잡음 환경하에서 환경 군집화를 이용한 고속화자 적응

김영국[1], 송화전[2], 김형순[1]
[1]부산대학교 전자공학과
[2]부산대학교 컴퓨터 및 정보 통신 연구소

Fast Speaker Adaptation in Noisy Environment using Environment Clustering

Young Kuk Kim[1], Hwa Jeon Song[2], Hyung Soon Kim[1]
[1]Department of Electronics Engineering, Pusan National University,
Geumjeong-gu, Busan 609-735, Korea
ykukim, kimhs @pusan.ac.kr
[2]Research Institute of Computer Information and Communications,
Pusan National University, Geumjeong-gu, Busan 609-735, Korea
hwajeon@pusan.ac.kr

Abstract

In this paper, we investigate a fast speaker adaptation method based on eigenvoice in several noisy environments. In order to overcome its weakness against noise, we propose a noisy environment clustering method which divides the noisy adaptation utterances into utterance groups with similar environments by the vector quantization based clustering using a cepstral mean as a feature vector. Then each utterance group is used for adaptation to make an environment dependent model. According to our experiment, we obtained 19-37 % relative improvement in error rate compared with the simultaneous speaker adaptation and environmental compensation method

I. 서론

음성인식 시스템에서 훈련 환경과 테스트 환경이 다른 경우 심각한 성능 저하가 발생한다. 이러한 환경 불일치를 발생시키는 예로는 발성 환경의 차이, 화자간의 차이 등이 대표적이며, 이를 보상하기 위해 제안된 방법들은 크게 feature space에서의 보상 및 model space에서의 보상 방법으로 나눌 수 있다. Feature space에서 채널왜곡 개선방법으로 간단하지만 효과적인 cepstrum mean subtraction(CMS)이 있으며, 부가 잡음에 대한 보상으로는 spectral subtraction(SS) 등이 대표적이다. 화자간의 차이를 보상하는 방법은 주로 model space에서 이루어지며 화자적응 방법이 대표적이다. 화자적응 방식은 크게 maximum a posteriori (MAP)[2], maximum likelihood linear regression (MLLR)[3], 그리고 화자 군집화(speaker clustering) 방식 등이 있다. 그 중에서 화자 군집화 방식의 하나인 eigenvoice 기법[4]이 추정할 파라미터 수가 다른 방법에 비해 적으므로 고속 화자적응에 유리한 것으로 알려져 있다. 그러나 eigenvoice 기법에 의해 적응된 모델은 화자공간에서 환경이 다른 새로운 화자에 대한 정확한 위치를 제공하지 못한다. 따라서 환경이 달라진 경우에 이를 보상할 수 있는 적응 방식이 필요하다.

본 논문에서는 잡음 환경 하에서 소규모의 적응 데이터로부터 음향모델을 적응시킴으로써 음성인식 성능을 향상시키는 방식을 연구하였다. 이를 위해 본 논문에서는 환경 군집화에 기반을 둔 고속화자 적응 방식을 제안 하였다. 이 방식은 화자 적응 전 동일한 환경에 속한 적응 데이터를 군집화 방식으로 분리하여 환경별 화자적응 모델을 만든다. 그리고 인식실험시 테스트 환경과 유사한 적응 모델을 선택해 인식실험을 함으로

서 인식성능 향상을 가져 올 수 있다. 본 논문의 구성은 다음과 같다. 서론에 이어 제 2장에서는 기존에 적은 적응 데이터로부터 강인한 파라미터를 추정하는데 많이 사용되는 Eigenvoice 기반 고속화자적응 방법에 대해 간략히 기술한다. 그리고 제 3장에서는 본 연구에서 제안한 환경 군집화에 기반을 둔 고속화자 적응 방식에 대해 기술하고, 제 4장에서는 다양한 잡음환경에서의 성능을 평가한다. 마지막으로 제 5장에서 결론을 맺는다.

II. Eigenvoice 고속화자 적응

최근 화자적응의 한 방법으로 speaker clustering 방식의 하나인 eigenvoice 적응방법이 널리 사용되고 있다. Eigenvoice는 각 화자들 간의 변동을 가장 잘 대표하는 기저벡터를 설정하고 적응화자에 대하여 기저벡터 성분의 가중치를 추정하는 방식이다. Eigenvoice 적응방식은 적응화자 모델에 대하여 추정할 파라미터가 적기 때문에 적응 데이터가 적은 경우에 기존의 적응 방식에 비하여 성능이 우수하다.

Eigenvoice 화자 적응 방식은 T 개의 잘 훈련된 speaker dependent(SD)모델을 벡터로 구성한다. 즉, T 개의 화자모델은 각각 차원 S의 벡터를 가지게 된다. 이러한 벡터를 "supervector"라고 하는데, HMM 파라미터가 supervector에 저장되는 순서는 상관없지만, T 개의 supervector가 저장되는 순서는 같아야 한다. 그리고 나서, Principal Component Analysis(PCA) 등과 같은 방법을 적용해서 차원 S를 가지는 T 개의 "eigenvoice"를 얻을 수 있다. 최초 몇 개의 eigenvoice들이 주어진 데이터가 가진 변동의 대부분을 설명하기 때문에 T의 eigenvoice 중 최초의 K 개, 즉, e(*1*),…, e(K)만으로 전체 변동을 대표할 수 있다($K < T << S$). 이와 같이 선택된 K 개의 eigenvoice는 "K-space"를 생성한다. 적응데이터가 주어지면 새로운 화자는 다음 식과 같이 K개의 eigenvoice로 나타낼 수 있다.

$$\mu = e(0) + \sum_{k=1}^{K} w(k)\, e(k) \qquad (1)$$

여기서 $e(0)$는 T 명의 SD 모델의 평균을 나타낸다. 그리고 가중치 $w(k)$는 MLED(Maximum Likelihood Eigen-Decomposition)라는 EM 알고리즘을 통해 모델을 추정한다.

III. 잡음 환경에서의 고속화자 적응

3.1 화자 및 환경 동시 적응 [5]

앞서 언급한 eigenvoice 방법의 경우 아주 적은 적응 데이터로 강인한 모델추정이 가능한 고속화자적응 방식으로 적합하다. 그러나 eigenvoice에 의해 적응된 모델은 화자공간에서 환경이 다른 새로운 화자에 대한 정확한 위치를 제공하지 못하기 때문에 환경이 달라진 경우에 이를 보상할 수 있는 eigenvoice기반 적응 방식이 필요하다. 이를 위하여 새로운 화자에 대한 음성은 화자의 특성을 적응하기 위해 eigenvoice 및 환경을 보상하기 위해 보상 벡터의 basis vector의 가중 합으로 나타낼 수 있으며 식(1)에서 상태 s, mixture m인 평균 부벡터는 다음과 같이 확장될 수 있다.

$$\hat{\boldsymbol{\mu}}_m^{(s)} = \mathbf{e}(0)_m^{(s)} + \sum_{j=1}^{K} w(j)\mathbf{e}(j)_m^{(s)} + \sum_{d=1}^{D} b(d)\mathbf{I}(d) \qquad (2)$$

여기서, $\mathbf{I}(d) = [\delta(d-1),\cdots,\delta(d-D)]^T$ 이며, $\delta(x)$는 delta 함수를 뜻한다. 따라서 eigenvoice 수가 D개만큼 증가하는 형태가 되며, 이들의 가중치는 MLED를 통하여 동시에 추정할 수 있게 된다.

3.2 환경 군집화에 기반한 화자적응

환경 변화에 따른 보상을 위해 3.1절에서는 보상 벡터의 basis vector의 가중 합으로 나타내었다. 그러나 다양한 잡음 환경의 데이터가 존재하는 경우 3.1절의 방식으로는 각 잡음별로 보상을 해 주는 것이 어렵다. 따라서 환경별 보상을 위해 유사한 환경별로 적응 데이터를 분류하고, 테스트시 유사한 환경 모델을 선택해서 인식하는 것이 필요하다. 이를 위해 본 절에서는 환경 군집화에 기반한 화자적응 방식을 제안하였다.

그림 1은 잡음 환경 하에서 화자적응의 성능을 향상시키기 위한 방법으로 본 논문에서 제안한 환경 군집화에 기반을 둔 화자적응 시스템을 나타낸다. 그림에서 보면 우선 크게 off-line과 on-line 두 가지 단계로 나눌 수 있다. Off-line 단계에서는 clean DB를 이용해 SI(Speaker Independent) Model을 만들며, noisy adaptation DB를 이용해 화자적응을 거친 후 SA(Speaker Adapted) Model을 만들게 된다. 여기서 noisy adaptation DB를 Classify Environment 모듈을 통해 유사한 잡음 데이터들로 분류할 수 있고, 분류된 adaptation DB를 이용해 앞서 언급한 환경별 SA Model을 만들게 된다. 실험에서 환경 분류를 위해 사용된 특징벡터로는 기존의 채널왜곡을 보상하기 위해 널리 사용되는 CMS의 cepstral mean 값을 사용하

였고, 분류기로는 VQ(Vector Quantization)을 사용하

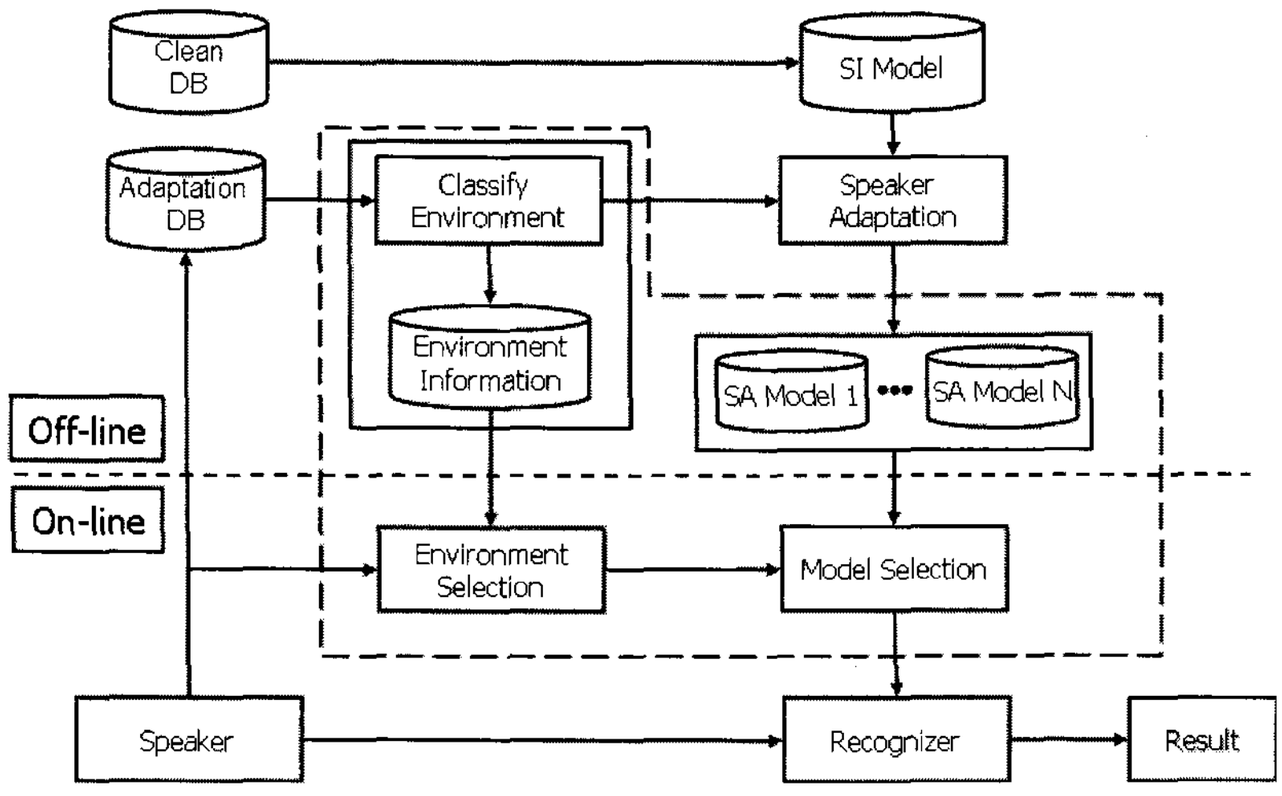

그림 1. 환경 군집화에 기반한 고속화자 적응

였다. 이를 통해 만들어진 코드북을 Environment Information에 저장한다. On-line 단계에서는 테스트 데이터와 환경 정보를 가지고 있는 코드북과의 비교를 통해 환경선택을 하게 되고, 선택된 환경 모델을 이용해서 테스트를 하게 된다.

Ⅳ. 인식실험 및 결과

4.1 실험환경

본 논문에서 사용한 음성 특징 파라미터로는 20ms Hamming window를 10ms씩 shift시키면서 12차 MFCC, delta 및 delta-delta를 구하여 총 36차의 파라미터를 사용하였다. 그리고 46 유사음소(PLU) set[4]을 기본으로 tree-based clustering(TBC)를 사용한 triphone을 기본 모델로 사용하였으며 모델 당 상태수는 3개로 정하였다.

가변어휘 음성 인식기의 훈련을 위해서는 음운 현상이 다양하게 포함된 데이터베이스를 사용하여야 우수한 성능을 얻을 수 있다. 본 실험에서는 훈련을 위하여 ETRI에서 구축한 3음소열 최적화 단어 (Phonetically Optimized Words, POW)[6] 음성 데이터베이스 중에서 남성 40명분의 음성 데이터베이스를 이용해서 모델을 훈련시켰다.

그리고 화자적응 및 인식 실험을 위해서는 훈련용 POW 3,848 음성 데이터베이스와는 어휘 내용이 다른 452 균일 음소 분포 단어 (Phonetically Balanced Words, PBW) 데이터베이스의 일부를 사용하였다. 남성화자 10명의 1회 발성분에 대해서 처음 50개 단어를 성능 평가에 사용하였다.

Eigenvoice를 생성시키기 위해 먼저 POW DB를 사용하여 SI 모델을 구성한 후 40명의 각각의 화자에 대해 MAP 적응 방식을 사용하여 40개의 SD 모델을 구성하였다. 각각을 supervector로 만든 후 PCA를 통하여 40개의 eigenvoice를 구성하였다. 본 논문에서 사용한 tied state 수는 4050개이다.

그리고 잡음 데이터를 생성하기 위해서 Aurora 2 DB 제작에서 사용한 방법과 동일하게 버블 잡음, 자동차 잡음, 그리고 로봇 잡음을 [7]에서 제시한 방법으로 20dB, 10dB가 되도록 부가하여 잡음 환경에서 제안한 방법의 성능을 평가하였다. 사용된 잡음 데이터로 버블 잡음은 NOISEX 92의 것을 사용하였으며, 자동차 잡음은 SiTEC 자동차 잡음 DB(CarNoiseDB01)의 일부로, 콘크리트로 포장된 고속도로를 100km/h로 주행하는 자동차에서 녹음한 것이다. 그리고 로봇 잡음은 실험실에서 데모용 로봇에서 직접 녹음한 것을 사용하였다.

4.2 실험결과

그림 2와 그림 3에 잡음 환경에서의 고속화자 적응 실험 결과를 나타내고 있다. 각각의 그림에서 "SI"는 clean 데이터를 사용한 모델로 인식 실험을 한 경우이고, "EV"는 eigenvoice 적응 방식을 이용한 방식이다. 그리고 "EV+Bias"는 3.1절에서 언급한 바이어스 보상을 통한 환경 화자 동시 적응 방식이며, "EV+Clustering"은 본 논문에서 제안한 환경 군집화에 기반한 적응 방식으로 환경 분리를 통한 뒤 화자적응을 적용한 경우이다.

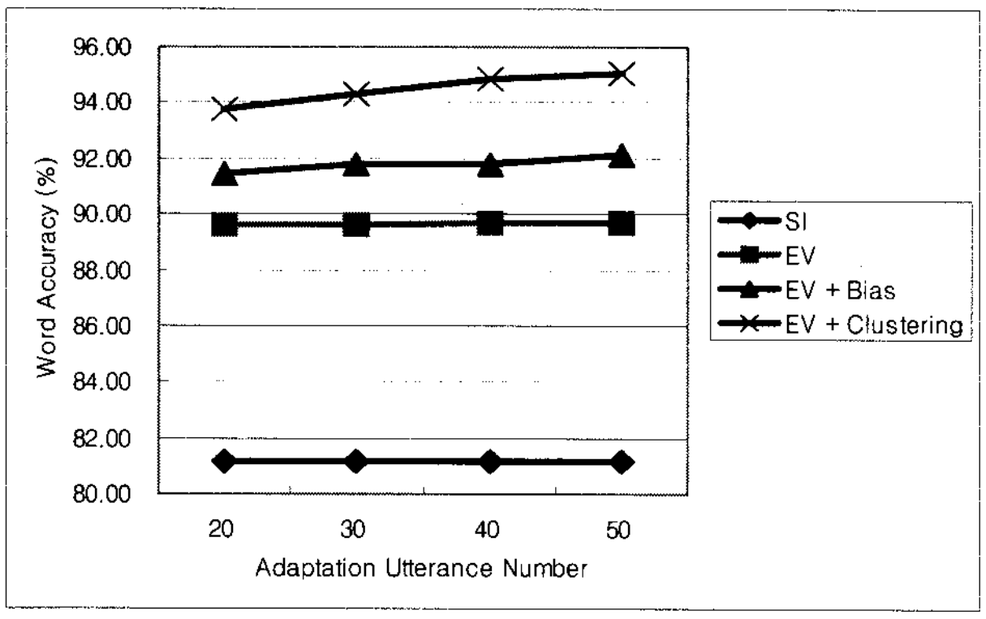

그림 2. SNR 20dB인 버블, 자동차, 그리고 로봇 잡음에 대한 고속화자 적응 실험

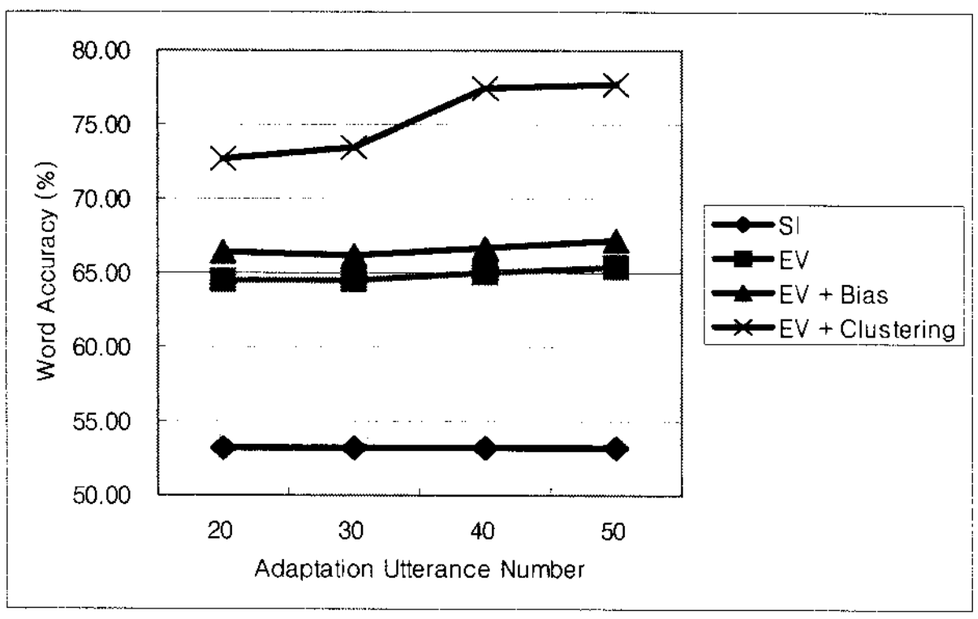

그림 3. SNR 10dB인 버블, 자동차, 그리고 로봇 잡음에 대한 고속화자 적응 실험

우선 그림 2는 화자 적응 및 테스트 데이터가 SNR 20dB인 버블, 지동차, 그리고 로봇 잡음이 동일 비율로 있는 경우에 대한 실험 결과이다. 그림에서 보면 "SI"의 경우 화자적응을 하기 전의 baseline 성능으로 81.8%의 단어 인식률을 나타낸다. "EV", "EV+Bias", 그리고 "EV+Clustering"의 성능은 모두 "SI"의 성능보다 우수한 것을 알 수 있으며, 적응 데이터의 수가 20에서 50까지 증가하면서 성능이 향상됨을 알 수 있다. 그리고 "EV+Bias"가 "EV"보다 우수한 성능을 얻었다. 이것은 bias 보상을 통해 잡음 환경에 대한 보상이 이루어 졌기 때문이라고 판단된다. "EV+Clustering"의 경우 여러 방식 중에서 가장 우수한 성능을 얻었고 "EV+Bias" 방식보다 각각의 적응 데이터 수에 따라 27.4%-37.4%의 오류율 감소를 얻었다.

그림 3은 SNR 10dB이고 "SI"성능이 53.3%인 것을 제외하고 그림 2와 동일한 환경의 실험이다. 실험결과 그림 2의 경우와 같이 "EV", "EV+Bias", 그리고 "EV+Clustering" 순으로 성능이 좋았으며 적응 데이터 수에 따라서 18.8%-32.4%의 오류율 감소율을 얻었다.

V. 결론

본 논문에서는 다양한 잡음 환경 하에서 화자적응을 통해 인식성능 향상을 얻을 수 있는 방식에 대해 연구하였다. 이를 위해 다양한 잡음 환경을 분리하기 위한 방법으로 환경 군집화 방법을 이용해 유사한 환경별로 군집화를 하였고, 군집화된 환경별로 SA 모델을 생성하여 화자 적응을 하는 방식을 제안하였다. 실험결과 제안한 방식이 기존의 eigenvoice를 보완한 화자 환경 동시 보상 방식보다 적응 데이터 수에 따라서 19%-37%의 오류율 감소를 얻었다. 향후 연구 과제로 본 논문에서는 화자 적응과 테스트에 사용된 잡음의 경우 동일한 환경들로 구성하였으나 서로 다른 잡음에 대한 추가적인 실험이 필요하다.

참고문헌

[1] A. Sanker, "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. Speech and Audio Processing, vol.4, no.3, pp.190 -202, May, 1996.

[2] 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, no.4, pp.806-814, April, 1991.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, no.1, pp.171-185, September, 1995.

[4] R. Kuhn, P. Nguyen, J. C. Jungua, L. Goldwasser, N. Niedzielski, S. Finche, K. Field and M. Contolini, "Eigenvoices for speaker adaptation," in Proc. ICSLP, vol.5, pp.1771-1774, 1998.

[5] H. J. Song and H. S. Kim, " Simultaneous Estimation of Weights of Eigenvoices and Bias Compensation Vector for Rapid Speaker Adaptation," ICSLP, October, 2004.

[6] Y. Lim and Y. Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," In Proc. ICASSP, vol.1, pp.89-91, 1995.

[7] ITU recommendation P.56, "Objective measurement of active speech level," March, 1993