhibition)와 각각의 5 단계의 잡음 레벨(clean, 20dB, 15dB, 10dB, 5dB)로 구성되어 있다. 테스트 DB 는 훈련에 나타난 4 가지 잡음종류(set A)에 새로운 4 가지 잡음 종류(set B)와 훈련과 다른 채널특성을 가지고 set A 와 set B 에 나타난 2 가지 잡음(set C)에 대해서 총 10 종류의 잡음으로 -5dB 에서 clean 까지의 7 가지의 잡음 레벨로 구성되어있다. 성능평가는 각 잡음 종류에 대해서 20dB 에서 0dB 까지의 잡음 레벨에 대해 수행된다.

Aurora baseline 시스템은 12 차 MFCC 와 log energy 를 포함하여 각각의 delta 와 delta-delta 파라메터를 포함한 총 39 차 MFCC 가 사용되었다. 특징벡터 추출은 W1007 front-end[4]를 이용한 것이다. 표 1 은 Aurora 2 데이터베이스의 baseline 실험 결과이다.

표 1. Aurora 2 baseline 실험결과

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	66.14	60.06
Average	74.58	71.01	74.96	73.23

표 2 는 기존의 SPLICE[5]를 적용하여 구현한 것이다. SPLICE 를 적용한 실험에서는 magnitude spectrum 대신에 power spectrum 을 사용하여 특징벡터를 추출하였으며, log-energy 대신 c0 를 사용하였다[5]. multi-condition 에 나타난 17 개의 잡음환경에 대해서 환경모델을 생성하였으며, 각 잡음환경에 대해 256 개의 mixtures 를 사용하여 잡음 음성 모델을 구하였다. 채널을 보상해주기 위해서 SPLICE 훈련과 테스트 과정에 모두 CMN 을 적용하였다. 보상벡터의 smoothing 은 하지 않았다.

표 2. 기존의 SPLICE 를 적용한 실험결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	90.77	88.25	89.37	89.48
Clean Only	85.90	85.53	84.90	85.55
Average	88.34	86.89	87.14	87.52
Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	24.22%	14.40%	34.45%	22.70%
Clean Only	63.54%	67.26%	55.42%	63.82%
Average	43.88%	40.65%	44.03%	43.26%

기존의 SPLICE 논문[5]에 제시된 전체 인식률은 86.38%로 본 논문의 baseline 으로 사용되는 87.52%보다는 약간 뒤떨어진다. 이는 환경선택 과정에서의 smoothing filter 계수 선정 방법 등의 차이 때문인 것으로 추정되며, 추가적인 검토가 필요하다.

4.2 음성학적 정보를 이용한 SPLICE 적용실험

음소 정보는 잡음이 섞이지 않은 음성 데이터를 이용하여 monophone 모델을 생성하여 Viterbi decoding 을 통해 구하였다. 영어 숫자 zero 에서 nine 까지 oh 를 포함하여 11 개의 숫자를 구성하는 21 개의 monophone 에,

묵음을 포함하여 22 개의 monophone 을 정의하였다. 기존의 256 개의 mixture 개수와 비슷한 환경에서 실험하기 위해 각 음소에 대해서 12 개의 mixture 로 구성된 Gaussian 모델을 사용하였다. 각 음소이 대한 VQ 로 GMM 을 초기화하고 EM 알고리즘을 통해 음소모델을 생성하였다. 이렇게 구성된 음향모델을 이용하여 각 음소에 대한 보상벡터를 구하였다. 표 3 은 각 set 에 대해서 요약된 결과이다. 전체 인식률은 87.52%에서 88.55%로 조금 향상되었으며, 상대적인 인식 향상률은 43.26%에서 46.64%로 향상되었음을 볼 수 있다. 그림 1 은 개별적인 SNR 에 대해서 인식 향상률을 비교해놓은 것이다. 잡음 레벨이 높은 음성에서 제안된 방식의 우수성이 더 큼을 알 수 있다.

표 3. 음성학적 정보를 추가한 SPLICE 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	90.70	88.86	89.90	89.80
Clean Only	87.90	87.04	86.61	87.30
Average	89.30	87.95	88.26	88.55

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	23.65%	18.83%	37.77%	25.07%
Clean Only	68.71%	70.72%	60.45%	68.20%
Average	46.18%	44.77%	49.11%	46.64%

4. 결론 및 향후 계획

SPLICE 방식은 잡음음성을 GMM 으로 모델링하고, 이것을 잡음과 미지의 채널 왜곡을 보상하는데 이용한다. 본 논문에서는 이러한 잡음 모델에 음성학적 정보를 구체적으로 도입하여 잡음 음성 모델과 보상벡터를 구하였다. 이것은 Aurora 2 DB 를 이용한 성능평가 결과 제안된 방법이 전체 인식 향상률을 43.26%에서 46.64%로 개선시켰다.

제안된 방법의 경우 모든 음소에 동일한 mixture 개수를 적용하였으나 전체 음향공간을 잘 표현하기 위해서 각 음소별 mixture 개수를 최적화하는 방법을 검토중이다.

참 고 문 헌

[1] L. Deng, A. Acero, M. Plumpe and X. Haung, "Large vocabulary continuous speech recognition under adverse conditions," *in Proc. of the ICSLP*, Beijing, Vol.3, pp.806-809, Oct. 2000.

[2] J. Droppo, A. Acero and L. Deng, "Efficient on-line acoustic environment estimation for FCDCN in a continuous speech recognition system," *in Proc. of the ICASSP*, Salt Lake City, Vol.1, pp.209-212, May, 2001.

[3] 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.

[4] 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 (2000-02), Feb. 2000.

[5] J. Droppo, L Deng and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora 2 database(web update)," *in Proc of the Eurospeech*, Aalborg, pp.217-220, Sep. 2001.

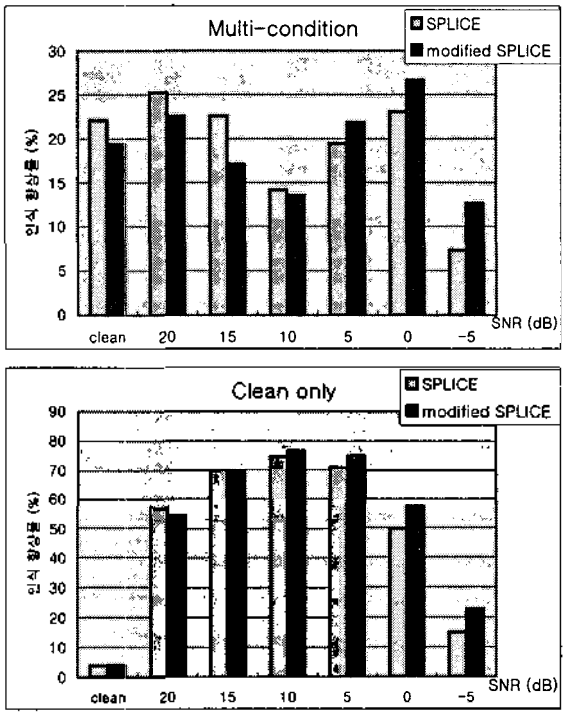

그림 1. 각 SNR 에 대한 성능 향상률 비교