EMAP에 기반한 화자적응을 위한 강인한 상관계수의 예측

전 유 진, 김 동 국, 김 남 수

서울대학교 전기공학부

Robust Correlation Estimation for Rapid Speaker Adaptation

Eugene Jon, Dong Kook Kim and Nam Soo Kim

School of Electrical Engineering, Seoul National University

Kwanak, P.O.Box 34,Seoul 151-742 Korea

Tel: +82-2-884-1824 Fax: +82-2-878-1452

e-mail : {box99,dkkim11,nkim}@snu.ac.kr

요약문

본 논문에서는 probabilistic principal component analysis (PPCA)를 이용하여 extended maximum *a posteriori* (EMAP)에 기반한 화자적응 시스템의 성능을 향상시키는 방법을 제시하고자 한다. PPCA는 각각의 hidden Markov model (HMM) 사이의 상관계수 행렬을 강인하게 예측하는데 적용된다. 이렇게 구한 상관계수 행렬은 화자적응 시스템에 사용된다. PPCA는 연산이 효율적이고, EMAP에서 기존에 사용되었던 방법에 비해 향상된 성능을 보여준다. 여러 차례의 음성인식 실험을 통하여, PPCA를 적용한 EMAP은 적은 양의 적응 데이타에서 좋은 성능을 보인다는 것을 확인할 수 있다.

1 서론

최근에 음성인식 기술의 발달로 인하여, 화자독립(Speaker Independent, SI) 연속 음성인식 시스템의 성능도 어느 정도 향상되었다. 하지만 훈련 데이타에 들어있지 않은 화자에 대해서는 여전히 성능이 떨어진다. 화자독립 시스템과 화자종속(Speaker Dependent, SD) 시스템의 성능 차이를 극복하기 위해 다양한 화자적응 방법이 연구되어 왔다 [1]. 이러한 방법 중의 하나가 maximum *a posteriori* (MAP) [2]인데, 이 방법은 음성인식 파라미터의 priori 정보를 활용한다. MAP 방식의 가장 큰 장점은 적응 데이터의 양이 늘어나면 적응된 파라미터들이 화자종속 모델로 수렴한다는 것이다. 하지만, MAP 방식은 단지 관측된 파라미터를 변형시키기 때문에, 빠른 화자적응에 적합하지 않다. 그러한 이유로 MAP에 기반한 화자적응에서 적은 양의 적응 데이타를 사용할 경우의 성능 향상을 위해 extended MAP (EMAP) [3]가 제안되었다. EMAP은 파라미터간의 상관관계를 이용하여 관측되지 않은 파라미터도 변환시킨다.

Principal component analysis (PCA) [4]는 데이타 집합의 분산도를 유지하면서 차원을 줄이기 위해서 많은 통계적 application에 사용된다. PCA의 이러한 특성은 적은 양의 적응 데이타를 충분히 활용할 수 있기 때문에, 빠른 화자적응에 유용하다. 확률 밀도 모델과 결합하게 되면 PCA는 probabilistic PCA (PPCA)가 된다.

본 논문에서는 PPCA를 사용해서 EMAP 방식에 이용되는 상관계수 행렬을 구하는 빠른 화자적응 방법을 제안한다. 여러 실험을 통해 기존의 EMAP 알고리즘과 비

교를 한 결과, 제안된 방법의 성능이 더 좋은 결과를 보인다.

2 Extended MAP

MAP 예측 방법은 예측될 파라미터의 prior 정보를 이용하는 것이다. 이것은 maximum likelihood (ML) 방법에 비하여, 강인한 파라미터 예측을 위해서 더 적은 양의 데이타를 필요로 한다. 하지만, MAP를 이용한 화자적응은 관측된 파라미터만 변환시키는 단점이 있다. HMM에 기반한 일반적인 음성인식 시스템은 보통 수백만개의 파라미터를 사용하기 때문에 실질적으로 MAP를 통한 빠른 화자적응은 불가능하다. EMAP은 빠른 화자적응에서의 MAP의 성능을 향상시키기 위해서 개발되었다. EMAP은 모든 Gaussian 확률 분포가 서로 상관관계가 있다고 가정하고, 상관관계 정보를 이용해서 관측되지 않은 파라미터를 변환한다. $\mathbf{m} = [\mathbf{m}_1^T, \ldots, \mathbf{m}_K^T]^T$ 라고 하자. 여기서 $\mathbf{m}_j$은 j번째 Gaussian 확률분포의 mean 벡터이고, K는 Gaussian 분포의 총 숫자이고, T는 행렬의 transpose를 나타낸다. 우리는 이 벡터 $\mathbf{m}$을 supervector라고 부르는데 그 이유는 이 벡터가 모든 Gaussian mean으로 구성되어있기 때문이다. 각각의 mean 벡터들은 서로 상관관계가 있고, $\mathbf{m}$ 의 확률밀도함수(pdf)는 다음과 같이 표현된다.

$$g(\mathbf{m}) = N(\mathbf{m}_0, \mathbf{S}_0) \tag{1}$$

여기서 $N(a,b)$는 평균 a와 공분산 b를 가진 정규분포를 나타낸다. 일반적으로, $\mathbf{m}_0$는 HMM 훈련을 통해서 얻어진다. HMM 훈련에서는 훈련 데이타를 바탕으로 ML을 사용하여 $\mathbf{m}_0$를 예측하게 된다. 이 과정은 [3]에 자세하게 소개되어 있고 여기서는 결과만 보여주겠다.

설명을 단순화하기 위해, 단일 관측 데이타 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L$만 적응에 사용된다고 하자. 여기서 L은 전체 프레임 수를 나타낸다. 만약 Gaussian mean 벡터의 prior pdf가 (1)처럼 주어진다면, MAP 방식에 의해 적응된 supervector $\hat{\mathbf{m}}_0$ 는 다음과 같이 표현됨을 알 수 있다 [3].

$$\hat{\mathbf{m}}_0 = \mathbf{S}(\mathbf{S} + \mathbf{S}_0\mathbf{C})^{-1}\mathbf{m}_0 + \mathbf{S}_0(\mathbf{S} + \mathbf{C}\mathbf{S}_0)^{-1}\mathbf{C}\mathbf{A} \tag{2}$$

여기서 $\mathbf{S} = diag(\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K)$이며, $\mathbf{\Sigma}_j$는 j번째 Gaussian의 공분산 행렬이다. $C = diag(c_1, \cdots, c_K)$이며, $c_k = \sum_{l=1}^{L} c_{kl}$는 관측된 Gaussian 중에 k번째 Gaussian의 count이다. $\mathbf{A} = ((\mathbf{m}_1)_{ML}^T, \cdots, (\mathbf{m}_K)_{ML}^T)^T$이며, $(\mathbf{m}_k)_{ML} = \frac{\sum_{l=1}^{L} c_{kl}\mathbf{x}_l}{c_k}$이다. 식 (2)의 오른편은 prior 정보와 주어진 데이타 간의 linear interpolation으로 생각할 수 있다. 효율적인 계산을 위해, shift된 평균값 $\hat{\mathbf{m}}_0 - \mathbf{m}_0$를 ML mean shift $\mathbf{A} - \mathbf{m}_0$의 함수로 표현한다. (2)의 식에서 우리는 다음과 같은 식을 얻을 수 있다.

$$\hat{\mathbf{m}}_0 - \mathbf{m}_0 = \mathbf{S}_0(\mathbf{S} + \mathbf{C}\mathbf{S}_0)^{-1}\mathbf{C}(\mathbf{A} - \mathbf{m}_0) \tag{3}$$

일반적으로, (3)의 상관계수 행렬 $\mathbf{S}_0$은 화자종속 모델 파라미터를 사용하여 다음과 같이 구한다:

$$\mathbf{S}_0 = \frac{1}{N_{sp}} \sum_{i=1}^{N_{sp}} (\mathbf{m}_i - \mathbf{m}_{SI})(\mathbf{m}_i - \mathbf{m}_{SI})^T \tag{4}$$

여기서 $\mathbf{m}_i$은 i번째 화자의 supervector이고, N_{sp}은 훈련 데이타 화자의 총 수이며, $\mathbf{m}_{SI}$는 모든 화자독립 Gaussian 벡터로 이루어진 화자독립 supervector를 나타낸다.

3 Probabilistic PCA

여기서 PPCA에 대한 개념과 관련된 수식들에 대해 살펴본다. $\mathbf{y} = [y_1, y_2, \cdots, y_D]^T$를 D차원의 관측벡터라 하자. 그리고 $\mathbf{y}$는 $P(\ll D)$ 차원의 latent 변수 $\mathbf{x} = [x_1, x_2, \cdots, x_P]^T$와 다음과 같은 관계를 갖는다고 가정하자.

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mu_{\mathbf{y}} + \epsilon \tag{5}$$

여기서 $\mathbf{W}$는 관측데이터의 주요 subspace을 나타내는 $D \times P$ 파라미터 행렬이고, $\mu_{\mathbf{y}}$는 $\mathbf{y}$의 mean 벡터이며, ϵ는 $\mathbf{x}$와 독립적인 Gaussian random noise이다. 일반적으로 latent 변수는 unit variance을 갖는 독립적인 Gaussian으로 다음과 같이 정의된다.

$$p(\mathbf{x}) = (2\pi)^{-P/2} \exp\{-\frac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{x}\} \tag{6}$$

Noise는 $\epsilon \sim N(0, \sigma^2\mathbf{I})$와 같은 한 개의 Gaussian으로 모델링되는데, 여기서 $\mathbf{I}$는 $D \times D$는 identity 행렬이다. 위와 같은 가정이 주어지는 경우, 관측벡터는 다음과 같이 normal 분포를 갖게된다.

$$p(\mathbf{y}) = (2\pi)^{-D/2}|\mathbf{\Sigma}_{\mathbf{y}}|^{-1/2} \cdot \exp\{-\frac{1}{2}(\mathbf{y} - \mu_{\mathbf{y}})^T\mathbf{\Sigma}_{\mathbf{y}}^{-1}(\mathbf{y} - \mu_{\mathbf{y}})\} \tag{7}$$

여기서 $\mathbf{\Sigma_y} = \sigma^2\mathbf{I} + \mathbf{WW}^T$ 이다. $\mathbf{x}$가 주어진 경우 $\mathbf{y}$에 대한 조건부 확률분포는 다음과 같이 구해진다.

$$p(\mathbf{y}|\mathbf{x}) = (2\pi\sigma^2)^{-D/2} \cdot \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{Wx} - \mu_\mathbf{y}||^2\} \quad (8)$$

관측열 $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\}$가 주어지는 경우, PPCA 기법은 latent 변수열 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$을 추정하고 최적의 모델 파라메터 $\hat{\lambda} = \{\hat{\mathbf{W}}, \hat{\mu}_\mathbf{y}, \hat{\sigma}^2\}$을 다음과 같은 ML 기준에 의해 구한다.

$$\hat{\lambda} = \underset{\lambda}{argmax}\ [\log p(\mathbf{Y}|\lambda)] \quad (9)$$

그러나 latent 변수 $\{\mathbf{x}_t\}$는 hidden 변수로 생각할 수 있으므로 파라미터 값을 반복적으로 갱신하는 EM 알고리즘을 적용할 수 있다. $\lambda^{(n)} = \{\mathbf{W}^{(n)}, \mu_\mathbf{y}{}^{(n)}, \sigma^{2,(n)}\}$을 n번째 과정에서 얻어진 파라미터 값이라 하자. 새로운 파라미터 값 $\lambda^{(n+1)} = \{\mathbf{W}^{(n+1)}, \mu_\mathbf{y}{}^{(n+1)}, \sigma^{2,(n+1)}\}$는 다음과 같은 식에 의해 얻어진다.

$$\lambda^{(n+1)} = \underset{\lambda}{argmax}\ Q(\lambda^{(n+1)}, \lambda^{(n)}) \quad (10)$$

여기서,

$$Q(\lambda^{(n+1)}, \lambda^{(n)}) = E\left[\log p(\mathbf{Y}, \mathbf{X}|\lambda^{(n+1)})|\mathbf{Y}, \lambda^{(n)}\right] \quad (11)$$

위의 식에 기반하여 다음과 같이 파라미터에 대한 update 식을 유도할 수 있다.

$$\mu_\mathbf{y}{}^{(n+1)} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{y}_t - \mathbf{W}^{(n)}\overline{\mathbf{x}_t}\right) \quad (12)$$

$$\mathbf{W}^{(n+1)} = \left[\sum_{t=1}^{T}(\mathbf{y}_t - \mu_\mathbf{y}^{(n+1)})\overline{\mathbf{x}_t}\right]\left[\sum_{t=1}^{T}\overline{\mathbf{x}_t\mathbf{x}_t^T}\right]^{-1} \quad (13)$$

$$\sigma^{2,(n+1)} = \frac{1}{DT}\sum_{t=1}^{T}\left\{||\mathbf{y}_t - \mu_\mathbf{y}^{(n+1)}||^2 - 2\overline{\mathbf{x}_t^T}\mathbf{W}^{T,(n+1)} \cdot \left(\mathbf{y}_t - \mu_\mathbf{y}^{(n+1)}\right) + tr\left(\overline{\mathbf{x}_t\mathbf{x}_t^T}\mathbf{W}^{T,(n+1)}\mathbf{W}^{(n+1)}\right)\right\} \quad (14)$$

여기서,

$$\overline{\mathbf{x}_t} \equiv E\left[\mathbf{x}_t|\mathbf{y}_t, \lambda^{(n)}\right] = \Sigma_\mathbf{x}^{-1}\mathbf{W}^T(\mathbf{y}_t - \mu_\mathbf{y}) \quad (15)$$

$$\overline{\mathbf{x}_t\mathbf{x}_t^T} \equiv E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t, \lambda^{(n)}\right] = \sigma^2\Sigma_\mathbf{x}^{-1} + \overline{\mathbf{x}_t} \cdot \overline{\mathbf{x}_t}^T \quad (16)$$

이고, $\mathbf{\Sigma_x} = \sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W}$ 그리고 tr는 행렬의 trace을 나타낸다. (9)번 식의 오른쪽의 log-likelihood의 값은 $\mathbf{W}$가 관측데이터의 주요 subspace을 span하는 경우에 최대값이 된다. [5]에서 증명한 것과 같이 ML 추정치, $\mathbf{W}_{ML}$ 값은 관측 데이터의 공분산 행렬의 주요 eigenvector들을 나타낸다.

4 Extended MAP에 사용된 PPCA

(4)에서 구한 화자독립 supervector $\mathbf{m}_1, \cdots, \mathbf{m}_{N_{sp}}$가 PPCA의 기준에서는 관측된 데이터라고 가정하자. (5)에서 주어진 latent 변수 모델을 바탕으로, 앞에서 계산했던 파라미터 W 과 σ^2를 이용해서 다음과 같이 상관계수 행렬을 구할 수 있다.

$$\mathbf{S}_0 = \sigma^2\mathbf{I} + \mathbf{WW}^T \quad (17)$$

(3)번 식의 $\mathbf{S}_0$에 (17)번 식의 등호 오른쪽 식을 대입하면 다음과 같다.

$$\hat{\mathbf{m}}_0 - \mathbf{m}_0 = (\sigma^2\mathbf{I} + \mathbf{WW}^T) \cdot (\mathbf{S} + \mathbf{C}(\sigma^2\mathbf{I} + \mathbf{WW}^T))^{-1}\mathbf{C}(\mathbf{A} - \mathbf{m}_0) \quad (18)$$

Matrix inversion lemma를 사용하여 위의 식을 보다 효율적으로 계산할 수 있다. 우선 (18)번 식의 역행렬 연산의 괄호 내에 있는 부분을 다음과 같이 표현해 보자.

$$(\mathbf{S} + \mathbf{C}(\sigma^2\mathbf{I} + \mathbf{WW}^T))^{-1} = (\mathbf{D} + \mathbf{W}_c\mathbf{W}^T)^{-1} \quad (19)$$

여기서 $\mathbf{D} = \mathbf{S} + \sigma^2\mathbf{CI}$ 이고, $\mathbf{W}_c = \mathbf{CW}$이다. Matrix inversion lemma를 이용하여 식을 전개하면 다음과 같다.

$$(\mathbf{D} + \mathbf{W}_c\mathbf{W}^T)^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{W}_c \cdot (\mathbf{I} + \mathbf{W}_c^T\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{D}^{-1} \quad (20)$$

이 방법을 사용하면 우리는 기존의 EMAP 방법에 비해 두 가지의 이득을 얻을 수 있다. PPCA는 데이타의 분산도를 더 작은 차원으로 압축하기 때문에 적은 양의 적응 데이타에 적합하다. 또한, 제안된 방법은 연산량을 줄일 수 있다. EMAP 화자적응을 수행할 경우에, 역행렬을 구하는 부분의 연산량이 제일 많다. (20)번 식의 경우 $P \times P$ 크기의 행렬의 역행렬을 구하는데, 이것은 (3)번 식의 $D \times D$ 크기의 행렬보다 상당히 작다. 그렇기 때문에 제안된 방법을 사용할 경우에 연산량이 어느 정도 준다고 예상된다.

표 I

화자독립 시스템과 MAP, EMAP, PPCA에 기반한 EMAP(P)을 이용해 각각 화자적응을 적용한 시스템의 인식률(%).

Method	SI	MAP	EMAP	PPCA (3)	PPCA (5)	PPCA (10)	PPCA (20)
2 sent		88.6	89.0	89.7	90.0	90.0	90.1
5 sent	87.5	88.8	89.8	90.5	90.6	90.6	90.9
10 sent		89.7	91.6	91.0	90.9	91.5	91.8

5 실험 및 결과

제안된 화자적응기법의 성능을 평가하기 위해 화자독립 연속숫자음 인식실험을 수행하였다. 학습과 테스트를 위해 각각 105명, 35명의 화자를 사용하였고, 각 화자는 3~7개의 숫자로 구성된 30~40 문장을 발성하였다. 음성은 8 kHz로 샘플링되어 20 ms overlap을 갖고 10 ms 마다 30 ms을 한 프레임으로 기준하여 분석되었다. 각 프레임은 12차 mel-frequency cepstral 계수와 일차 미분치로 구성된 24차 특징벡터로 변환되었다. 숫자와 묵음은 각각 7개, 1개의 state로 된 left-to-right HMM로 구성하였고 각 state 당 2개의 mixture 성분을 사용하였다. EMAP 화자적응을 구현하기 위하여, 모든 Gaussian mean들을 cepstrum과 delta-cepstrum 2개의 supervector로 구성했다. Mean 벡터가 supervector에 배열되는 순서는 MAP 적응에 의해 자동적으로 정렬되게 하였다.

적응실험을 위해 각 화자에 대해 2, 5 그리고 10 문장(2~20 sec)을 적응데이터로 사용하고 나머지 30 문장을 인식데이터로 이용하였다. 화자적응은 레이브링된 적응데이터를 사용한 static supervised 모드로 수행하였다. PPCA latent 변수의 차원은 3,5,10,20으로 설정했다.

화자적응 실험 결과는 SI 시스템, MAP를 이용해 적응된 시스템, 기존의 EMAP을 이용해 적응된 시스템과 비교했다. 이 결과는 표 1에 자세히 나와있다. 여기서 P는 PPCA latent 변수의 차원을 나타낸다. 결과를 보면, 제안된 방법은 기존의 EMAP 방법에 비해 적은 양의 적응데이터에서 특히 좋은 성능을 보임을 알 수 있다. 적응 데이터가 2 문장일 경우 기존의 EMAP에 비해 word error rate(WER)이 6.1% ($P = 3$) 에서 10.0% ($P = 20$) 감소한다. 또한 P의 값이 커질 수록 제안된 방법의 성능이 향상됨을 알 수 있다.

6 결론

본 논문에서는 EMAP에 기반한 화자적응의 성능을 향상시키기 위해 PPCA를 사용한 새로운 기법을 제안하였다. PPCA를 사용하여, EMAP 화자적응에 사용될 강인한 상관계수 행렬을 구한다. 제안 된 방법은 화자적응에서 향상된 성능을 보이고, 특히 적은 양의 적응 데이타에서 좋은 성능을 보인다. 또한, 연산량을 어느 정도 줄였다. 이러한 이유로, 제안된 방법이 빠른 화자 적응에 적합하다고 주장한다.

참고문헌

[1] P. C. Woodland, "Speaker adaptation: techniques and challenges," *ASRU Workshop*, vol. 1, pp. 85-90, 1999.

[2] J. L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 291-298, Apr. 1994.

[3] G. Zavaliagkos, "Maximum a Posteriori Adaptation Techniques for Speech Recognition," PH.D thesis,Northeastern University, Oct 1995.

[4] I. T. Jolliffe, *Principal Component Analysis.* Springer-Verlag, 1986.

[5] M. E. Tipping and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analysers," *Technical Report* NCRG/97/003, July 1998.