Kim et al. EURASIP Journal on Advances in Signal Processing (2015) 2015:102
DOI 10.1186/s13634-015-0290-2

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

# Scaled norm-based Euclidean projection for sparse speaker adaptation

Younggwan Kim*, Myung Jong Kim and Hoirin Kim

**Abstract**

To reduce data storage for speaker adaptive (SA) models, in our previous work, we proposed a sparse speaker adaptation method which can efficiently reduce the number of adapted parameters by using Euclidean projection onto the $L_1$-ball (EPL1) while maintaining recognition performance comparable to maximum *a posteriori* (MAP) adaptation. In the EPL1-based sparse speaker adaptation framework, however, the adapted Gaussian mean vectors are mostly concentrated on dimensions having large variances because of assuming unit variance for all dimensions. To make EPL1 more flexible, in this paper, we propose scaled norm-based Euclidean projection (SNEP) which can consider dimension-specific variances. By using SNEP, we also propose a new sparse speaker adaptation method which can consider the variances of a speaker-independent model. Our experiments show that the adapted components of mean vectors are evenly distributed in all dimensions, and we can obtain sparsely adapted models with no loss of phone recognition performance from the proposed method compared with MAP adaptation.

**Keywords:** Euclidean projection onto the $L_1$-ball, MAP adaptation, Scaled norm-based Euclidean projection, Sparse speaker adaptation

## 1 Introduction

In these days, modern server-based speech recognition systems (SRSs) serve millions of users. For this reason, reducing data storage for speaker adaptive (SA) acoustic models becomes an important issue when considering speaker adaptation to enhance speech recognition performance. There are various adaptation methods for Gaussian mixture model-hidden Markov model (GMM-HMM)-based SRS [1–5]. Among those methods, maximum *a posteriori* (MAP) speaker adaptation is the most conventional and powerful method when relatively large amount of adaptation data that is about 20 min to 10 h long is available [6, 7].

SA models obtained by MAP adaptation require the data storage as much as a speaker-independent (SI) model needs, and the SI model typically has billions of parameters. Olsen et al. showed that most of the adapted parameters obtained by MAP adaptation are not closely related to speech recognition performance [6, 7]. To restrict the redundant parameter adjustments, they proposed sparse MAP (SMAP) adaptation in which a typical MAP problem is maximized with certain sparse constraints. In the SMAP approach, two sets of optimization parameters need to be controlled. The first set of the optimization parameters are related to parameter regularization which is used for typical MAP adaptation. The second set of the parameters are used to restrict the redundant parameter adjustments. However, the more parameters we have, the harder it becomes to tune those parameters because the parameters are empirically chosen to show the best recognition performance.

To resolve the aforementioned problem, in our previous work, we first reinterpreted the MAP adaptation as a constrained optimization problem with an $L_2$ norm-based constraint [8, 9]. To obtain sparsely updated SA models, we replace the $L_2$ norm-based constraint with an $L_1$ norm-based constraint. From the modification, we proposed a sparse adaptation method based on Euclidean projection onto the $L_1$-ball (EPL1) [10], which only requires a single control parameter. By using the proposed sparse adaptation method, we showed that less data storage for SA models can be obtained with almost no loss of phone recognition performance than the SMAP adaptation method. Although the number of control parameters can be dramatically reduced, EPL1-based speaker

* Correspondence: cleanthink@kaist.ac.kr
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-Ro, Yuseong-Gu, Daejeon 305-701, South Korea

adaptation still has a limitation that variances cannot be considered. Because of the limitation, parameters having large variances are only adapted during the adaptation step. However, we believe that parameters with small variances can also reflect speaker characteristics. Thus, in this paper, we propose scaled norm-based Euclidean projection (SNEP) which is a generalized version of EPL1, utilizing dimension-specific variances. From the SNEP framework, we also propose a new sparse speaker adaptation method. From our experiments, it is shown that the proposed SNEP-based speaker adaptation method can sparsely adapt the SI model (only about 9 % of the total number of parameters) with no loss of phone recognition performance against MAP adaptation.

The rest of this paper is organized as follows. In Section 2, we introduce EPL1 and a piecewise root finding (PRF) method which is a well-known solver for EPL1 [11, 12]. In Section 3, from the derivation of EPL1, we describe the modified optimization problem and how to find the optimal solution of SNEP. In Section 4, we briefly review MAP- and EPL1-based speaker adaptation. In Section 5, we describe our SNEP-based sparse speaker adaptation method using the variances of the SI model. In Section 6, we analyze our experimental results on adapted mean vectors and speech recognition performance. We conclude this paper in Section 7.

## 2 Euclidean projection onto the $L_1$-ball

Euclidean projection onto the $L_1$-ball (EPL1) is widely used for gradient projection methods [13–18] which are used to find the optimal sparse solution of a constrained optimization problem which is given by

$$\min_{\mathbf{x}\in\mathbb{R}^D} \mathscr{L}(\mathbf{x}) \qquad \text{s.t. } \|\mathbf{x}\|_1 \le c \tag{1}$$

where $\mathscr{L}: \mathbb{R}^D \to \mathbb{R}$ is a convex and differentiable loss function, $||\cdot||_1$ indicates an $L_1$ norm operator enforcing the sparse solution, and $c$ is a constant for controlling regularization and sparsity, meaning how many zeros are in the optimal solution vector. Gradient projection with Nesterov's method [19–22] is an optimal first-order black-box method and can find the optimal solution of (1) by generating a sequence $\{\mathbf{x}^k\}$ which is obtained from

$$\mathbf{x}^{k+1} = \prod\nolimits_{L_1}\left(\mathbf{s}^k - \eta_k \nabla\mathscr{L}\left(\mathbf{s}^k\right)\right) \tag{2}$$

where $\mathbf{s}^k = \mathbf{x}^k + \alpha_k(\mathbf{x}^k - \mathbf{x}^{k-1})$, $\alpha_k$, and $\eta_k$ are learning rates selected by certain rules [23], $\nabla\mathscr{L}(\mathbf{s}^k)$ is the gradient of $\mathscr{L}(\cdot)$ at $\mathbf{s}^k$, and $\prod_{L_1}(\mathbf{y})$ is the EPL1 problem defined as

$$\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v}-\mathbf{y}\|_2^2 \quad \text{s.t. } \|\mathbf{v}\|_1 \le c \tag{3}$$

where $||\cdot||_2^2$ is squared $L_2$ norm operator. In practice, (3) is modified into another constrained optimization problem which is given by

$$\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u}-\mathbf{z}\|_2^2 \quad \text{s.t. } \|\mathbf{u}\|_1 \le c,\ \mathbf{u} \succcurlyeq \mathbf{0} \tag{4}$$

where $\mathbf{z}$ is composed of absolute values of components in $\mathbf{y}$, $\succcurlyeq$ denotes component-wise inequality, and $\mathbf{0}$ is a vector with all zero components. The optimal solution of (3) can be obtained by

$$\mathbf{v}^* = sign(\mathbf{y}) \odot \mathbf{u}^* \tag{5}$$

where $sign(\boldsymbol{\rho})$ returns the vector whose components are signs of all components in $\boldsymbol{\rho}$, $\odot$ is component-wise multiplication of two vectors, and $\mathbf{u}^*$ is the optimal solution of (4) which can be solved by Lagrangian function given by

$$L(\mathbf{u},\lambda) = \frac{1}{2}\|\mathbf{u}-\mathbf{z}\|_2^2 + \lambda(\|\mathbf{u}\|_1 - c) - \boldsymbol{\kappa}^T\mathbf{u} \tag{6}$$

where $\lambda$ and $\boldsymbol{\kappa}$ are the Lagrangian multipliers. We assume that optimal value $\lambda^*$ is known and $||\mathbf{z}||_1 > c$. Since the components in (6) can be decoupled, the closed form solution is as follows [10]:

$$u_i^* = \max(0,\ z_i - \lambda^*),\ \ i = 1, \ldots, D. \tag{7}$$

According to the optimal vector $\mathbf{u}^*$, $i$ is the component index; the constraints of (4) can be expressed as

$$\sum_{i=1}^{D} \max(0, z_i - \lambda^*) = c. \tag{8}$$

To find the optimal value of $\lambda$, a piecewise linear function [11, 12] is used, which is given by

$$f(\lambda) = \sum_{i=1}^{D} \max(0, z_i - \lambda) - c = \sum_{i\in R_\lambda} z_i - |R_\lambda|\lambda - c \tag{9}$$

where $R_\lambda = \{i | i \in \{1, \ldots, D\}, z_i > \lambda\}$ and $|R|$ is the number of elements in the set $R$. Figure 1 shows an illustration of $f(\lambda)$ and a first-order gradient-based iterative method called piecewise root finding (PRF) [12] for the optimal value of $\lambda$. With the PRF method, we can generate a sequence $\{\lambda^k\}$ via

$$\lambda^k = \frac{\sum_{i\in R_{\lambda^{k-1}}} z_i - c}{\left|R_{\lambda^{k-1}}\right|} \tag{10}$$

until $f(\lambda^k) = 0$ is satisfied. As shown in Fig. 1, each $\lambda^k$ for $k \ge 1$ represents the root of a tangent line. To determine the set $R_{\lambda^k}$, every component of $\mathbf{z}$ needs to be compared with $\lambda^k$. If we set an initial value of $\lambda$ to 0, the sequence $\{\lambda^k\}$ could have a non-decreasing property. According to the property, in the $k$th step, we can skip the comparing operations for the components decided as less than $\lambda^{k-1}$.

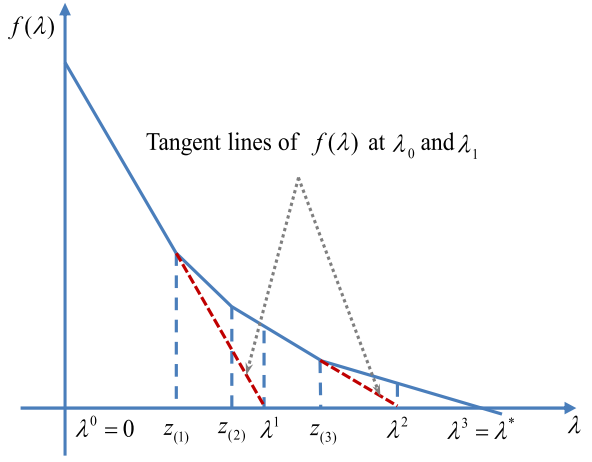


**Fig. 1** Illustration of $f(\lambda)$ and piecewise root finding method with an initial condition $\lambda^0 = 0$. $z_{(i)}$ is the $i$th smallest component of a vector **z**

## 3 Scaled norm-based Euclidean projection

Basically, $L_2$ and $L_1$ norm for EPL1 can be interpreted as a multivariate Gaussian distribution with unit variance and a multivariate Laplace distribution with unit standard deviation [24]. Hence, every component in EPL1 is equally treated for optimization without considering any scaling parameters such as dimension-specific variances and standard deviations. For this reason, we propose a scaled norm-based Euclidean projection (SNEP) method which is a more generalized version of EPL1. The proposed constrained optimization problem for SNEP is given by

$$\min_{\mathbf{u}} \frac{1}{2}\sum_{i=1}^{D}\left\{\frac{(u_i - z_i)}{\sigma_{2,i}}\right\}^2 \quad \text{s.t.} \quad \sum_{i=1}^{D}\frac{u_i}{\sigma_{1,i}} \le c, \quad \mathbf{u} \succcurlyeq \mathbf{0} \tag{11}$$

where $\sigma_{2,i}$ and $\sigma_{1,i}$ denote scaling parameters for $L_2$ and $L_1$ norm, respectively. As shown in (11), we can apply any dimension-specific scaling parameters to the SNEP framework. The Lagrangian function of (11) and its differentiation with respect to $u_i$ are given by

$$L^{\text{SNEP}}(\mathbf{u}, \lambda) = \frac{1}{2}\sum_{i=1}^{D}\left\{\frac{(u_i - z_i)}{\sigma_{2,i}}\right\}^2 + \lambda\left(\sum_{i=1}^{D}\frac{u_i}{\sigma_{1,i}} - c\right) - \boldsymbol{\kappa}^T\mathbf{u} \tag{12}$$

$$\frac{dL^{\text{SNEP}}(\lambda, \mathbf{u})}{du_i} = \frac{u_i - z_i}{\sigma_{2,i}^2} + \frac{\lambda}{\sigma_{1,i}} - \kappa_i. \tag{13}$$

By setting $dL^{\text{SNEP}}(\lambda, \mathbf{u})/du_i = 0$ and considering the complementary slackness KKT condition, the optimal value $u_i^*$ is given by

$$u_i^* = \max\left(0, z_i - \frac{\sigma_{2,i}^2}{\sigma_{1,i}}\lambda^*\right), \quad i = 1, \ldots, D \tag{14}$$

with optimal value $\lambda^*$. By using (14), the piecewise linear function for SNEP is given by

$$\begin{aligned} f^{\text{SNEP}}(\lambda) &= \frac{dL^{\text{SNEP}}(\lambda, \mathbf{u})}{d\lambda} = \sum_{i=1}^{D}\frac{u_i}{\sigma_{1,i}} - c \\ &= \sum_{i=1}^{D}\frac{1}{\sigma_{1,i}}\max\left(0, z_i - \frac{\sigma_{2,i}^2}{\sigma_{1,i}}\lambda\right) - c \\ &= \sum_{i \in R_\lambda^{\text{SNEP}}}\left(\frac{z_i}{\sigma_{1,i}} - \frac{\sigma_{2,i}^2}{\sigma_{1,i}^2}\lambda\right) - c \end{aligned} \tag{15}$$

where $R_\lambda^{\text{SNEP}} = \left\{i \middle| i \in \{1, \ldots, D\}, z_i > \lambda\sigma_{2,i}^2/\sigma_{1,i}\right\}$. By setting $f^{\text{SNEP}}(\lambda) = 0$, the sequence $\{\lambda^k\}$ from $f^{\text{SNEP}}(\lambda)$ is generated as

$$\lambda^k = \frac{\sum_{i \in R_{\lambda^{k-1}}^{\text{SNEP}}}\frac{z_i}{\sigma_{1,i}} - c}{\sum_{i \in R_{\lambda^{k-1}}^{\text{SNEP}}}\frac{\sigma_{2,i}^2}{\sigma_{1,i}^2}} \tag{16}$$

, and the PRF method can also be used to find the optimal solution of SNEP with the initial condition, $\lambda^0 = 0$.

## 4 Previous work for speaker adaptation

For better understanding, our previous sparse speaker adaptation, MAP-based speaker adaptation is described first. Let $\boldsymbol{\Phi} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\Theta}\}$ be the whole parameter set of HMMs, where $\boldsymbol{\pi}$ is the initial state distribution, $\boldsymbol{A}$ is the transition probability matrix, and $\boldsymbol{\Theta}$ is the set of GMMs for every state. The GMM distribution of state $s$ is given as follows:

$$p(\mathbf{x}|\boldsymbol{\Theta}_s) = \sum_{g=1}^{M} w_{g,s}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{g,s}, \boldsymbol{\Sigma}_{g,s}) \tag{17}$$

where $\mathcal{N}(\cdot)$ is a normal distribution, $M$ is the number of Gaussian components, and $w_{g,s}$, $\boldsymbol{\mu}_{g,s}$, and $\boldsymbol{\Sigma}_{g,s}$ are the weight, mean vector, and covariance matrix of Gaussian component $g$, respectively. In this paper, $\boldsymbol{\Sigma}_{g,s}$ is set as diagonal matrix whose diagonal components are represented as $[(\sigma_{1,g,s})^2, (\sigma_{2,g,s})^2, \ldots, (\sigma_{D,g,s})^2]^T$. Since MAP adaptation is typically performed on single state to adjust GMM parameters, we will omit the state index $s$ and describe every procedure in terms of GMM framework. Since, in addition, it is well known that adapting mixture weights and variances is not helpful for recognition performance, we focus on how to adapt mean vectors only.

In order to adapt mean vectors of the SI model, the MAP adaptation process is composed of two major stages. In the

first stage, mean vectors based on maximum likelihood (ML) criterion are computed for each mixture component of the SI model. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of acoustic feature vectors extracted from utterances of a target speaker. The *a posteriori* probability of Gaussian component $g$ for SI model is given by

$$p(g|\mathbf{x}_n) = \frac{w_g^{\mathrm{SI}}\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_g^{\mathrm{SI}}, \boldsymbol{\Sigma}_g^{\mathrm{SI}})}{\sum_{g'=1}^{M} w_{g'}^{\mathrm{SI}}\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_{g'}^{\mathrm{SI}}, \boldsymbol{\Sigma}_{g'}^{\mathrm{SI}})}. \tag{18}$$

With the probability of Gaussian component $g$, we then compute the ML mean vector:

$$\boldsymbol{\mu}_g^{\mathrm{ML}} = \frac{1}{n_g}\sum_{n=1}^{N} p(g|\mathbf{x}_n)\mathbf{x}_n \tag{19}$$

where $n_g = \Sigma_{n=1}^{N} p(g|\mathbf{x}_n)$ which is called posterior sum.

In the second stage, $\boldsymbol{\mu}_g^{\mathrm{ML}}$ is used to obtain the adapted mean vector from the SI model, which is given by

$$\boldsymbol{\mu}_g^{\mathrm{MAP}} = \frac{n_g\boldsymbol{\mu}_g^{\mathrm{ML}} + \tau\boldsymbol{\mu}_g^{\mathrm{SI}}}{n_g + \tau} \tag{20}$$

where $\tau$ is called the relevance factor which controls the balance between $\boldsymbol{\mu}_g^{\mathrm{ML}}$ and $\boldsymbol{\mu}_g^{\mathrm{SI}}$. By modifying (20), we can obtain

$$\boldsymbol{\mu}_g^{\mathrm{MAP}} = (\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}})\frac{n_g}{n_g + \tau} + \boldsymbol{\mu}_g^{\mathrm{SI}} = \boldsymbol{\phi}_g^{\mathrm{MAP}} + \boldsymbol{\mu}_g^{\mathrm{SI}}. \tag{21}$$

From (21), it is noticeable that $\boldsymbol{\phi}_g^{\mathrm{MAP}}$ is same as the optimal solution of the following constrained optimization problem, which is given by

$$\begin{aligned} &\min_{\boldsymbol{\phi}_g} \frac{1}{2}\|\boldsymbol{\phi}_g - (\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}})\|_2^2 \\ &\text{s.t.} \quad \|\boldsymbol{\phi}_g\|_2 \le \|\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}}\|_2 \frac{n_g}{n_g + \tau}. \end{aligned} \tag{22}$$

This constrained optimization problem is described in Fig. 2 from a geometric perspective. The shaded region implies the constraint part of (22), and the outer circle indicates the constraint when $n_g$ goes to infinity. As also shown in Fig. 2, the $L_2$ norm-based constraint can cause most of the small and redundant adjustments which can be negligible in terms of speech recognition performance. By replacing the constraint part of (22) with an $L_1$ norm-based constraint, we can efficiently restrict the redundant adjustments. The modified constrained optimization problem is given by

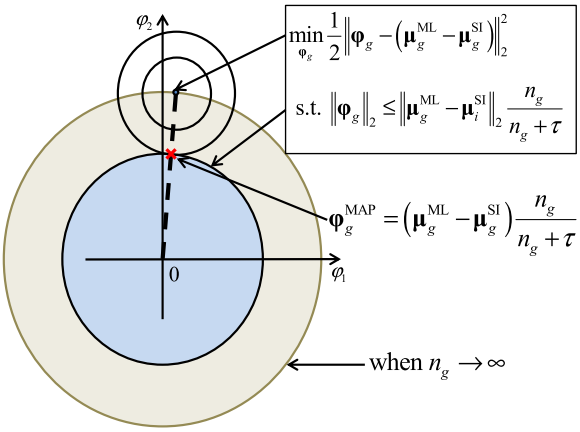


**Fig. 2** Geometric interpretation for MAP-based speaker adaptation

$$\begin{aligned} &\min_{\boldsymbol{\phi}_g} \frac{1}{2}\|\boldsymbol{\phi}_g - (\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}})\|_2^2 \\ &\text{s.t.} \quad \|\boldsymbol{\phi}_g\|_1 \le \|\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}}\|_1 \frac{n_g}{n_g + \tau}. \end{aligned} \tag{23}$$

The constrained optimization problem in (23) is exactly same as EPL1 except for the constraint part. As you can see in (23), the right-hand side of the constraint part is not the constant $c$ in previous section but variables depending mostly on $n_g$ and $\tau$. The posterior sum $n_g$ is naturally determined by the amount of adaptation data. Also, $n_g$ is used for considering the asymptotic property of adaptation, which means relaxation of regularization effect including sparsity as adaptation data increase. Thus, the parameter $\tau$ takes charge of controlling the sparsity and regularization instead of parameter $c$ for speaker adaptation. Figure 3 shows how the optimal solution can have sparse vectors indicated by the

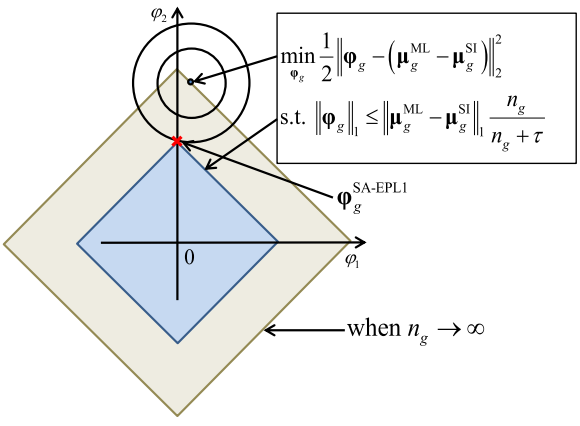


**Fig. 3** Geometric interpretation for EPL1-based sparse speaker adaptation

red cross. Before finding the optimal solution of (23), we first define a vector which is given by

$$\boldsymbol{\psi}_g = |\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}}| \tag{24}$$

where $|\boldsymbol{\rho}|$ returns the vector of absolute values in $\boldsymbol{\rho}$. To find the optimal solution of (23), we use $\boldsymbol{\psi}_g$ for the following steps. The Lagrangian form of (23) is given by

$$L^{\mathrm{SA-EPL1}}(\boldsymbol{\phi}_g, \lambda) = \frac{1}{2}\|\boldsymbol{\phi}_g - \boldsymbol{\psi}_g\|_2^2 + \lambda\left(\|\boldsymbol{\phi}_g\|_1 - \|\boldsymbol{\psi}_g\|_1 \frac{n_g}{n_g+\tau}\right) + \boldsymbol{\kappa}^T\boldsymbol{\phi}_g. \tag{25}$$

As described in Section 2, after being decoupled, the closed form solution of (23) with the optimal value $\lambda^*$, and the piecewise linear function in terms of $\lambda$ are given by

$$\phi_{i,g}^{\text{SA-EPL1}} = \max\left(0,\ \psi_{i,g} - \lambda^*\right),\ \ i = 1, \ldots, D \tag{26}$$

$$f^{\mathrm{SA-EPL1}}(\lambda) = \sum_{i \in R_\lambda^{\mathrm{SA-EPL1}}} \psi_{i,g} - |R_\lambda^{\mathrm{SA-EPL1}}|\lambda - \|\boldsymbol{\psi}_g\|_1 \frac{n_g}{n_g+\tau} \tag{27}$$

where $R_\lambda^{\text{SA-EPL1}} = \left\{ i | i \in \{1, \ldots, D\}, \psi_{i,g} > \lambda \right\}$. As also described in Section 2, $\lambda^*$ can be obtained by the sequence $\{\lambda^k\}$ from $f^{\text{SA - EPL1}}(\lambda)$, which is given by

$$\lambda^k = \frac{\sum_{i \in R_\lambda^{\mathrm{SA-EPL1}}} \psi_{i,g} - \|\boldsymbol{\psi}_g\|_1 \frac{n_g}{n_g+\tau}}{|R_\lambda^{\mathrm{SA-EPL1}}|} \tag{28}$$

when $f(\lambda^k) = 0$ is satisfied. Thus, the final adapted mean vector from EPL1-based sparse speaker adaptation is given as follows:

$$\boldsymbol{\mu}_g^{\mathrm{SA-EPL1}} = sign(\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}}) \odot \boldsymbol{\phi}_g^{\mathrm{SA-EPL1}} + \boldsymbol{\mu}_g^{\mathrm{SI}}. \tag{29}$$

## 5 SNEP-based sparse speaker adaptation

In GMM-HMM SRS, each dimension of Gaussian components typically has different variance denoting the dynamic range of each component. In Section 4, we describe the procedure for EPL1-based speaker adaptation which is unable to consider the dimension-specific variances. As a result, the adapted dimensions of mean vectors are mostly concentrated on the dimensions having large variances. Without considering the variances, the mean vectors adapted by EPL1 are not able to fully represent speaker-specific variability, which may cause loss of recognition performance. In this paper, we propose a new sparse speaker adaptation method using SNEP which can apply the variances of the SI model. Again, the proposed method utilizes $\boldsymbol{\psi}_g$ in all steps. The proposed constrained optimization problem for sparse speaker adaptation is given by

$$\begin{aligned} &\min_{\boldsymbol{\phi}_g} \frac{1}{2}(\boldsymbol{\phi}_g - \boldsymbol{\psi}_g)^T (\boldsymbol{\Sigma}_g^{\mathrm{SI}})^{-1} (\boldsymbol{\phi}_g - \boldsymbol{\psi}_g) \\ &\text{s.t.} \quad \sum_{i=1}^{D} \frac{\phi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}} \le \frac{n_g}{n_g+\tau} \sum_{i=1}^{D} \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}}, \quad \boldsymbol{\phi}_g \succcurlyeq \mathbf{0}. \end{aligned} \tag{30}$$

In (30), note that the same standard deviation $\sigma_{i,g}^{\mathrm{SI}}$ is shared by the objective function and the sparse constraint. The Lagrangian function of (30) is given by

$$\begin{aligned} L^{\mathrm{SA-SNEP}}(\boldsymbol{\phi}_g, \lambda) &= \frac{1}{2}(\boldsymbol{\phi}_g - \boldsymbol{\psi}_g)^T (\boldsymbol{\Sigma}_g^{\mathrm{SI}})^{-1} (\boldsymbol{\phi}_g - \boldsymbol{\psi}_g) \\ &+ \lambda\left(\sum_{i=1}^{D} \frac{\phi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}} - \frac{n_g}{n_g+\tau} \sum_{i=1}^{D} \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}}\right) - \boldsymbol{\kappa}^T\boldsymbol{\phi}_g. \end{aligned} \tag{31}$$

As described in Section 4, the closed form solution of (30) is

$$\phi_{i,g}^{\text{SA-SNEP}} = \max\left(0,\ \psi_{i,g} - \sigma_{i,g}^{\mathrm{SI}}\lambda^*\right),\ \ i = 1, \ldots, D. \tag{32}$$

Next, the piecewise linear function and the sequence $\{\lambda^k\}$ are given as follows:

$$\begin{aligned} f^{\text{SA-EPL1}}(\lambda) &= \sum_{i \in R_\lambda^{\text{SA-SNEP}}} \left(\frac{\psi_{i,g}}{\sigma_{i,g}} - \lambda\right) - \frac{n_g}{n_g+\tau} \sum_{i=1}^{D} \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}} \\ &= \sum_{i \in R_\lambda^{\text{SA-SNEP}}} \frac{\psi_{i,g}}{\sigma_{i,g}} - \left|R_\lambda^{\text{SA-SNEP}}\right|\lambda - \frac{n_g}{n_g+\tau} \sum_{i=1}^{D} \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}} \end{aligned} \tag{33}$$

$$\lambda^k = \frac{\sum_{i \in R_{\lambda^{k-1}}^{\text{SA-SNEP}}} \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}} - \frac{n_g}{n_g+\tau} \sum_{i=1}^{D} \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}}}{\left|R_{\lambda^{k-1}}^{\text{SA-SNEP}}\right|} \tag{34}$$

where $R_\lambda^{\text{SA-SNEP}} = \left\{ i | i \in \{1, \ldots, n\}, z_i > \sigma_{i,g}^{\mathrm{SI}}\lambda \right\}$. Since the objective function and the constraint share the same standard deviations, (32)-(34) are slightly modified from related equations in Section 3. For simple implementation, (32) can be changed into following form:

$$\frac{\phi_{i,g}^{\text{SA-SNEP}}}{\sigma_{i,g}^{\mathrm{SI}}} = \max\left(0,\ \frac{\psi_{i,g}}{\sigma_{i,g}^{\mathrm{SI}}} - \lambda^*\right),\ \ i = 1, \ldots, D. \tag{35}$$

Note that the right-hand sides of (34) and (35) are composed of scaled $\psi_{i,g}$ by $\sigma_{i,g}^{\mathrm{SI}}$. Thus, if we find the optimal solution with $\psi_{i,g}/\sigma_{i,g}^{\mathrm{SI}}$ by EPL1, the solution would be $\phi_{i,g}^{\text{SA-SNEP}}/\sigma_{i,g}^{\mathrm{SI}}$. By multiplying $\sigma_{i,g}^{\mathrm{SI}}$ with the solution, we can obtain exactly same result with (32).

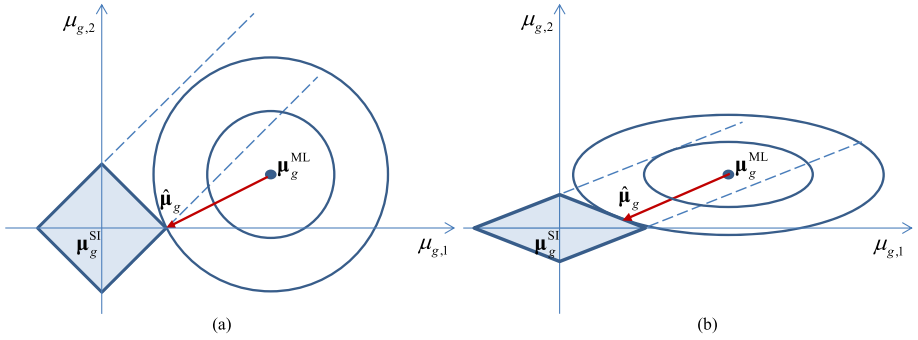


**Fig. 4** Illustration of **a** EPL1 and **b** SNEP-based sparse speaker adaptation from the same ML mean vector

Finally, the adapted mean vector is given as follows:

$$\boldsymbol{\mu}_g^{\mathrm{SA-SNEP}} = sign(\boldsymbol{\mu}_g^{\mathrm{ML}} - \boldsymbol{\mu}_g^{\mathrm{SI}}) \odot \boldsymbol{\phi}_g^{\mathrm{SA-SNEP}} + \boldsymbol{\mu}_g^{\mathrm{SI}}. \quad (36)$$

In Fig. 4, each figure shows how the sparse speaker adaptation methods work by EPL1 and SNEP. In the figure, the red arrows from the ML mean indicate the adapted mean vectors, and the region of non-sparse solution is surrounded by the two dashed lines. As can also be seen in Fig. 4, from the same ML mean vector, the differently adapted mean vectors are obtained because of the shared standard deviations.

## 6 Experimental results

The experiments were conducted on the ETRI Korean conversation speech database collected at 16 kHz sampling rate and 16-bit resolution by two types of smart phone devices in clean condition. We used about 100 h of speech data spoken by 300 speakers to train the SI triphone-based GMM-HMM acoustic model. For adaptation and evaluation, we used 50 speakers' 350 sentences (300 sentences for adaptation and 50 sentences for the phone recognition test) and each sentence is roughly 4–5 s long. We used 12-dimensional Mel-frequency cepstral coefficients with log energy and concatenated their first and second derivatives as a feature vector to constitute 39-dimensional feature vectors. We applied a phone level unigram language model in terms of 39 Korean phonemes to our phone recognition experiments. The SI model had 11,848 tied-state triphone-based HMMs including three states per each HMM and GMM with 32 Gaussian components per state. All phone recognition tests were performed according to various values of hyperparameter $\tau$.

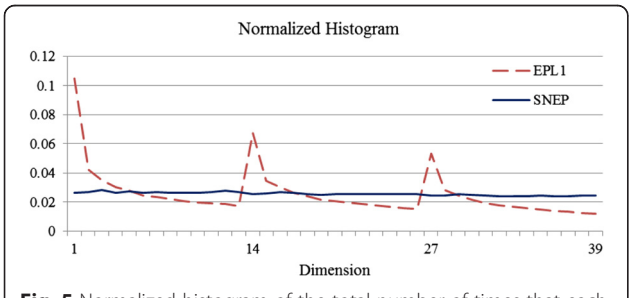


**Fig. 5** Normalized histogram of the total number of times that each dimension is adapted after adaptation with $\tau = 2$

To observe the effects of the variances of the SI model for SNEP compared with EPL1, we counted the number of times that each dimension of mean vectors was adapted during the adaptation process. In Fig. 5, $x$-axis indicates each dimension of the mean vector and normalized histogram of the counts is shown on $y$-axis. For EPL1, three distinct peaks are observed, and their dimensions are related to the log energy and its first and second derivatives. On the other hand, it is noticeable that there is no peak with SNEP and every dimension is evenly adapted. As mentioned earlier, we believe that speaker characteristic is not mainly concentrated on the three dimensions which are related to log energy. Therefore, it can be said that SNEP-based sparse speaker adaption can reflect more the speaker variability than the EPL1-based method.

In Table 1, phone error rate (PER) and sparsity of various methods are summarized, and the sparsity indicates the percentage of the number of parameters which are not adjusted after adaptation. For comparison purpose,

**Table 1** Phone error rate (%) and sparsity (%) for different $\tau$'s

| $\tau$ | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 |
|---|---|---|---|---|---|---|---|
| Phone error rate (%) | | | | | | | |
| SI | 31.45 | | | | | | |
| MLLR | 21.77 | | | | | | |
| MAP | 17.87 | 17.76 | 17.63 | 17.71 | 17.68 | 17.94 | 18.06 |
| EPL1 | 18.19 | 17.99 | 17.99 | 18.12 | 18.26 | 18.24 | 18.37 |
| SNEP | 18.37 | 18.23 | 17.98 | 17.85 | 17.63 | 17.75 | 17.74 |
| Sparsity (%) | | | | | | | |
| MAP | 50.96 | | | | | | |
| EPL1 | 89.45 | 91.37 | 92.62 | 93.52 | 94.19 | 95.13 | 95.48 |
| SNEP | 87.42 | 88.72 | 89.67 | 90.46 | 91.08 | 91.60 | 92.05 |

we also did phone recognition tests on SI model and maximum likelihood linear regression (MLLR) adaptation. For MLLR, we used full matrix and 65 regression classes which showed the best PER. The best PER for EPL1 is 17.99 % with 91.37 % sparsity. In contrast, the SNEP shows no recognition performance degradation against MAP adaptation with 91.08 % sparsity. From our experimental results, it is proven that sparse speaker adaptation with the dimension-specific variances can adapt the SI model more accurately than EPL1-based sparse speaker adaptation.

## 7 Conclusions

In this paper, we propose the SNEP method which is a more generalized version of EPL1 in which certain scaling parameters can be applied to the EPL1 framework. In addition, by using the SNEP method, we also propose sparse speaker adaptation. In our experiments, we show that a small number of dimensions are mostly adapted by EPL1-based speaker adaptation and the proposed speaker adaptation method can evenly adapt every dimension of the mean vectors by using the variances of the SI model. With the proposed methods, it is also shown that we can obtain sparsely adapted model with no loss of phone recognition performance compared with MAP adaptation. Our further work is to apply the EPL1 and SNEP framework to deep neural network-based acoustic model adaptation [25–28] with the gradient projection method.

#### Competing interests

The authors declare that they have no competing interests.

#### Acknowledgements

This work was partly supported by the ICT R&D program of MSIP/IITP [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90 % for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning] and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) [No. 2014R1A2A2A01007650].

Received: 24 June 2015 Accepted: 23 November 2015
Published online: 01 December 2015

#### References

1. CJ Leggetter, PC Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang **9**(2), 171–185 (1995)
2. MJF Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
3. R Kuhn, JC Junqua, P Nguyen, N Niedzielski, Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process. **8**(6), 695–706 (2000)
4. P Kenny, G Boulianne, P Dumouchel, Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process. **13**(3), 345–354 (2005)
5. JL Gauvain, CH Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. **2**(2), 291–298 (1994)
6. PA Olsen, J Huang, SJ Rennie, V Goel, Sparse maximum a posteriori adaptation, in *Proc. Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 53–58
7. PA Olsen, J Huang, SJ Rennie, V Goel, Affine invariant sparse maximum a posteriori adaptation, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2012, pp. 4317–4320
8. Y Kim, H Kim, Constrained MLE-based speaker adaptation with L1 regularization, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2014, pp. 6369–6373
9. C.M. Bishop, Pattern recognition and machine learning (Springer-Verlag, 2nd edition, 2006)
10. J Duchi, S Shalev-Shwartz, Y Singer, T Chandra, Efficient projections onto the L1-ball for learning in high dimensions, in *Proc. Int. Conf. Mach. Learn*, 2008, pp. 272–279
11. J Liu, J Ye, Efficient Euclidean projections in linear time, in *Proc. Int. Conf. Mach. Learn*, 2009, pp. 657–664
12. P Gong, K Gai, C Zhang, Efficient Euclidean projections via piecewise root finding and its application in gradient projection. Neurocomputing **74**(17), 2754–2766 (2011)
13. C Cortes, V Vapnik, Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
14. V. Vapnik, The nature of statistical learning theory (Springer, 2000)
15. S Shalev-Shwartz, Y Singer, N Srebro, A Cotter, Pegasos: primal estimated sub-gradient solver for SVM. Math. Program. **127**(1), 3–30 (2011)
16. R Tibshirani, Regression selection and shrinkage via the lasso. J. R. Stat. Soc. B. **58**(1), 267–288 (1996)
17. J Duchi, Y Singer, Efficient online and batch learning using forward backward splitting. J. Mach. Learn. Res. **10**, 2899–2934 (2009)
18. MAT Figueiredo, RD Nowak, SJ Wright, Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Top. Signal Process. **1**(4), 586–597 (2007)
19. Y Nesterov, A method of solving a convex programming problem with convergence rate O (1/k2). Sov. Math. Dokl. **27**(2), 372–376 (1983)
20. Y Nesterov, *Introductory lectures on convex optimization: a basic course* (Springer, Netherlands, 2004)
21. Y Nesterov, Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
22. Y Nesterov, Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
23. D.P. Bertsekas, *Nonlinear programming* (Athena Scientific, 1995)
24. S Kotz, T Kozubowski, K Podgorsk, *The Laplace distribution and generalizations* (Birkhäuser, Boston, 2001)
25. K Yao, D Yu, F Seide, H Su, L Deng, Y Gong, 2012, in *Proc. Spoken Language Technology Workshop*, 2012, pp. 366–369
26. A Mohamed, GE Dahl, G Hinton, Acoustic modeling using deep belief networks. IEEE Trans. Audio. Speech. Lang. Processing. **20**(1), 14–22 (2012)
27. GE Dahl, D Yu, L Deng, A Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech Lang. Process **20**(1), 30–42 (2012)
28. H Liao, Speaker adaptation of context dependent deep neural networks, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2013, pp. 7947–7951