Eigenvoice 병합을 이용한 효율적인 고속 화자 적응

최동진, 오영환
한국과학기술원

Efficient Rapid Speaker Adaptation Using Merging Eigenvoices

Dong-jin Choi, Yung-Hwan Oh
Division of Computer Science, Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology
{cdjin, yhoh}@speech.kaist.ac.kr

요약

음성 인식 분야에서는 화자 적응을 통해 화자 독립 시스템의 성능을 화자 종속 시스템에 근접시키려는 여러 가지 노력이 시도되고 있다. 특히 30 초미만의 매우 적은 양의 적응 자료를 이용하는 고속 화자 적응에 대한 관심이 증가하고 있다. 고속 화자 적응에 적합한 eigenvoice 를 이용한 적응 방법은 eigenvoice 를 구성하기 위해 너무 많은 계산량과 메모리를 요구한다. 본 논문에서는 각각 따로 계산된 eigenvoice 들을 한 번에 구성한 eigenvoice 들과 거의 같은 정확도를 갖도록 병합하여 고속 화자 적응에 이용하는 방법을 제안한다. 이 방법을 이용하면 훈련 자료의 추가시 처음부터 새롭게 eigenvoice 를 구하는 대신 추가된 자료에 대한 eigenvoice 를 구하고 병합함으로써 계산량과 메모리양을 현저히 줄일 수 있다. 실험 결과, 메모리와 계산량은 추가되는 화자 종속 모델의 수에 따라 감소하며 성능 저하는 거의 없었다.

본 연구는 정보통신부 대학 IT연구센터 육성.지원사업의 연구결과로 수행되었습니다.

1. 서론

화자 독립(SI) 음성 인식 시스템은 여러 화자로부터 수집된 많은 자료를 이용하여 훈련된 시스템으로 어떠한 사람에 대해서도 고른 성능을 나타낸다. 화자 적응이란 특정 화자의 음성을 이용하여 화자 독립 음성 인식 시스템을 그 화자에 특화되도록 보정하여 더 좋은 성능을 얻는 기술이다.

화자 적응의 대표적인 방법은 maximum a posteriori (MAP) 화자 적응 [1] 과 Maximum Likelihood Linear Adaptation (MLLR) [2] 로 대표되는 변환 기반 화자 적응이 있다. 하지만 이 방법들은 적응 자료의 양이 매우 제한적인 경우에는 각각 성능 향상이 거의 없거나 오히려 성능 저하가 일어나는 문제점이 있다.

Eigenvoice(EV) 화자 적응 [3] 은 a prior 지식을 사용하여 추정해야 하는 파라미터의 수를 줄임으로써 매우 적은 양의 적응 자료에도 성능 향상을 나타낸다.

그러나 EV 화자 적응도 eigenvoice 들을 추정하기 위해 많은 계산량을 필요로 하며, a prior 지식으로 사용할 수 있는 화자 종속 (SD) 모델이 추가 되었을 때 전체 SD 모델에 대해 다시 eigenvoice 들을 추정해야 하는 문제점이 있다.

본 논문에서는 SD 모델이 추가되었을 때, 전체 SD 모델에 대해 eigenvoice 들을 추정하지 않고 추가된 SD 모델에 대한 eigenvoice 들을 구하고 기존의 eigenvoice 들과 병합함으로써 계산량과 메모리양을 줄이는 방법을 제안한다.

2. Eigenvoice 화자 적응

EV 화자 적응에서는 화자 종속 모델로부터 principal component analysis (PCA) 를 이용한 a priori 지식을 추출하여 화자 적응에 사용한다. 자세한 알고리즘은 다음과 같은 순서를 따른다.

첫째, 화자 독립 모델과 화자 종속 모델을 구성한다. 이때 화자 종속 모델은 화자 독립 모델을 구성하기 위해 사용하였던 데이터베이스를 이용하여 각각의 화자에 대해 HMM 을 훈련시키므로써 구축할 수 있으며, mixture 의 수, state 의 수, 모델 수 등은 모두 동일하게 한다.

둘째, 각각의 화자 종속 모델로부터 supervector 를 구성한다. supervector 란 HMM output Gaussian 의 모든 평균값들을 일정한 순서로 나열한 vector 이다. μ_p^i 를 화자 p 의 화자 종속 모델에 있는 i 번째 가우시안 component 의 평균값, M 을 모든 가우시안 component 의 수라고 하면 화자 p 에 대한 supervector 는 다음과 같은 식으로 나타낼 수 있다.

$$X_p = [\mu_p^1 \mu_p^2 \cdots \mu_p^M]^T \quad (1)$$

셋째, R 개의 supervector 들에 대해 PCA 를 이용하여 eigenvector, $e(1), e(2), \cdots, e(R)$ 를 추출한다. 이 R 개의 vector 중에 eigenvalue 가 큰 K 개와 R 개의 supervector 의 평균 벡터를 $e(0)$ 를 eigenvoice 라 하며 화자 적응에 사용한다.

마지막으로, 적응된 모델의 supervector 는 다음과 같은 식으로 나타낼 수 있으며,

$$X_s = e(0) + w(1)e(1) + \cdots + w(K)e(K) \quad (2)$$

$w(i)$ 는 maximum likelihood eigen decomposition (MLED) 방법을 이용하여 계산할 수 있다.

3. Eigenvoice 의 병합

eigenvoice 를 이용한 화자 적응 방법의 가장 큰 문제점들 중 하나는 eigenvoice 를 계산하는 데에 너무 많은 계산량을 필요로 하며, 추가로 화자 종속 모델이 추가되어 eigenvoice 를 갱신하려면 기존의 eigenvoice 를 이용할 수 없고 기존에 사용하였던 화자 종속 모델과 추가된 화자 종속 모델을 합하여 다시 eigenvoice 를 계산해야 한다는 점이다. 이번 장에서는 추가된 화자 종속 모델만으로 부터 eigenvoice 를 계산하고 기존에 사용하던 eigenvoice 와 병합하여 eigenvoice 를 갱신하므로써 위의 문제점을 해결하는 방법을 제안한다.

기존에 eigenspace 모델을 병합하거나 나누는 많은 방법이 있었지만, 이 방법들은 평균 벡터를 갱신할 수는 없었다. 음성 인식의 경우 각각의 가우시안 모델의 평균이 매우 중요하므로 평균 벡터가 갱신되지 않는다면 좋은 성능 향상을 기대할 수 없다.

하지만, 최근 들어 평균 벡터를 갱신하면서 eigenspace 모델을 병합하는 방법이 제안되었다 [4,5]. 이 방법은 기존의 eigenspace 구성 전에 사용하였던 자료를 참조하지 않는다.

다음과 같이 singular value decomposition (SVD) 를 이용하여 구성된 두 개의 eigenspace 가 있다고 가정하자.

$$\Theta(X) = (\mu(X), U(X)_{np}, \Sigma(X)_p, V(X)_{Np}, N(X)) \quad (3)$$
$$\Theta(Y) = (\mu(Y), U(Y)_{nq}, \Sigma(Y)_q, V(Y)_{Mq}, N(Y)) \quad (4)$$

이때, μ 는 평균 벡터, $U(Y)$ 는 eigenspace, $\Sigma(Y)$ 는 singular value, $V(Y)$ 는 eigenspace 에 투영된 벡터, N 과

M 은 각각 X, Y 의 자료의 수, n 은 자료의 차원수, p 와 q 는 각각 X, Y 의 eigenspace 의 차원수, N 과 M 은 각각 X, Y 의 자료의 수를 나타낸다.

$\Theta(X)$, $\Theta(Y)$ 의 자료만을 이용하여 병합된 eigenspace 는 다음 식과 같이 표현할 수 있으며,

$$\Theta(Z) = \Theta(X) \oplus \Theta(Y) \tag{5}$$

$\Theta(Z)$ 는 다음과 같은 방법으로 계산할 수 있다.

$$N(Z) = N(X) + N(Y) \tag{6}$$

$$\mu(Z) = (N(X)\mu(X) + N(Y)\mu(Y))/N(Z) \tag{7}$$

$$\begin{aligned} & R\Sigma V^T \\ &= \begin{bmatrix} \Sigma(X)_{pp} V(X)^T_{Np} & G_{pq}\Sigma(Y)_{qq} V(Y)^T_{Mq} \\ o_{tN} & (\nu^T_{nt} U(Y)_{nq})\Sigma(Y)_{qq} V(Y)^T_{Mq} \end{bmatrix} \\ &+ \begin{bmatrix} U(X)^T_{np}(\mu(X)-\mu(Z))1_{N(X)} & U(X)^T_{np}(\mu(Y)-\mu(Z))1_{N(Y)} \\ \nu^T_{nt}(\mu(X)-\mu(Z))1_{N(X)} & \nu^T_{nt}(\mu(Y)-\mu(Z))1_{N(Y)} \end{bmatrix} \end{aligned} \tag{8}$$

여기서

$$g_p = U(X)^T(\mu(X) - \mu(Y)) \tag{9}$$

$$G_{pq} = U(X)^T U(Y) \tag{10}$$

$$H_{nq} = [U(Y) - U(X)G_{pq}] \tag{11}$$

$$h_n = (\mu(X) - \mu(Y)) - U(X)g_p \tag{12}$$

$$\nu_{nt} = Orthobasis(\zeta[H_{nq}, h_n]) \tag{13}$$

이고, ζ 는 매우 작은 벡터를 제거하는 함수이고, t 는 이때 남은 벡터의 개수를 나타내며, *Orthobasis* 는 Gram-Schmidt orthogonalization 을 이용하여 상호 orthogonal 한 벡터를 계산한다.

SVD 를 이용하여 위 식을 계산하면, 아래 식을 이용하여 구하고자 하는 병합된 eigenspace 의 eigenvector 와 eigenvalue 를 계산할 수 있다.

$$U(Z) = [U(X)\nu]R \tag{14}$$

$$\Sigma(Z) = \Sigma \tag{15}$$

위에서 설명한 계산법을 eigenvoice 에 적용하면 원 자료를 알지 못해도, 기존의 eigenvoice 들과 추가된 자료만으로 병합된 eigenvoice 들을 계산할 수 있다.

Hall 의 논문 [5] 에 따르면 두개의 eigenspace 를 먼저 계산하고 이를 병합하는 데에 걸리는 시간은 모든 원 데이터를 이용하여 전체 eigenspace 를 계산하는데 걸리는 시간과 거의 같다. 따라서 이미 하나의 eigenvoice 들이 계산되어 있는 경우에는 그만큼의 시간을 덜 사용하고 전체 eigenvoice 들을 구성할 수 있다.

4. 실험 및 결과

실험에 사용된 데이터베이스는 ARPA resource Management task (RM)이다. 이 자료는 약 1000 단어로 이루어진 미국식 영어 문장으로 구성되어 있고, 미국의 여러 가지 사투리가 포함되어 있다. 화자 독립 자료는 109 명의 화자로부터 녹음되었고, 화자 종속 자료는 12 명의 화자로 이루어져 있다.

모든 음성 자료로부터 12 차 MFCC 와 로그 에너지, 그리고 이 값들의 1 차, 2 차 차분을 이용해 총 39 차 벡터를 추출하여 사용하였다. 음소 집합과 발음 사전은 CMU 에서 만든 것을 사용하였고, 47 개의 음소는 3 개의 상태를 가지도록 하였다.

109 명분의 훈련 자료를 이용하여 화자 독립 모델을 훈련시켰고, 구성된 모델에 훈련자료로 사용했던 각각의 화자별 훈련자료로 MLLR 과 MAP 화자 적응을 적용하여 109 개의 화자 종속 모델을 구성하였다.

제안한 방법을 테스트하기 위한 적응 자료로는 RM DB 의 화자 종속 모델용 훈련자료인 12 명분의 자료를 이용하였다.

그림 1 은 화자 적응 방법으로 널리 사용되는 MLLR 과 MAP 방법을 eigenvoice 를 이용한 화자 적응 방법과 비교하는 실험 결과이다.

앞에서 언급하였듯이 MLLR 은 적응 자료의 수가 4 문장 이하 일 때에는 적응 전보다 오히려 낮은 성능을 나타내었고, MAP 의 경우는 성능향상 속도가 매우 느린 것을 확인할 수 있다. eigenvoice 를 이용한 화자 적응 방법은 적은 적응 자료에 대해서도 성능

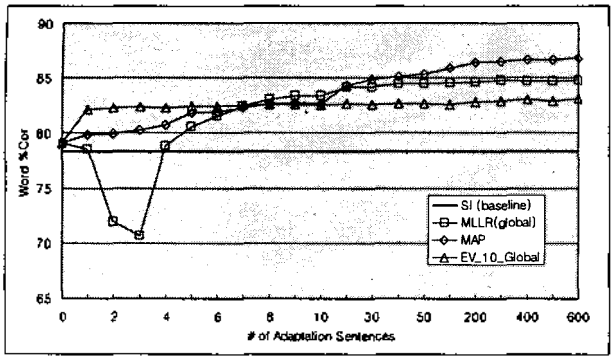

그림 1. 적응 자료 양의 증가에 따른 여러가지 화자 적응 방법의 인식률 비교

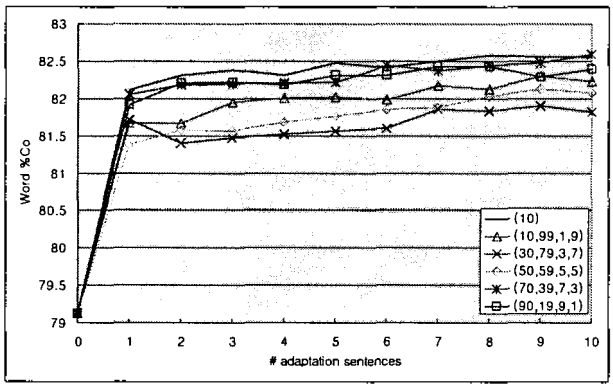

그림 2. eigenvoice의 병합을 이용한 화자 적응 방법의 인식률 비교

향상을 가져오지만 성능이 매우 빨리 수렴하여 적응 자료의 양이 많아져도 성능 향상이 거의 없는 특성을 확인할 수 있다.

그림 2 는 eigenvoice 들을 병합한 후 MLED 를 이용해 적응된 모델을 구성한 경우에 대한 실험 결과이다. eigenvoice 는 10 개를 사용하였고, 그림의 (a,b,c,d)에서 a,와 b 는 각각 기존 eigenvoice 를 구성할 때 사용한 화자 종속 모델의 수와 추가된 eigenvoice 를 구성할 때 사용한 화자 종속 모델의 수를 나타내고, c 와 d 는 각각 기존 eigenvoice 의 수, 추가된 eigenvoice 의 수를 나타낸다. 실험 결과에서 알 수 있듯이 기존의 eigenvoice 수나 추가된 eigenvoice 의 수와 관계없이 비슷한 성능을 나타냄을 알 수 있다.

5. 결론

본 논문에서는 eigenvoice 를 이용한 화자 적응에서 문제되었던 계산량을 줄이기 위해 화자 종속 모델의 추가 시에 처음부터 새로 eigenvoice 를 추정하지 않고 추가된 자료에 대한 eigenvoice 만 계산한 후 기존의 것과 병합하는 방법을 제안하였다.

이를 이용하여 eigenvoice 를 계산하기 위한 시간을 추가된 자료의 양에 따라 감소시켰으며, 성능 저하는 거의 일어나지 않았다.

참고문헌

1. C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806-814, 1991.

2. C. Leggetter and P.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

3. R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Tran. Speech and Audio Proc.*, vol. 8, no. 6, pp. 695-707, 2000.

4. P. Hall, D. Marshall, and R. Martin, "Merging and Splitting Eigenspace Models," *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 1042-1049, 2000.

5. P. Hall, D. Marshall, and R. Martin, "Adding and Subtracting Eigenspaces with Eigenvalue Decomposition and Singular Value Decomposition," *Image and Vision Computing*, vol. 20, pp. 1009-1016, 2002.