말소리와 음성과학 제7권 제3호 (2015.09.30)
ISSN 2005-8063, pp. 131~138

ksss.jams.or.kr
http://dx.doi.org/10.13064/KSSS.2015.7.3.131

L1-norm regularization을 통한 SGMM의 state vector 적응

L1-norm Regularization for State Vector Adaptation of Subspace Gaussian Mixture Model

구 자 현[1] · 김 영 관[2] · 김 회 린[3]
Goo, Jahyun · Kim, Younggwan · Kim, Hoirin

ABSTRACT

In this paper, we propose L1-norm regularization for state vector adaptation of subspace Gaussian mixture model (SGMM). When you design a speaker adaptation system with GMM-HMM acoustic model, MAP is the most typical technique to be considered. However, in MAP adaptation procedure, large number of parameters should be updated simultaneously. We can adopt sparse adaptation such as L1-norm regularization or sparse MAP to cope with that, but the performance of sparse adaptation is not good as MAP adaptation. However, SGMM does not suffer a lot from sparse adaptation as GMM-HMM because each Gaussian mean vector in SGMM is defined as a weighted sum of basis vectors, which is much robust to the fluctuation of parameters. Since there are only a few adaptation techniques appropriate for SGMM, our proposed method could be powerful especially when the number of adaptation data is limited. Experimental results show that error reduction rate of the proposed method is better than the result of MAP adaptation of SGMM, even with small adaptation data.

Keywords: L1-norm regularization, speaker adaptation, state vector adaptation, subspace gaussian mixture model, automatic speech recognition

1. 개요

음성인식 시스템의 성능 저하를 불러오는 주된 원인으로는 화자간 차이를 꼽을 수 있는데, 이는 주로 강세나 빠르기와 같은 각 음소 발음 방법의 차이로 드러난다 [1]. 기존에 널리 사용되던 음향모델링 기법인 GMM-HMM에는 이미 이러한 화자간 차이를 보상하기 위한 다양한 방법이 존재한다. 우선 고전적 적응 기법인 MAP (Maximum A Posteriori)이나 MLLR (Maximum Likelihood Linear Regression)은 GMM 내 모든 변수를 적응시킬 수 있으나, 대부분의 성능 향상은 가우시안 평균 벡터의 적응에서 드러나는 편이다 [2]. 또한 CAT (Cluster Adaptive Training) [3]나 Eigenvoice [4]는 군집(cluster) 기반의 적응 기법으로, 화자간 변이를 모델링하는 군집 모델을 만들어 두고 그들의 선형 결합을 이용하여 가우시안 평균 벡터를 적응하게 된다. 이들 적응 기법을 잘 살펴보면, GMM-HMM에서 화자 간 차이를 보상할 때는 가우시안 평균 벡터를 적응하는 방법을 주로 활용한다. 하지만 이 때 가우시안 평균 벡터는 음소에 관한 정보 뿐 아니라 음높이나 운율과 같은 준언어적(paralinguistic) 정보 또한 포함하고 있고 기존 화자적응 기법 역시 이들 모두를 다 같이 보정하는 것을 목표하고 있다. 만일 음향 모델링시 음소 정보를 보다 효율적으로 모델링할 수 있다면, 해당 음소 정보를 보정하는 것을 통해 화자 간 차이에서 특히 음소 발음 방법의 차이로 나타나는 부분을 더욱 잘 보상할 수 있을 것이다.

한편 2010년 Povey 등은 GMM을 약간 변형한 모델인 SGMM (Subspace Gaussian Mixture Model)을 제안하였다 [5], [6]. 이 모델은 본디 다국어 데이터베이스를 이용하여 목표 언어의 음성인식 성능을 끌어올리기 위해 제안되었는데, 해당 모델에서 가우시안 평균 벡터는 모든 HMM state가 공유하는 변수와 각 state를 대표하는 변수로 분해(decompose)된다. 이 경우 음향모델 학습시 각 state를 대표하는 변수는 음소에 직접적

1) KAIST, jahyun.goo@kaist.ac.kr
2) KAIST, cleanthink@kaist.ac.kr
3) KAIST, hoirkim@kaist.ac.kr, 교신저자

접수일자: 2015년 8월 4일
수정일자: 2015년 9월 4일
게재결정: 2015년 9월 21일

으로 대응하므로 목표하는 언어로만 훈련하고, 대신 모든 state가 공유하는 변수는 녹음환경이 같은 다국어 데이터로 훈련하는 방식을 생각할 수 있고, 실제로 이런 방식으로 성능 향상을 기록한 바 있다 [6], [7].

이 때 모든 state가 공유하는 변수는 projection matrix라 불리며, 기하학적으로는 특징벡터가 위치할 수 있는 특정한 부분공간(subspace)을 정의한다. 각 state를 대표하는 변수는 state vector라 불리고, 해당 부분공간 내에서 특정 HMM state와 대응하는 곳의 좌표(coordinate)로써 정의된다. 이 때문에 SGMM 음향모델을 채용할 경우 가우시안 평균 벡터가 지닌 다양한 정보 중에서 음소 정보는 state vector를 통해 주로 모델링되고, 이는 화자적응을 수행하는 과정에서 각 음소 발음 차이를 보상할 때 state vector를 적응시킨다는 생각으로 자연스레 생각할 수 있다.

SGMM 음향모델에서 화자적응을 수행하는 방법은 기존에 다음과 같이 제안된 바 있다. 우선 Povey 등은 SGMM 음향모델을 제안할 당시 화자부분공간(speaker subspace) 확장을 함께 제안하였는데, 이는 Eigenvoice와 유사한 화자종속(speaker dependent) 오프셋을 가우시안 평균 벡터에 더하는 방식이다 [6]. 그러나 이 경우 화자종속 오프셋을 사용하기 위해 미리 화자 부분공간을 훈련시켜야 하는 어려움이 있고, 또한 이 접근방법은 각 음소 발음의 차이를 보상하고자 하는 본 논문과는 다소 차이가 있다. 한편 Lu 등은 다국어 데이터베이스를 통한 음성인식의 연장선상에서 projection matrix에 대한 MAP 적응을 제안했다 [8]. 그 기법은 화자적응에 적용할 경우 성능은 좋으나 역시 음소 발음을 주로 보상하고자 하는 본 논문과는 접근 방법 측면에서 다르다고 할 수 있다. Hamidi G. 등은 선형회귀(linear regression)를 통한 state vector의 적응을 제안했는데 [9], 이 방법은 효과적인 적응을 위해 경험적으로 정해야 하는 설정이나 변수가 너무 많아 좋은 성능을 끌어내기가 쉽지 않다.

본 논문에서는 SGMM 음향모델에서 화자적응을 위해 state vector를 보상하는 MAP 기반 적응을 두 가지 제안한다. 우선 state vector에 일반적인 MAP 적응을 적용하는 방법을 논하고, 그 후 L1-norm regularization을 통해 sparse solution을 추정하는 방법을 제안할 것이다. 일반적인 MAP 적응이 적응 과정에서 기존 음향모델과 같은 양의 파라미터를 요구하는데, 이는 요즘의 대규모 다중 사용자 음성인식 환경에서 화자적응 시스템을 구현하는 데 큰 제약이 될 수 있다 [10]. 여기서 BPDN (Basis Pursuit De-noising) [11] 혹은 LASSO [12]로 불리는 L1-norm regularization을 통한 추정은 sparse solution을 얻도록 하는데, 이는 일반적인 MAP 적응에 비해 조정하는 파라미터의 수를 상당량 줄이면서도 거의 유사한 수준의 성능을 보장할 수 있도록 한다. 또한 적응 데이터의 양이 적어질수록 L1-norm regularization을 통한 추정이 더 좋은 성능을 보이기도 한다.

본 논문은 다음과 같이 구성된다. 2장에서는 SGMM 음향모델에 대한 기본적인 설명과 state vector에 MAP 적응을 수행하는 방법에 대해 설명한다. 3장에서는 그 연장선상에서 모델 파라미터 적응에 사용된 L1-norm regularization에 대해 보다 구체적으로 밝히고, 4장에서 실험을 통한 성능 검증을 수행한 후 5장에서 결론을 맺고 논문을 마무리한다.

2. SGMM 음향모델 및 MAP 적응

1장에서 언급했듯이, SGMM은 GMM-HMM의 일종이다. 따라서 각 state의 우도(likelihood)는 GMM-HMM 음향 모델과 같이 다차원 가우시안 혼합(Gaussian mixture)으로 주어지지만, 각 가우시안의 평균 벡터는 대응하는 state vector와 projection matrix 쌍으로부터 유도되는 특징이 있다. 거기에 더해 SGMM에서는 GMM-HMM모델과 달리 공분산행렬이 공유되고, 실제 사용에서는 각 state에 속한 substate가 가우시안 혼합으로 구성되며 그 substate들의 weighted sum이 매 state를 이루게 되는 것 또한 SGMM의 특징이다. 각 state에서의 우도 수식은 아래와 같이 주어진다.

$$p(\mathbf{x}_t \mid j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} N(\mathbf{x}_t; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i) \tag{1}$$

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} \tag{2}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}} \tag{3}$$

여기서 j,m,i는 각각 상태, 부분상태, 가우시안 성분 지수(index)를 가리킨다. $\mathbf{v}_{jm}$은 각 부분상태에 대응하는 substate vector이고, $\mathbf{M}_i$와 $\mathbf{w}_i$는 projection matrix, projection vector로 불리며 HMM상태와 무관한 전역변수로서 substate vector와 곱해 가우시안 평균 벡터와 가중치를 각각 만들어내게 된다. 위에서 언급했듯 식 (1)에서 각 상태 i에 대해 M_i개의 부분상태가 있는 것을 확인할 수 있다.

substate vector가 S차원, 특징벡터는 D차원 벡터라고 할 경우 위 식은 $\mathbf{M}_i$의 D차원 열벡터 S개를 $\mathbf{v}_{jm}$ 성분에 따라 적절히 가중합(weighted sum)하여 가우시안 평균 $\boldsymbol{\mu}_{jmi}$를 만든다고 볼 수 있고, 이 구조는 $\mathbf{M}_i$의 열벡터로 정의된 부분공간을 통해 좌표벡터 $\mathbf{v}_{jm}$를 가우시안 평균으로 투사(projection) 혹은 사상(mapping)하는 것으로 볼 수 있다. $\mathbf{M}_i$을 구하는 과정에서 LDA (Linear Discriminant Analysis)와 유사한 과정을 거치기 때문에 substate vector는 음소 클래스의 관점에서 잘 분리되고, 동일 state에 속한 특징벡터의 중심(centroid)인 가우시안 평균에 비해 각 음소를 더 잘 대표한다.

LDA는 클래스로 구분되는 데이터를 포함하는 공간에서 클래스내 분산을 작게 하면서 클래스간 분산을 되도록 크게 하는 축(axis)을 잡아 데이터를 새로 표현하는 일종의 패턴인식 방법론으로, 해당 축을 이용해 새로운 좌표로 데이터를 표현할 경우 클래스간 구분을 훨씬 용이하게 할 수 있다. 음성인식 과정에서 음향모델이 수행하는 역할을 개략적으로 표현하면 각 특징벡터의 음소 state 매칭이므로, 클래스가 잘 구분되도록 하는 기저벡터(basis vector)를 축으로 삼아 특징벡터가 속한 공간[4]을 표현한다면 음향모델의 전반적인 성능에 영향을 수 있을 것이다. 식 (2)에서는 $\mathbf{M}_{\mathrm{j}}$을 구성하는 D차원 열벡터 S개가 매 state에 대한 가우시안 평균들을 보다 잘 구분하는 기저벡터가 되고, 결과적으로 state vector는 feature space에서의 가우시안 평균보다 각 state를 더 잘 대표하는 값이 된다[5].

관련하여 Burget 등 또한 비슷한 것을 언급했는데, 그들은 [7]에서 모음에 대응하는 state vector들 클래스 간에 잘 분리되어 있으며, 그 분포가 모음사각도와 유사함을 보인 바 있다.

이 SGMM substate vector는 E-M 알고리즘을 통해 업데이트되고, 그 수식은 아래와 같다 [6].

$$Q_{\mathrm{MLE}}(\mathbf{v}_{\mathrm{jm}}) = \mathbf{g}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{v}_{\mathrm{jm}} - \frac{1}{2}\mathbf{v}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{H}_{\mathrm{jm}}\mathbf{v}_{\mathrm{jm}} \tag{4}$$

$$\hat{\mathbf{v}}_{jm}^{\mathrm{MLE}} = \mathrm{arg\,max}_{\mathbf{v}_{\mathrm{jm}}} Q_{\mathrm{MLE}}(\mathbf{v}_{\mathrm{jm}}) \tag{5}$$

$$= \mathbf{H}_{\mathrm{jm}}^{-1}\mathbf{g}_{\mathrm{jm}} \tag{6}$$

여기서 식 (4)는 수렴성을 위해 2차 근사를 적용한 것으로 substate vector의 업데이트 수식은 식 (6)과 같고, $\mathbf{g}_{jm}$, $\mathbf{H}_{jm}$은 E-M 알고리즘의 E-step에서 축적하는 중간 변수[6]이다 [13].

이러한 ML (Maximum Likelihood) 기반 수식을 화자적응에 바로 이용하는 것도 가능하나, 이는 특정 화자의 데이터로 재훈련하는 것으로 대부분의 경우 데이터 수에 비해 훈련할 변수가 많아 쉽게 과적합(over-fitting)에 빠질 우려가 있다. 따라서 그를 보완하기 위해 substate vector의 분포가 가우시안 분포를 이룬다고 가정하고 아래와 같이 MAP 적응을 구성할 수 있다.

$$\log P(\mathbf{v}_{\mathrm{jm}}) = \mathrm{c} + \bar{\mathbf{v}}_{\mathrm{jm}}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{jm}}^{-1}\mathbf{v}_{\mathrm{jm}} - \frac{1}{2}\mathbf{v}_{\mathrm{jm}}^{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{jm}}^{-1}\mathbf{v}_{\mathrm{jm}} \tag{7}$$

$$Q_{\mathrm{MAP}}(\mathbf{v}_{jm}) = Q_{\mathrm{MLE}}(\mathbf{v}_{jm}) + \tau \log P(\mathbf{v}_{jm}) \tag{8}$$

$$\hat{\mathbf{v}}_{jm}^{\mathrm{MAP}} = \mathrm{arg\,max}_{\mathbf{v}_{\mathrm{jm}}} Q_{\mathrm{MAP}}(\mathbf{v}_{\mathrm{jm}}) \tag{9}$$

$$= (\mathbf{H}_{\mathrm{jm}} + \tau\boldsymbol{\Omega}_{\mathrm{jm}}^{-1})^{-1}(\mathbf{g}_{\mathrm{jm}} + \tau\boldsymbol{\Omega}_{\mathrm{jm}}^{-1}\bar{\mathbf{v}}_{\mathrm{jm}}) \tag{10}$$

여기서 $\bar{\mathbf{v}}_{jm}$은 사전확률분포의 평균이며 substate vector의 현재 값을 나타내고, $\boldsymbol{\Omega}_{jm}$은 사전확률분포의 공분산을 가리킨다. 사전확률분포 변수는 같은 state에 속한 substate vector들이 한 가우시안 분포를 이룬다고 가정하여 구했다. 업데이트 수식은 ML 기반 업데이트와 유사하기 때문에 기존 시스템의 적은 개선을 통해 손쉽게 구현할 수 있다.

이 SGMM의 MAP 적응은 기존 GMM-HMM에서의 MAP 적응과 많은 공통점을 가지는데, 우선 업데이트 수식인 식 (10)이 분수의 형태를 하고 있으며 분모와 분자 모두 화자독립변수와 화자종속변수의 선형결합 꼴을 하고 있다는 점이 그러하다. 조절 hyperparameter τ의 역할 또한 유사하여 $\tau = 0$일 때는 MAP 추정치가 화자종속모델 (ML 추정치)과 같아지며 τ가 커질수록 MAP 추정치가 화자독립모델에 가까운 값을 갖게 된다.

사전확률의 공분산 $\boldsymbol{\Omega}_{jm}$으로는 전행렬 (full matrix)을 사용하나 복잡도를 줄일 목적으로 대각행렬 (diagonal matrix)과 단위행렬의 사용을 생각해볼 수 있다. 특히 단위행렬을 사용하는 경우는 사전확률분포의 분산을 추정하지 않는 것으로, 이는 같은 state에 속한 substate vector들이 동일한 분포에서 생성되었다는 불필요한 가정을 하지 않는 것일뿐더러 MAP 적응에 있어 직전 substate vector만 사용하게 되는 것으로 이 경우 추정은 변수 변화량에 대한 L2-norm 기반 penalty를 식 (5)에 도입한 것과 같아진다.

$$Q_{\mathrm{L2}}(\mathbf{v}_{jm}) = Q_{\mathrm{MLE}}(\mathbf{v}_{jm}) - \frac{\tau}{2}\|\mathbf{v}_{jm} - \bar{\mathbf{v}}_{jm}\|_2^2 \tag{11}$$

$$= Q_{\mathrm{MLE}}(\mathbf{v}_{\mathrm{jm}}) + \tau(\bar{\mathbf{v}}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{v}_{\mathrm{jm}} - \frac{1}{2}\mathbf{v}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{v}_{\mathrm{jm}}) \tag{12}$$

이러한 penalty 도입은 기하학적으로 ML 추정치와 기존 변수의 L2-norm 기반 거리를 제한하는 효과를 주어 과적합을 막게 된다. 그러나 L2-norm 기반 제약항을 사용할 경우 추정치에 크기가 0에 가까운 수치가 많이 만들어지며, 이는 1장에 언급한 것처럼 적응 효과에 비해 많은 변수를 적응시켜야 하는 단점을 낳는다. 따라서 이를 해결하기 위해 penalty로 L1-norm을 사용하여 부분공간벡터를 sparse adaptation하는 것을 고려해볼 수 있다.

3. L1-norm regularization

L1-norm regularization은 sparse solution을 추정하는 방법 중 하나이다. 음향모델 적응에 sparse adaptation을 사용하는 것은

4) Feature space. 가우시안 평균 벡터가 속한 공간과 같다.

5) 이를 앞서 1절에서는 "state vector가 특정 부분공간의 좌표가 된다"고 표현했고, 바로 앞단락에서는 $\mathbf{M}_{\mathrm{j}}$의 열벡터에 대한 가중치(weight)로 언급하였다.

6) 값은 각각 다음과 같다. (γ는 posterior count를 나타낸다.)
$\mathbf{g}_{jm} = \sum_{t,i}\gamma_{jmi}(t)\mathbf{M}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x}(t) + \sum_i \mathbf{w}_i(\gamma_{jmi} - \gamma_{jm}w_{jmi} + \gamma_{jm}w_{jmi}\mathbf{w}_i^{\mathrm{T}}\mathbf{v}_{jm})$
$\mathbf{H}_{jm} = \sum_i(\gamma_{jmi}\mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i + \gamma_{jm}w_{jmi}\mathbf{w}_i\mathbf{w}_i^T)$

GMM-HMM 음향모델에 대해서는 기존에 Olsen 등이 가우시안 평균 벡터의 MAP 적응에서 L0-norm 제약항을 사용해 sparsity를 조절하는 방식을 제안한 바 있고 [14], Kim 등이 L1-norm regularization을 적용한 화자적응을 제안하기도 했다 [10]. 본 논문에서는 L1-norm regularization을 SGMM substate vector 적응에 사용하기로 한다. 그를 위해 조함수 (11)에서 L2-norm을 L1-norm으로 바꾸면 아래와 같다.

$$Q_{\mathrm{L1}}(\mathbf{v}_{\mathrm{jm}}) = \mathrm{Q}_{\mathrm{MLE}}(\mathbf{v}_{\mathrm{jm}}) - \lambda\|\mathbf{v}_{\mathrm{jm}} - \bar{\mathbf{v}}_{\mathrm{jm}}\|_1 \tag{13}$$

적응으로 인해 변하는 값을 $\mathbf{d}_{jm} \triangleq \mathbf{v}_{jm}^{\mathrm{L1}} - \bar{\mathbf{v}}_{jm}$으로 대치하여 L1-norm 항이 하나의 변수만을 갖도록 하면 조함수는 다음과 같은 표준적인 L1-norm regularization 문제가 된다.

$$Q_{\mathrm{L1}}(\mathbf{d}_{\mathrm{jm}}) = (\mathbf{g}_{\mathrm{jm}} - \mathbf{H}_{\mathrm{jm}}\bar{\mathbf{v}}_{\mathrm{jm}})^{\mathrm{T}}\mathbf{d}_{\mathrm{jm}} - \frac{1}{2}\mathbf{d}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{H}_{\mathrm{jm}}\mathbf{d}_{\mathrm{jm}} - \lambda\|\mathbf{d}_{\mathrm{jm}}\|_1 \tag{14}$$

$$\hat{\mathbf{v}}_{jm}^{\mathrm{L1}} = \bar{\mathbf{v}}_{jm} + \hat{\mathbf{d}}_{jm}^{\mathrm{L1}} \tag{15}$$

L1-norm regularization 문제를 풀기 위해 본 논문에서는 Figueiredo 등이 [17]에서 제안한 GPSR (Gradient Projection for Sparse Reconstruction) 알고리즘을 사용하기로 한다. 해당 알고리즘은 앞서 Lu 등이 SGMM의 sparse estimation에 활용한 방법이기도 하다 [15], [16]. 알고리즘을 사용하기 위해서는 우선 아래와 같은 과정을 통해 L1-norm regularization 문제를 보다 간단한 문제로 바꾸는 과정을 거친다.

$\mathbf{x}_{\mathrm{jm}}, \mathbf{y}_{\mathrm{jm}}$을 각각 $\mathbf{d}_{\mathrm{jm}}$의 양수, 음수 부분의 절댓값을 취한 nonnegative 벡터로 정의하면 $\mathbf{d}_{\mathrm{jm}} = \mathbf{x}_{\mathrm{jm}} - \mathbf{y}_{\mathrm{jm}}$으로 나타낼 수 있고 $\mathbf{z}_{\mathrm{jm}} = (\mathbf{x}_{\mathrm{jm}}^{\mathrm{T}}, \mathbf{y}_{\mathrm{jm}}^{\mathrm{T}})^{\mathrm{T}}$라 하면 식 (14)를 아래와 같은 constrained quadratic minimization 문제로 쓸 수 있다.

$$\hat{\mathbf{z}}_{jm} = \mathrm{argmax}_{\mathbf{z}_{\mathrm{jm}}} \mathbf{c}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{z}_{\mathrm{jm}} + \frac{1}{2}\mathbf{z}_{\mathrm{jm}}^{\mathrm{T}}\mathbf{B}_{\mathrm{jm}}\mathbf{z}_{\mathrm{jm}} \tag{16}$$

$$\mathbf{c}_{\mathrm{jm}} = \lambda\mathbf{1}_{2\mathrm{S}} + \begin{pmatrix} -(\mathbf{g}_{\mathrm{jm}} - \mathbf{H}_{\mathrm{jm}}\bar{\mathbf{v}}_{\mathrm{jm}}) \\ \mathbf{g}_{\mathrm{jm}} - \mathbf{H}_{\mathrm{jm}}\bar{\mathbf{v}}_{\mathrm{jm}} \end{pmatrix} \tag{17}$$

$$\mathbf{B}_{\mathrm{jm}} = \begin{pmatrix} \mathbf{H}_{\mathrm{jm}} & -\mathbf{H}_{\mathrm{jm}} \\ -\mathbf{H}_{\mathrm{jm}} & \mathbf{H}_{\mathrm{jm}} \end{pmatrix} \tag{18}$$

이 때 $\mathbf{z}_{\mathrm{jm}}$의 모든 항은 0보다 크거나 같기 때문에 nonnegative 제약이 걸린 경사도 상승(gradient ascent) 방법을 적용하여 문제를 해결할 수 있고, $\mathbf{z}_{jm}$을 정의하면서 차원은 두 배가 되지만 수식에서는 L1-norm 항에서 비롯되는 비선형성이 제거되어 조금 더 수월하게 문제를 풀 수 있게 된다.

MAP 적응의 hyperparameter τ와는 달리 L1-norm regularization의 λ를 통해서는 SGMM 변수의 적응 정도뿐 아니라 sparsity를 동시에 조절할 수 있다. 따라서 L1-norm regularization을 이용하면 훨씬 적은 수의 변수를 적응시키더라도 기존의 MAP 적응에 비견할 만한 좋은 화자적응 성능을 얻을 수 있다. 이에 대해서는 4장에서 더 자세히 살펴보기로 한다.

반면에 L1-norm regularization은 MAP 적응을 약간 개선한 것이기 때문에 MAP 적응이 가지는 단점을 똑같이 가지고 있다. 즉, 두 적응 기법은 모두 각 HMM state를 독립적으로 조정하기 때문에 적응 데이터에서 관측하지 못한 음소에 대응하는 state는 적응시키지 못한다는 단점을 갖고 있다.

또한 L1-norm 제약항은 모든 차원을 동등하게 규제하는데 반해, SGMM substate vector의 각 성분의 중요도는 서로 다르다 [13]. 이는 앞서 언급했듯이 projection matrix $\mathbf{M}_i$를 구하는 데 LDA와 유사한 방법이 사용되기 때문으로, state vector를 적응시킬 때 L1-norm 제약항에서 각 차원별 가중치를 다르게 줄 경우 더 나은 성능을 보일 수도 있다. 이는 2장에서 언급했듯 MAP 적응은 유클리드 거리 대신에 보다 일반화된 마할라노비스 거리 (Mahalanobis distance)를 적용한 L2-norm regularization 문제로 볼 수 있고, 이러한 일반화를 L1-norm regularization에도 적용할 수 있는데 이에 대한 개념도가 <그림 1>과 같다.

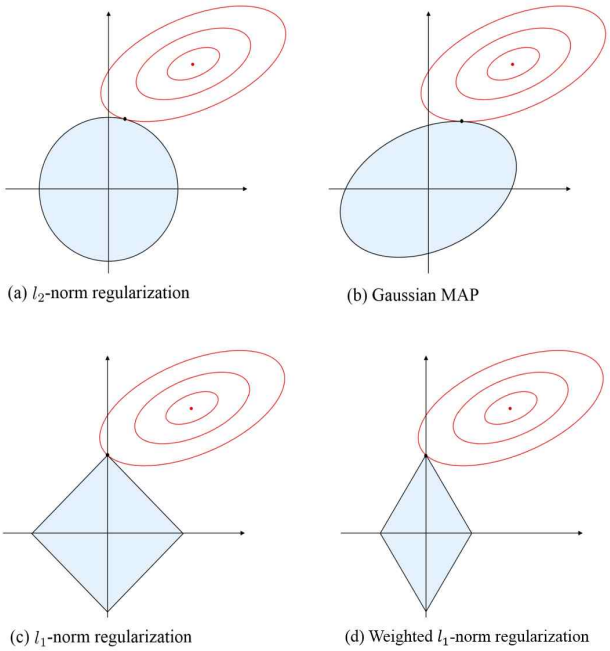

그림 1. 여러 MAP 기반 적응 기법의 기하학적 표현
Figure 1. Geometric interpretation of various MAP-based adaptation

<그림 1>[7]에서는 설명을 위해 2차원 변수의 추정을 예로 들어 각 적응기법을 나타냈다. 예를 들어 식 (11), (13)을

7) 그림의 기본적인 얼개는 [12]를 참조하였다.

constrained optimization 측면에서 보면 식 (4)의 파라미터 추정 과정에서 norm의 크기를 제한하는 constraint를 준 것으로 볼 수 있다. <그림 1>의 각 그래프에서 오른쪽 위 타원은 식 (6)과 같이 적응 이전에 추정된 파라미터 및 그 분산을 나타내고, 원점을 중심으로 하는 도형은 constraint를 가리킨다. 즉, (a)와 (c)는 식 (11), (13)에 대응하며 constraint가 norm ball로 표현되고, 여기서 constraint를 만족하면서 추정된 해는 각 그래프에 나타난 두 도형의 교점이 된다. <그림 1>에서 L1-norm 기반의 constraint를 사용한 경우인 (c)를 보면 (a)에 비해 비교적 크기가 작은 차원의 값이 0이 되는 차이를 볼 수 있다. 이는 앞서 언급했듯 일반적인 unconstrained optimization 항목에 L1-norm panelty를 적용할 경우 sparse solution을 얻을 수 있다고 말한 것과 일치한다.

또한 (b)는 (a)의 constraint가 L2-norm, 다시 말해 유클리드 거리를 제한하는 것을 일반화해 마할라노비스 거리를 제한하는 것으로 개념을 보다 일반화시킨 경우이고, (d) 또한 (c)를 조금 더 일반화시켜 L1-norm constraint에서 차원별 제약 정도를 다르게 준 경우로 볼 수 있다.

이렇게 일반화된 L1-norm 문제는 weighted L1으로도 불리며 선형계획법의 해로 이미 존재하지만, 이는 차원별 가중치를 경험적으로 정해 두어야 하는 단점이 있다. Weighted L1 문제에서 기존의 L1 문제와 같이 hyperparameter를 자동적으로 정하면서 weighted L1을 수행하는 알고리즘이 re-weighted L1으로 제안된 바 있지만 [18], [19], 이는 아직 음향모델링 기법에 적용된 바 없고 본 논문에서도 추후 과제로 두기로 한다.

4. 실험 결과 및 분석

제안한 SGMM의 state vector에 대한 MAP 적응 및 L1-norm regularization에 대한 성능 평가는 다음과 같은 환경에서 수행하였다. 전반적으로 Wall Street Journal 낭독체 영어음성 데이터베이스를 사용했으며 화자독립 모델 훈련에는 7,138 문장, 15시간 분량의 SI-84 subset을 사용하였고, 적응 및 검증에는 Dev93 및 Eval92 두 subset을 사용하였으며 이들은 각각 화자 10명 (총 503문장), 8명 (총 333문장)으로 구성되어 있다. 화자당 데이터 수가 적기 때문에 화자적응 모델을 훈련하고 검증할 때는 검증에 사용하는 2문장을 뺀 나머지를 적응에 사용하는 교차검증(cross-validation)을 활용하였고, 각 화자별로 처음 20문장 분량의 단어 오류율 (WER, word error rate) 평균값을 성능 비교에 사용하였다. 특징은 정적 값에 1차, 2차 차분값을 포함한 39차 MFCC에 CMVN을 적용하여 사용했고, full covariance를 사용하는 SGMM에 비해 diagonal matrix를 covariance로 사용하는 GMM-HMM을 보상하기 위해 해당 모델에 사용할 특징에는 LDA/MLLT를 추가로 적용하였다. 언어모델로는 데이터베이스에서 기본 제공하는 2만단어급 pruned trigram 언어모델을 모든 음향모델에서 동일하게 사용하였다. 실험은 GMM-HMM과 SGMM 모델 각각을 우선 화자독립 데이터베이스를 이용해 훈련하고, 두 화자독립 음향모델에 각각 ML 추정, MAP 적응 및 L1-norm regularization을 적용하였다. 실험은 전반적으로 오픈소스 툴킷인 Kaldi를 이용하여 진행하였다 [20].

표 1. 여러 적응 기법의 단어 오류율 (%)
Table 1. WER(%) of various adaptation approaches

		Dev93	Eval92	평균
[GMM-HMM]	SI	18.00	10.47	14.48
	MLE	15.53	9.48	12.70
	MAP ($\tau=20$)	14.97	9.45	12.39
	L1 ($\lambda=20$)	16.84	9.30	13.32
[SGMM]	SI	14.40	8.93	11.85
	MLE (all)	19.57	18.16	18.91
	MLE ($\mathbf{v}_{jm}$)	14.89	10.22	12.71
	MAP ($\mathbf{v}_{jm}$, iden, $\tau=20$)	12.22	7.62	10.07
	MAP ($\mathbf{v}_{jm}$, diag, $\tau=10$)	13.80	8.57	11.36
	MAP ($\mathbf{v}_{jm}$, full, $\tau=20$)	13.89	8.31	11.28
	L1 ($\mathbf{v}_{jm}$, $\lambda=20$)	12.09	8.09	10.22

<표 1>에 제시한 기본적인 적응 기법은 다음과 같이 적용되었다. GMM-HMM 모델에 대해서는 언급된 적응 기법 모두 가우시안 평균벡터만을 적응시켰다. SGMM 모델에 ML 기반 재훈련을 하는 것은 앞서 소개한 것과 같이 substate vector $\mathbf{v}_{jm}$만을 업데이트하는 것과 더불어 전체 변수를 업데이트하는 것(all로 표시)을 비교해 보았다. SGMM 음향모델에서는 음높이, 억양과 같은 준언어적 정보나 환경과 같은 언어 외적 정보가 주로 projection matrix에 반영된다고 보기 때문에, 이러한 비교를 통해 화자 간 차이가 음소 아닌 부분에 미치는 영향력을 알아볼 수 있다. 또한 MAP 적응의 경우 역시 세 가지를 제시했는데 이는 substate vector의 사전확률분포 공분산행렬 (식 (7)의 $\boldsymbol{\Omega}_{jm}$)을 각각 단위행렬(iden.), 대각행렬(diag.), 전행렬(full)으로 간주하고 계산한 것이다. 각 적응 방법에서 hyperparameter τ 및 λ는 1에서 30까지 약 3 간격으로 차이를 벌려 설정해가며 실험해 가장 낮은 오류율을 기록한 결과를 수록하였다.

<표 1>을 보면 우선 SGMM에서 ML 기반 적응의 성능이 단순히 기존의 화자독립 모델을 인식에 사용한 것보다도 더 좋지 못하다는 것을 볼 수 있다. 이는 SGMM이 substate라는 특수한 구조를 갖고 있는데, ML 기반 적응이 그에 적합하지 않아 음소별로 과적합 혹은 부적합이 일어났기 때문인 것으로 보인다. SGMM의 각 state는 1~13개의 substate로 구성되어 있는데, ML 기반 적응은 사실상 적은 양의 화자종속 데이터를 이용한 재훈련을 통해 화자종속 모델을 새로 만드는 것임에도 불구하고 화자독립 모델의 부분상태 구조를 동일하게 사용한

다. 따라서 적은 수의 substate 구조를 갖춘 음소에 해당하는 데이터가 적응 데이터에서 비교적 많은 경우에는 과적합이 발생할 수 있고, 반대로 적응 데이터의 분량이 부족한 음소에서는 부적합이 일어날 수 있다. 이는 MAP이나 L1-norm regularization과 같은 훈련 상의 제약조건을 걸었을 때 성능이 확연히 개선되는 모습을 통해 알 수 있다.

이제 두 가지 음향모델에서 MAP 적응과 L1-norm regularization 기반 적응을 비교한다. GMM-HMM 시스템에서는 MAP 적응에 비해 L1-norm regularization의 성능이 단어 오류율 기준 상대적으로 8%가량 나빠진 데 반해, SGMM 시스템에서는 그 차이가 1% 정도로 성능이 거의 비슷함을 알 수 있다. 이는 SGMM 시스템에서는 L1-norm regularization 기반 적응이 MAP 적응에 비해 실제로 업데이트하는 변수의 수는 적으면서 거의 동등한 성능을 보장한다는 것을 뜻한다. 게다가 sparse solution의 특성상 적응 데이터의 분량이 적을수록 L1-norm regularization은 다른 적응 기법에 비해 더 나은 성능을 보여준다.

특이한 사항으로 MAP 적응에서 단위행렬 covariance를 사용했을 때, 즉 L2-norm regularization을 진행했을 때가 L1-norm regularization보다 성능이 좋은데 이는 적응 데이터의 분량이 충분해 state vector 파라미터를 높은 비율로 업데이트할 수 있는 경우, 후자가 λ값에 따라 sparsity를 어느 정도는 유지하는 데 반해 전자는 성능 향상에 기여하는 실질적인 파라미터 수가 많아지기 때문이다.

관련하여 <그림 2>에서 적응 데이터의 수에 따른 각 적응 기법의 단어 오류율을 비교한 결과를 제시한다. 가로축은 적응 데이터의 수로 화자당 12, 24, 36, 48개의 발화를 각각 사용한 경우를 나타내고 세로축은 각 적응 데이터 분량과 적응 방법에 대한 단어 오류율의 크기를 가리킨다. 점선은 GMM-HMM 음향모델, 실선은 SGMM 음향모델을 각각 가리키고 삼각형 점은 MAP 적응, 사각형 점은 L1-norm regularization을 나타낸다. 모든 적응 데이터 분량에 대해 GMM-HMM에서는 MAP 적응이 더 좋은 성능을 보이는 반면 SGMM에서는 L1-norm regularization이 MAP 적응에 비해 유사하거나 더 좋은 성능을 보이는 것을 볼 수 있다. 단순히 성능이 좋을 뿐 아니라 sparsity를 비교할 경우 SGMM 음향모델에서 L1-norm regularization을 쓰는 것의 장점은 더 도드라진다.

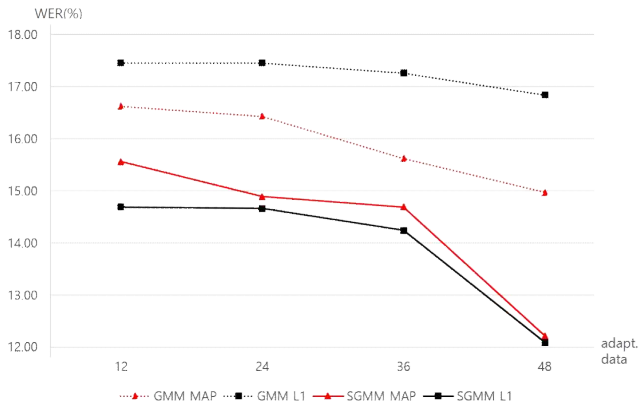

그림 2. 적응 방법별 적응 데이터의 분량에 따른 단어 오류율 추이 (Dev93)
Figure 2. WER trend line of various adaptation approaches in relation to the amount of adaptation data (Dev93)

표 2. SGMM 음향모델에서 적응 데이터의 분량에 대한 세 가지 적응 방법의 sparsity 비교 (Dev93)
Table 2. Sparsity trend of three approaches in SGMM acoustic model (Dev93)

	12	24	36	48
SGMM MLE	28.26	19.74	13.82	8.39
SGMM MAP	28.26	19.74	13.82	10.81
SGMM L1	47.36	46.36	45.59	44.74

그 맥락에서 sparsity를 대략적으로 비교할 수 있도록 <표 2>를 수록하였다. <표 2>에는 SGMM에 적용한 세 가지 적응 방법의 sparsity를 비교해서 기록해 두었는데, 여기서의 sparsity는 모든 state vector의 전체 파라미터의 수 대비 값이 변하지 않은 파라미터의 수로 나타내었으며 비슷한 수준의 단어 오류율을 기록했을 경우에는 sparsity가 높을수록 적은 수의 숫자를 적응시켜 비견할 만한 음향모델 성능을 얻었다는 의미이다. 이 때 sparse solution 추정을 하지 않는 ML 및 MAP 추정은 유사한 sparsity값을 가지는데, 이것은 적응 데이터에서의 미발견 음소(unseen phone) 비율로 생각할 수 있다. 이 때 L1-norm regularization 기반 적응에서는 미발견 음소량에 비해 상당수의 파라미터가 변하지 않은 것을 볼 수 있고, 특히 주목할 만한 점은 데이터량에 무관하게 유사한 sparsity를 유지하고 있다는 것으로 이는 많은 사용자에 대해 맞춤형 음성인식기를 제공하는 상황에서 특정인에게 맞게 적응시킨 모델을 비교적 적은 용량으로 안정적으로 저장할 수 있게 한다.

<표 3>에서는 projection matrix $\mathbf{M}_i$의 MAP 적응 ([8])을 이용한 화자적응 성능과, substate vector 적응 및 projection matrix 적응의 여러 조합을 종합적으로 제시한다. 또한 모든 조합에 디코딩시 fMLLR (feature-space MLLR)을 추가로 수행한 결과도 덧붙인다.

가장 주목할 만한 사항은 projection matrix에 대한 적응으로, 이는 ML 기반 추정의 경우에도 기본적으로 효과가 좋은데 그것은 각 substate vector의 훈련이 음소별로 이루어지는데 반해 projection matrix는 변수 전체의 훈련에 적응 데이터 전체가 대응되는 전역(global) 적응이므로 더욱 안정적인 훈련이 가능했기 때문으로 보인다. 또한 음소 발음 방식만으로 설명되지 않는 화자의 차이를 SGMM의 projection matrix가 반영하고, 그

표 3. 여러 적응 기법의 조합
Table 3. Combinations of various adaptation methods

	WER (%)	
	normal	fMLLR
[GMM-HMM] SI	14.48	12.55
MLE	12.70	11.29
MAP ($\tau=20$)	12.39	11.37
L1 ($\lambda=20$)	13.32	11.62
[SGMM] SI	11.85	10.96
MLE ($\mathbf{v}_{jm}$)	12.71	12.43
MLE ($\mathbf{M}_i$)	10.53	10.88
MAP ($\mathbf{v}_{jm}$, iden, $\tau=20$)	10.07	9.56
MAP ($\mathbf{M}_i$ $\tau=10$)	9.85	9.88
L1 ($\mathbf{v}_{jm}$, $\lambda=20$)	10.22	9.81
MAP ($\mathbf{v}_{jm}$), MAP ($\mathbf{M}_i$) (6)	10.10	10.46
L1 ($\mathbf{v}_{jm}$), MAP ($\mathbf{M}_i$) (10)	9.83	9.76

음소 발음 외적인 요소도 화자적응에서 상당한 중요성을 차지함을 말해준다.

그러나 projection matrix의 적응이 반영하는 화자간 변이는 fMLLR과 같은 다른 전역 적응이 거의 유사한 수준으로 반영하고 있음을 알 수 있는데, $\mathbf{v}_{jm}$에 MAP 적응을 한 후 fMLLR을 적용한 것이 $\mathbf{v}_{jm}$과 $\mathbf{M}_i$를 동시에 적응시킨 것과 비등한 성능을 보이고 있다는 사실 및 $\mathbf{M}_i$를 MAP 적응한 후 fMLLR을 적용하면 적용 전보다 오히려 성능이 떨어지는 것에서 추측할 수 있다. 반면에 같은 관찰을 통해 state vector에 대한 적응은 다른 전역 적응 방식으로 대체할 수 없다는 것을 알 수 있다.

<표 2>의 마지막 두 열은 논문에서 제안한 두 가지 state vector 적응 방법론에 projection matrix 적응을 같이 수행한 것이다. projection matrix 적응이 더해질 경우에는 substate vector의 MAP 적응에 비해 L1-norm regularization의 성능이 더 뛰어난 모습을 볼 수 있는데 이는 많은 변수를 업데이트해야 하는 상황에서 regularization을 통해 업데이트 대상을 한정짓는 것이 좋은 결과를 가져온 것으로 추측할 수 있다. 같은 맥락에서 fMLLR을 추가로 적용할 경우에 후자에서만 추가적 성능 향상이 이루어진 것도 설명할 수 있다. 즉, 여러 적응 기법을 적용하는 경우 가능한 한 sparse adaptation을 하는 것이 적절함을 나타낸다. 또한 이는 부분공간이 적절하게 재훈련된 경우 좌표벡터는 변화가 크지 않도록 sparse하게 추정하는 것이 바람직하다고 표현할 수도 있다. 이를 통해 간단하고 적은 용량을 차지해야 하는 적응을 수행할 때나 점진적 학습을 할 때는 제안한 방법론이 적절함을 알 수 있다.

5. 결론

본 논문에서는 MAP 적응의 연장선상에서 L1-norm regularization 기반 SGMM state vector 적응에 대해 논하고 또한 화자적응의 관점에서 성능 검증을 위한 실험을 수행하고 결과를 분석하였다. 제안한 방법을 통해, 특히 적응 데이터가 제한된 상황에서 기존의 GMM-HMM 음향모델을 사용한 경우, 혹은 SGMM 음향모델에 기존의 MAP 적응을 사용한 경우에 비해 향상된 성능을 얻을 수 있었다. 특히 SGMM에서는 아직 SGMM에 특화된 화자적응 방법론이 많이 제안되지 않았기 때문에, 본 논문에서 제안한 방법은 다른 전역 적응 기법과 같이 사용하는 경우 유용하게 사용될 것으로 보인다.

향후에는 대규모의 자연어 발화 데이터베이스를 사용하고, 더 다양한 적응기법과 성능을 견주어 본 연구의 의의를 재검증할 예정이다. 또한 미발견 음소 문제를 해결하고, reweighted L1-norm regularization을 적응하고 점진적 적응 프레임워크를 만들어 실시간 적응을 구현하는 것에 대해서도 연구할 계획이다.

감사의 글

본 연구는 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행되었습니다(No. 2014R1A2A2A01007650).

참고문헌

[1] Benzeghiba, M. et al. (2007), Automatic speech recognition and speech variability: A review, *Speech Comm.*, Vol. 49, No. 10-11.

[2] Huang, X., Acero, A., and Hon, H.-W (2001), *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall.

[3] Gales, M. (2000), Cluster adaptive training of hidden Markov models, *IEEE Trans. Speech and Audio Process.*, Vol. 8, No. 4.

[4] Kuhn, R. et al. (1998), Eigenvoices for speaker adaptation. in *Proc. ICSLP.*

[5] Povey, D. et al. (2010), Subspace Gaussian Mixture Models for speech recognition, in *Proc. ICASSP.*

[6] Povey, D. et al. (2011), The subspace Gaussian mixture model −A structured model for speech recognition, *Computer Speech and Language*, Vol. 25, No. 2.

[7] Burget, L. et al. (2010), Multilingual acoustic modeling for speech recognition based on subspace Gaussian Mixture Models, in *Proc. ICASSP.*

[8] Lu, L. et al. (2012), Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition, in *Proc. ICASSP.*

[9] Hamidi, S. and Rose, R. C. (2013), Phonetic subspace

adaptation for automatic speech recognition, in *Proc. ICASSP*.
[10] Kim, Y. and Kim, H. (2014), Constrained mle-based speaker adaptation with l1 regularization, in *Proc. ICASSP*.
[11] Chen, S. S. et al. (1998), Atomic Decomposition by Basis Pursuit, *SIAM J. Scientific Computing*, Vol. 20.
[12] Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *J. Roy. Stat. Soc. Series B (Methodological)*, Vol. 58, No. 1.
[13] Povey, D. (2009), *A tutorial-style introduction to subspace Gaussian mixture models for speech recognition*, Microsoft research, Redmond, WA, Tech. Rep.
[14] Olsen, P. A. et al. (2011), Sparse Maximum A Posteriori adaptation, in *Proc. ASRU*.
[15] Lu, L. et al. (2011), Regularized subspace Gaussian mixture models for cross-lingual speech recognition, in *Proc. ASRU*.
[16] Lu, L. et al. (2011), Regularized Subspace Gaussian Mixture Models for Speech Recognition, *IEEE Signal Processing Letters*, Vol. 18, No. 7.
[17] Figueiredo, M. A. et al. (2007), Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems, *IEEE J. Selected Topics Signal Process.*, Vol. 1, No. 4.
[18] Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008), Enhancing sparsity by reweighted L1 minimization, *J. Fourier Analysis Applicat.*, Vol. 14.
[19] Asif, M. S. and Romberg, J. (2013), Fast and Accurate Algorithms for Re-Weighted L1-Norm Minimization, *IEEE Trans. Signal Process.*, Vol. 61, No. 3.
[20] Povey, D., Ghoshal, A., and Boulianne, G. (2011), The Kaldi speech recognition toolkit, in *Proc. ASRU*.

• **구자현 (Goo, Jahyun)**
한국과학기술원 전기 및 전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: jahyun.goo@kaist.ac.kr
관심분야: 음성인식, 화자적응
현재 전기 및 전자공학부 박사과정 재학 중

• **김영관 (Kim, Younggwan)**
한국과학기술원 전기 및 전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: cleanthink@kaist.ac.kr
관심분야: 음성인식, 화자적응, 화자인식
현재 전기 및 전자공학부 박사과정 재학 중

• **김회린 (Kim, Hoirin)** 교신저자
한국과학기술원 전기 및 전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7417 Fax: 042-350-7619
Email: hoirkim@kaist.ac.kr
관심분야: 음성인식, 화자적응, 화자인식, 음성신호처리
2000~현재 전기 및 전자공학부 교수