스테레오 데이터에 기반한 차원별 가중 보상에 의한 음성 인식 성능 향상

김종현, 송화전, 김형순

부산대학교 전자공학과

Performance Improvement of Speech Recognition based on Stereo Data with Dimensionally Weighted Bias Compensation

Jong Hyeon Kim, Hwa Jeon Song, Hyung Soon Kim

Dept. of Electronics Engineering, Pusan National University

E-mail : {jhstudio,hwajeon,kimhs}@pusan.ac.kr

요 약

훈련 과정과 인식 과정사이의 주변 잡음과 채널 특성으로 인한 환경의 불일치는 음성 인식 성능을 급격히 저하시킨다. 이러한 차이를 극복하기 위해 다양한 전처리 방법이 제안되어 왔으며, 최근에는 스테레오 데이터와 잡음 음성의 Gaussian Mixture Model(GMM)을 이용하여 보상벡터를 구하는 SPLICE 방법이 좋은 성능을 보여주고 있다. 하지만 차원별로 특징벡터를 보상해주는 추정된 보상벡터는 underestimation되는 경향이 있으며, 그 정도가 각각의 차원마다 달라짐이 관찰되었다.

본 논문에서는 SPLICE 방법에 기반하여 추정된 보상벡터와 실제 보상벡터 사이의 관계를 관찰하여 차원별로 다른 가중치를 적용하는 차원별 가중 보상 방법을 제안하였다. 제안한 방법은 Aurora2 Clean-condition인 경우 baseline 실험 결과에 비해 68%의 높은 상대적인 인식 향상율을 얻었다.

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 있으며, 문제의 해결책으로 잡음환경에 강인한 음성인식을 위한 다양한 접근 방법이 제안되어 왔다. 특히 특징벡터 영역 기반의 잡음 보상방법은 기존의 음성 인식 시스템의 구조를 변화 시키지 않고 적용할 수 있는 장점이 있다.

최근 이러한 전처리 방법 중 하나인 Stereo-based Piecewise LInear Compensation for Environments (SPLICE) 방법이 제안되어 우수한 성능을 보여주고 있다[1][2]. 하지만 차원별로 특징벡터를 보상해주는 추정된 보상벡터는 underestimation되는 경향이 있으며 그 정도가 각각의 차원마다 달라짐이 관찰되었다. 본 논문에서는 이러한 사실에 기반하여 각 차원별로 underestimation의 정도를 관찰하여 그에 맞는 가중치를 적용하는 차원별 가중 보상 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2절에서 기존의 SPLICE 방법에 대해 살펴보고, 3절에서는 본 논문에서 제안한 차원별 가중 보상 방법에 대해 설명한다. 4절에서는 실험 환경 및 결과에 대해서 언급하고, 마지막으로 5절에서 결론을 맺는다.

2. SPLICE 방법

2.1 음성 모델과 왜곡

SPLICE 방법은 두 가지 가정을 전제로 한다. 그 중

첫번째 가정은 각각의 잡음 음성의 특징벡터 분포는 다음과 같은 Gaussian mixture로 모델링 될 수 있다는 것이다.

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

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

여기서, $p(k)$, $\boldsymbol{\mu}_k$ 및 $\boldsymbol{\Sigma}_k$ 는 각각 k 번째 Gaussian mixture의 사전 확률, 평균벡터 그리고 공분산 행렬이다. 그리고 각각의 잡음환경은 Gaussian Mixture Model (GMM)로 훈련된다.

두 번째 가정은 잡음 음성 $\mathbf{y}$와 특정 mixture가 주어졌을 때 원음성 $\mathbf{x}$의 조건부 확률 분포는 Gaussian이라는 것이다.

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

여기서 $\bar{\mathbf{r}}_k$은 깨끗한 원음성과 잡음이 섞인 음성 두 가지가 동시에 녹음된 스테레오 데이터를 사용하여 구한 보상벡터이다.

2.2 SPLICE 훈련

잡음 음성의 분포 $p(\mathbf{y})$ 는 각각의 잡음에 대해 Gaussian mixture로 모델링 될 수 있으며, 분포 $p(\mathbf{x} \mid \mathbf{y}, k)$ 에 대한 보상벡터 $\bar{\mathbf{r}}_k$ 는 스테레오 데이터가 주어진다면, maximum likelihood criterion에 의해서 다음과 같이 추정할 수 있다.

$$\bar{\mathbf{r}}_k = \frac{\sum_n p(k \mid \mathbf{y}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_n p(k \mid \mathbf{y}_n)} \tag{4}$$

여기서

$$p(k \mid \mathbf{y}_n) = \frac{p(\mathbf{y}_n \mid k) p(k)}{\sum_n p(\mathbf{y}_n \mid k) p(k)} \tag{5}$$

이다.

2.3 특징벡터 보상

2.1절에서의 두 가지 가정은 SPLICE 방법에서 잡음 음성이 주어졌을 때 원음성의 Minimum Mean Square Estimation(MMSE)을 간단하게 해준다. 잡음 음성이 주어졌을 때 구한 원음성의 MMSE는 다음과 같이 정리된다.

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{y} + \sum_k p(k \mid \mathbf{y}) \bar{\mathbf{r}}_k \tag{6}$$

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

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

SPLICE 잡음 보상은 다음의 두 단계로 적용된다. 첫 단계에서 잡음 음성의 매 프레임마다 식 (7)에 의해 최적 mixture를 찾는다. 다음 단계로 식 (6)을 사용하여 그 mixture에 대응하는 보상벡터를 잡음 음성의 특징벡터에 더해준다.

3. 차원별 가중 보상 방법

3.1 차원별 보상벡터의 underestimation

본 논문에서 특정 차원에 대한 보상벡터의 estimation 오차에 대한 척도를 식 (8)과 같이 사용하였다.

$$\mathrm{CE}(d) = \frac{1}{N} \sum_n \frac{[\bar{r}(n,d) - r(n,d)]}{\sigma(d)} \tag{8}$$

여기서 n은 프레임 인덱스, N은 전체 프레임 수이고, d는 차원 인덱스이다. $\bar{r}$와 r은 각각 추정된 보상벡터와 스테레오 데이터를 이용해 구한 실제 보상 벡터값이다. 그리고, $\sigma(d)$는 각 차원별 표준편차이다. 추정오차 척도인 $\mathrm{CE}(d)$ 값이 0에 가까우면 보상이 제대로 된

것이고, 양수 쪽으로 멀어지면 overestimation, 음수 쪽으로 멀어지면 underestimation 되는 것이다.

그림 1 은 Aurora 2 데이터베이스의 훈련데이터에서 잡음이 섞인 음성과 깨끗한 원음성사이의 각 차원에 따른 보상벡터의 CE(d) 를 나타낸 것이다. 특징벡터의 차원이 높을수록 보상벡터의 underestimation되는 정도가 증가하는 경향을 관찰할 수 있다. 본 논문에서는 높은 차원의 보상벡터에 대해 가중치를 주어 좀더 잡음에 강인한 차원별 가중 보상 방법을 제안한다.

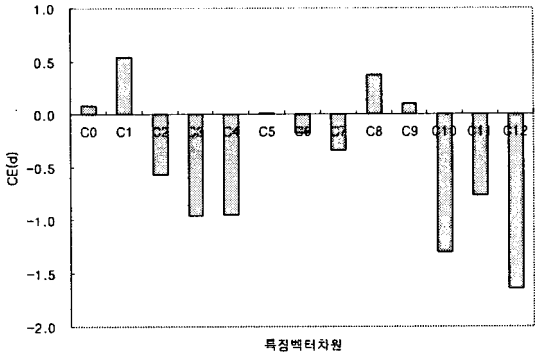

그림 1. 차원별 보상벡터의 estimation 오류

3.2 차원별 가중 보상 방법

본 논문에서는 차원별 가중 보상 방법을 제안한다. d 번째 차원에 대한 특징벡터를 보상하는 방법은 식 (9)과 같이 간략화 할 수 있다.

$$\hat{x}(d) = y(d) + \tilde{r}(d) \tag{9}$$

여기서 $\hat{x}(d)$ 및 $y(d)$ 는 각각 보상된 특징벡터와 보상되기 전의 특징벡터의 d 번째 차원이며, $\tilde{r}(d)$ 는 본 논문에서 제안한 보상벡터의 d 번째 차원 이다. 이는 SPLICE에 의해 구해진 보상벡터의 높은 차원에 대해 가중치를 크게한 형태이며 식 (10)과 같이 정리될 수 있다.

$$\tilde{r}(d) = w(d)\bar{r}(d) \tag{10}$$

여기서 $\bar{r}(d)$ 는 식 (4)에 의해 추정된 보상벡터의 d 번째 차원이며 $w(d)$ 는 차원별 가중 함수이다.

가중 함수의 형태는 여러 가지가 있을 수 있으나 본 논문에서는 식 (11)과 같은 시그모이드 함수를 사용하였다. 그림 2 에 가중함수의 예를 나타내었다.

$$w(d) = \alpha + \frac{1}{1+\exp(\beta - d)} \tag{11}$$

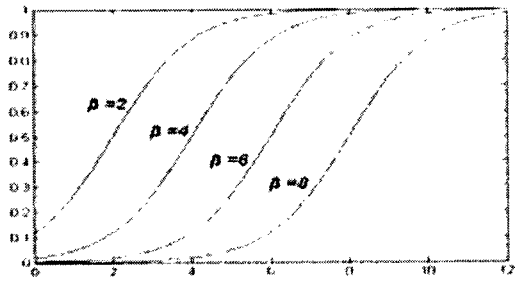

그림 2. β 값의 변화에 따른 가중 함수 특성

($\alpha = 0$ 인 경우)

4. 실험 및 결과

4.1 Aurora 2 데이터베이스

제안된 방법의 평가를 위해서 Aurora 2 데이터베이스 [3]가 사용되었다. Aurora 2 데이터베이스는 1자리에서 7자리까지의 영어 연결숫자로 구성된 TI Digit에 다양한 잡음을 인공적으로 부가한 것이다. 인식 모델은 2가지 방법으로 훈련이 되는데, 8440개의 clean 발성으로 구성된 clean-condition과 동일한 발성을 20개의 잡음환경에 나누어 각 422개의 noisy utterance로 구성된 잡음 발성으로 훈련된 multi-condition이 있다. 20개의 잡음 환경은 4가지의 잡음종류(subway, babble, car, exhibition)와 각각의 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 7가지의 잡음 레벨로 구성된다. 성능 평가는 각 잡음의 종

류에 대해서 20dB에서 0dB까지의 잡음 레벨에 대해 수행된다.

4.2 차원별 가중 보상 방법 적용 결과

본 논문에서는 식 (11)에 표현된 차원별 가중함수에서 α 값으로 1을 사용했으며, β 값을 변화시키면서 실험하였다. 그림 3 은 clean-condition인 경우 본 연구실에서 구현한 normalized SPLICE 방법[4]과 제안한 차원별 가중 보상 방법의 실험 결과의 비교이다. 가중함수의 β 값이 작아질수록 인식 성능이 향상됨을 볼 수 있다. 즉, 보상벡터의 낮은 차원에 비해 높은 차원에 상대적으로 가중치를 크게 할수록 인식 성능이 향상된다.

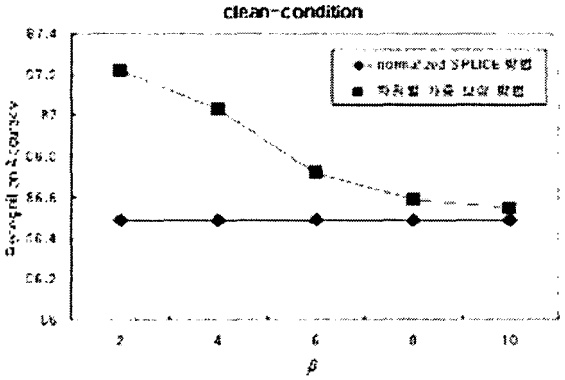

그림 3. 가중치의 변화에 따른 인식 성능

표 1 은 특징벡터 추출에 WI007 front-end[5]를 이용하여 실험한 Aurora 2 baseline 결과와 제안한 방법에 대한 인식 실험 결과를 비교해서 나타낸 것이다. Clean-condition인 경우 Aurora 2 baseline 실험 결과에 비해 68%의 높은 상대적인 인식 향상율을 얻었다.

5. 결론 및 향후 계획

본 논문에서는 차원별로 추정된 보상벡터의 underestimation을 관찰하여 이를 보상하는 차원별 가중 보상방법을 제안하였으며, Aurora2 clean-condition인 경우 68%의 음성 인식 성능 향상율을 얻었다.

제안된 방법은 모든 SNR에 대해 보상벡터의 특정 차원에 동일한 가중치를 적용하였으나, 향후 계획으로 SNR을 추정하여 그에 적절한 가중치를 적용함으로 음성 인식 성능을 추가적으로 향상시키는 방법을 검토중이다.

표 1. Baseline 및 제안 방식의 인식 성능 비교

(a) Aurora 2 baseline 실험 결과

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

(b) 차원별 가중 보상 방법을 적용한 결과

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	90.98	88.27	89.40	89.58
Clean Only	87.80	87.02	86.47	87.22
Average	89.39	87.65	87.93	88.40

(c) Baseline에 비해 제안 방식의 성능 향상 정도

Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	25.94%	14.59%	34.66%	23.44%
Clean Only	68.43%	70.68%	60.03%	68.00%
Average	47.18%	42.63%	47.35%	45.72%

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, 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.

[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] 김두희, 송화전, 김형순, "음성학적인 정보를 포함한 SPLICE를 이용한 잡음환경에서의 음성 인식," 한국음향학회 하계학술발표대회 논문집 제 21권 제 1호, pp.83-86, 2002년 11월.

[5] 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.