부분공간 분포 군집화를 이용한 SPLICE의 실시간 적용

정규준, 오영환
한국과학기술원

Real-time implementation of SPLICE using subspace distribution clustering

Gue Jun Jung, Yung-Hwan Oh
Department of Electrical Engineering & Computer Science Division of Computer Science
Korea Advanced Institute of Science and Technology
{sylph, yhoh}@speech.kaist.ac.kr

요약

본 논문에서는 훈련과정과 인식과정에서 발생하는 환경 불일치에 의한 성능 저하에 효과적으로 대처할 수 있는 방법 중의 하나인 Stereo-based Piecewise LInear Compensation for Environments(SPLICE)를 실시간으로 적용하기 위한 방법을 제안한다. SPLICE는 깨끗한 환경과 잡음 환경에서 동시에 수집한 음성 자료를 이용하여 잡음으로 인해 발생한 음성 왜곡을 Gaussian Mixture Model(GMM)로 표현하고 잡음 음성의 음질을 개선시키는 방법으로 잡음에 대한 특별한 가정 없이 켑스트럼 영역에서 효과적으로 음질을 개선시킬 수 있다. 하지만 학습된 GMM으로부터 매 프레임 출력확률을 얻어야 하기 때문에 많은 연산이 요구되어 실시간으로 적용하기 부적합하였다. 본 논문에서는 출력확률 획득 과정에 부분공간 군집화 방법을 적용하여 연산량을 효율적으로 줄여 SPLICE를 실시간으로 적용시킨다. 제안한 방법의 유효성을 검증하기 위해 Aurora2 데이터베이스를 이용한 실험 결과 기존 방법에 비해 인식성능을 유지하면서도 출력확률 계산에서 연산량을 1/15로, 음성 왜곡 모델 표현을 위한 가용 메모리를 1/10로 감소시켰다.

1. 서론

인간의 가장 편리한 의사전달 수단인 음성은 차세대 사용자 인터페이스를 위한 핵심 요소 기술로 그 필요성이 증대되고 있으나 사용 환경에서 발생하는 주변 잡음의 영향으로 인해 급격히 성능이 저하되어 실용화에 어려움을 겪고 있다. 이러한 환경 불일치에 대응하기 위해 잡음이 포함된 관측 자료에서 잡음을 제거하는 음질 개선 방법에서부터 음성 모델을 사용 환경에 적응시키는 방법까지 다양한 연구가 진행되고 있다.[1]

최근 음질개선방법 중 하나인 Stereo-based Piecewise Linear Compensation for Environments(SPLICE) 방식이 제안되어 우수한 성능을 보여주었다.[2] SPLICE는 프레임 기반의 특징 벡터 영역에서의 잡음을 제거 방법으로 깨끗한 환경과 잡음 환경에서 동시에 녹음된 음성 자료를 이용하여 음성 왜곡 모델을 Gaussian Mixture Model (GMM)으로 표현한다. 이 방식은 매 프레임마다 음질 개선을 위해 음성 왜곡 모델의 출력 확률을 요구하기 때문에 많은 연산이 필요하여 실시간으로 적용하기에는 부적합하다.

본 논문에서는 이러한 제약을 극복하고 SPLICE를 실시간으로 적용하기 위해 부분공간 군집화 방법을 음성 왜곡 모델에 적용한다. 부분공간 군집화는 GMM이 표현하고 있는 공간을 직교 부분공간 (orthogonal subspace)으로 분할하고 해당 부분공간에 포함된 부분확률 분포들 중 비슷한 분포들을 하나로 묶어(tying) 양자화 하는 방법으로 가우시안 분포에 비해 모델 표현에 필요한 메모리 공간을 최소화 하면서도 모델의 표현 능력을 유

지 시킬 수 있고 연산량 감소의 장점을 가지고 있기 때문에 음성인식 모델을 이동통신 기기에 적합하도록 구성할 수 있다.[3][4] 이러한 장점을 가진 부분공간 군집화는 확률 분포에서 부분공간을 어떻게 구성하는가 하는 것이 중요한 문제가 된다. 본 논문에서는 양자화 오류를 줄이면서, 각 부분공간의 차원을 자동으로 결정하는 부분공간 구성 방법을 적용한다.

2장에서는 기존 SPLICE를 개괄적으로 살펴보고, 3장에서는 본 논문에서 제안한 부분공간 분포 군집화에 관해 설명한다. 4장에서는 기존 방법과 제안한 방법을 적용한 실험 결과를 제시하고, 5장에서 결론을 맺는다.

2. SPLICE

2.1 음성 모델과 왜곡

SPLICE는 잡음이 섞이지 않은 깨끗한 음성 x와 잡음에 의해 왜곡된 음성 y에 대해서 다음과 같은 두 가지를 가정한다.

첫 번째 가정은 잡음 음성의 켑스트럼 벡터의 분포는 다음과 같이 GMM으로 표현할 수 있다는 것이다.

$$p(y)=\sum_s p(y|s)p(s)\ \ (1\le s\le N) \tag{1}$$

식에서 $p(y|s)=N(y;\mu_s,\Sigma_s)$이고 $p(s)$, μ_s, Σ_s는 각각 s번째 가우시안 분포의 사전 확률, 평균 벡터 그리고 공분산을 의미한다. GMM은 각각의 잡음 환경에 대해서 개별적으로 훈련된다.

두 번째 가정은 잡음 음성 y가 주어졌을 때 깨끗한 음성 x의 평균 벡터는 잡음 음성의 평균 벡터와 선형 변환 관계를 가진다는 것이다. 이때 선형 행렬을 단위 행렬로 가정하면 원음성의 잡음 음성에 대한 조건부 확률을 다음과 같은 형태로 표현할 수 있다.

$$p(x|y,s)=N(x;y+\gamma_s,\Gamma_s) \tag{2}$$

식에서 γ_s와 Γ_s는 각각 s번째 가우시안 분포에 의존하는 보상 벡터와 공분산을 의미한다.

2.2. 훈련과정

잡음 음성의 특징 벡터의 분포 $p(y)$는 첫 번째 가정에서 GMM을 따른다고 하였으므로 EM알고리즘을 이용하여 μ_s와 Σ_s를 추정할 수 있으며 초기 파라미터는 VQ clustering을 이용하여 결정한다. stereo 자료가 주어졌을 경우 분포 $p(x|y,s)$에 대한 보상 벡터 γ_s는 MMSE (Minimum Mean Squared Error)에 의해서 다음과 같이 추정할 수 있다.

$$\gamma_s=\frac{\sum_n p(s|y_n)(x_n-y_n)}{\sum_n p(s|y_n)} \tag{3}$$

식에서 $p(s|y_n)=\dfrac{p(y_n|s)p(s)}{\sum_s p(y_n|k)p(s)}$ 를 의미한다.

2.3. 켑스트럼 보상

2.1절의 두 가정은 SPLICE에서 잡음 음성에 대한 원음성의 MMSE 추정을 간단하게 만들어 준다. 잡음 음성이 주어졌을 때 MMSE 추정에 의한 원음성의 기댓값은 다음과 같다.

$$\hat{x}_{MMSE}=E_x[x|y]=\sum_s p(s|y)E_x[x|y,s] \tag{4}$$

식(2)를 이용하면 $E_x[x|y,s]$는 다음과 같이 정리된다.

$$E_x[x|y,s]=y+\gamma_s \tag{5}$$

식(5)을 식(4)에 대입하면 식(5)는 다음과 같이 정리할 수 있다.

$$\hat{x}_{MMSE}=y+\sum_s p(s|y)\gamma_s \tag{6}$$

즉, 원 음성은 각각의 가우시안 분포에 포함되는 보상 벡터들의 가중 합으로 표현할 수 있다.

빠른 구현을 위해서 식(6)의 $p(s|y)$를 식(7)과 같이 가정하면 원 음성은 MAP 방법으로 추정하는 것과 동일하게 되며 식(8)과 같이 간략화 된다.

$$\hat{p}(s|y)=\begin{cases}1 & ,\hat{s}=\arg max_s p(s|y)\\ 0 & ,others\end{cases} \tag{7}$$

$$\hat{x}_{MAP}=y+\gamma_s \tag{8}$$

3. 부분공간 군집화

3.1. 부분공간 군집화 개요

부분공간 군집화는 SDCHMM (Subspace Distribution Clustering HMM)를 통해 제안된 방법으로 <그림1>과 같이 D차원 공간을 표현하는 가우시안 분포를 부분공간 정의에 따라 분할한 후 비슷한 형태의 부분 공간 분포들

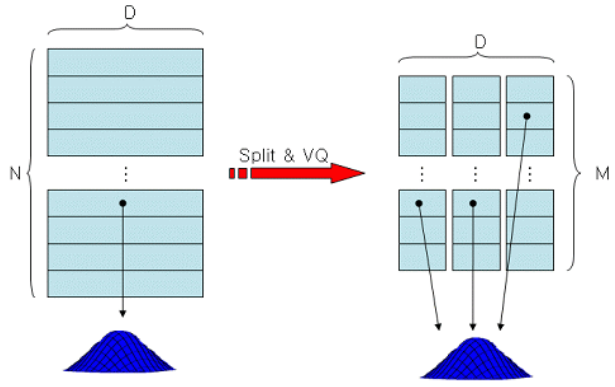

그림 1. 부분공간 군집화

을 분포의 정형(prototype)으로 묶는다.[3] 따라서 원래의 전체 공간 분포는 식(9)와 같이 부분공간 분포의 정형들의 조합으로써 원래 분포 값에 가깝게 추정된다.

$$p(O) = \sum_{m=1}^{M} w_m N(O;\mu_k,\Sigma_k) \approx \sum_{m=1}^{M} w_m \prod_{k=1}^{K} N^{tied}(O_k;\mu_{mk},\Sigma_{mk}) \quad (9)$$

3.2. 양자화 오류를 고려한 부분공간 구성

3.2.1. 부분공간과 벡터 양자화

벡터 양자화를 적용할 때 벡터의 차원이 작을수록 양자화 오류가 감소하므로 벡터를 부벡터(sub-vector)로 분할한 다음, 각 부벡터 별로 양자화 하는 경우가 전체 벡터를 양자화 하는 경우보다 양자화 오류가 더 작게 된다.[5] 이 때, 부벡터를 구성하는 방법에 따라 양자화 오류의 크기가 변화함으로 부분공간을 나누어 주는 기준이 중요한 요소가 된다. 먼저 가우시안 분포를 부분공간으로 나누어 준 다음 양자화 되는 가정을 살펴보면 <그림 1>과 같다. 이 과정에서 벡터 양자화 오류가 최소가 되기 위해서는 벡터의 각 차원의 값들이 비슷한 분포를 가져야 한다. <그림 2>의 경우를 살펴보면 (b)의 경우보다 (a)의 경우에 더 양자화 오류가 적음을 알 수 있다. 따라서 전체 가우시안 분포를 부분공간으로 분할

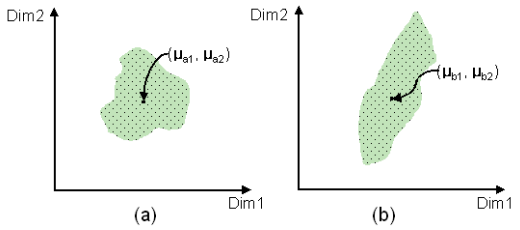

그림 2. 분포에 다른 양자화 오류

할 때 각 차원에 포함된 값들의 분산이 비슷한 차원들을 하나의 부분공간으로 만들어 주어야 한다. 또한 가우시안 분포의 경우 평균값의 오류에 민감하게 반응하기 때문에 부분공간을 설정할 때 가우시안 분포의 각 차원의 평균값들에 대한 분산의 비가 비슷한 차원들을 같은 부분공간으로 해 주도록 한다.[6]

3.2.2. 평균값의 분산에 따른 부분공간 구성

앞 절에서 살펴본 것과 같이 가우시안 분포의 평균값들의 분포를 고려하여 부분공간을 나누어주기 위해서 agglomerative clustering algorithm을 이용하였다. 이때 각 부분공간의 차원은 군집화의 결과에 따라 자동적으로 결정된다. 군집화 알고리즘에서 사용되는 거리 척도는 식 (10)과 같다. 식(10)에서 σ_i는 i번째 차원의 평균값의 분산을 의미한다.

$$Distance(i,j) = \left|\log\frac{\sigma_i}{\sigma_j}\right| \quad (10)$$

분산의 차이를 비교하는데 있어서 전체 파라미터들이 퍼져있는 영역의 차이를 고려하기 위하여 두 분산의 비를 사용하였으며, 이 값이 0에 가까울수록 두 분포들은 비슷한 분산 값을 갖는다. 제안한 거리 척도를 이용하여 부분공간을 결정하는 과정을 요약하면 다음과 같다.

- 각 차원에 대해서 평균값의 분산 계산
- $\log\sigma_1, \cdots, \log\sigma_D$를 agglomerative clustering algorithm을 적용하여 K개의 군집으로 구성
- 군집화의 결과를 이용하여 부분 공간 구성

4. 실험 및 결과

4.1. 실험 환경

본 논문에서 제안한 부분공간 군집화 방법을 적용한 SPLICE의 유효성을 검증하기 위하여 Aurora2 데이터베이스를 사용하여 실험하였다. Aurora2 데이터베이스는 1자리에서 7자리까지의 연결 숫자로 구성된 TI DIGIT 음성에 다양한 잡음을 인공적으로 부가한 자료이다.[7]

특징 벡터는 알고리즘 비교를 위해 Aurora2 데이터베이스에 포함된 W1007 front-end를 변형하여 power spectrum에서 추출한 12차 MFCC 계수와 c0를 추출하였다.(D=13) 음성 왜곡모델의 생성을 위해 multi-condition 훈련 자료에 나타난 17개의 잡음 환경에 대해서 GMM

을 생성하였으며, 각 잡음환경에 대해 256개의 가우시안 분포를 사용하였다.(N=4352) 채널 보상해 주기 위해서 SPLICE 훈련과 테스트 과정에 모두 Cepstral Mean Normalization (CMN)을 적용하였으며 보상벡터의 smoothing은 적용하지 않았다.

부분공간 수는 실험을 통해서 메모리 감소와 인식률, 및 속도를 고려하여 4개(K=4)로 정해주었으며, 각 부분공간의 코드북 사이즈는 256개(M=256)로 정해 주었다.

4.2. 실험 결과

먼저 기존 SPLICE와 제안한 부분공간 군집화를 적용한 SPLICE의 인식률을 비교해 보았으며 결과는 <표 1>과 같다. 다음으로 출력확률 계산에 필요한 연산량을 정량적으로 비교해 보았으며 결과는 <표 2>와 같다. 마지막으로 음성왜곡 모델을 표현하기 위해 필요한 메모리 요구량을 비교해 보았으며 결과는 <표 3>과 같다.

실험 결과에서 확인할 수 있듯이 제안한 방법은 인식 성능의 저하 없이 출력확률에 필요한 연산량을 약 1/15로 감소시켰으며, 모델 표현을 위한 메모리 또한 1/10로 감소시켰다. 특히 모바일 기기에 구현할 경우 제안한 방식이 더욱 큰 이점을 가지게 됨을 알 수 있다.

표 1. 기존 SPLICE와 부분공간 군잡화를 적용한 SPLICE의 인식률 (단위: %)
(SDC: Subspace Distribution Clustering)

Absolute performance (Accuracy)				
종류	set A	set B	set C	Overall
SPLICE	88.42	88.07	87.54	88.01
SDC_SPLICE	88.08	87.89	87.53	87.83
차이	0.34	0.18	0.01	0.18

표 2. 출력확률 계산에 필요한 연산량
(N: 기존SPLICE에 사용되는 가우시안 분포 수
M: 부분공간 군집화에 사용되는 코드북 크기
D: 전체벡터의 차원, K: 부분공간 수)

연산자	SPLICE	SDC_SPLICE
*,/	4352 * 13 (N*D)	256 * 13 (M*D)
+,-	4352 * 13 (N*D)	4352*4 + 256*13 (N*K + M*D)

표 3. 음성왜곡 모델의 표현을 위한 메모리 크기

SPLICE	SDC_SPLICE
452,608 ≈ 442 KB ($N*D*4*2$)	44,032 ≈ 43 KB ($M*D*4*2+N*(\log_2 M)/8*K$)

5. 결론 및 향후 연구

본 논문에서는 켑스트럼 영역에서 효과적으로 음질 개선 방법인 SPLICE를 실시간으로 적용하기 위한 방법을 모색하였다. 이를 위해 기존 가우시안 분포의 모델 표현에 필요한 메모리를 최소화하고 출력확률 계산에 필요한 연산량을 효율적으로 감소시켜주면서도 모델의 표현 능력을 유지시켜주는 부분공간 분포 군집화를 SPLICE에 적용하여 인식 성능을 유지하면서도 출력확률 연산에 필요한 연산과 음성왜곡 모델 표현을 위해 필요한 메모리를 대폭 감소시키는 방법을 제안하였다.

SPLICE는 stereo자료를 사용하여 우수한 성능을 보이므로 학습에 사용된 자료와 상반된 잡음을 처리할 경우 최적의 성능을 이끌어내지 못할 것이다. 앞으로 이러한 제약을 완화하기 위한 방법의 연구가 필요할 것으로 판단된다.

참고문헌

1. Y. Gong, "Speech recognition in noisy environments: A survey", Speech Communication, vol. 16, pp. 261-291, 1995.
2. J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE algorithm on Aurora2 database", Proc. of Eurospeech, pp217-220, 2001
3. Enrico bocchieri, Brian Kan-Wing Mak,, “Subspace Distribution Clustering Hidden Markov Model”, IEEE Trans. Speech Audio Processing, vol.9, No3, 2001
4. Imre Varga, Stefanie Aalburg, “ASR in Mobile Phones - An Industrial Approach”, IEEE Trans. Speech and Audio Processing, vol.10, No8, 2002.
5. M. Ravishankar, “Sub-Vector Clustering To Improve Memory and Speed Performance of Acoustic Likelihood computation”, Proc. of Eurospeech, pp151-154, 1997
6. Gue Jun Jung, Su-Hyun Kim, and Yung-Hwan Oh, "An efficient codebook design in SDCHMM for mobile communication environments", Proc. of ICSLP, pp713 - 716, 2004.
7. H. G. Hirsh and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", ISCA ITRW AST2000, 2000