



# MAP Estimation of Subspace Transform for Speaker Recognition

Donglai Zhu, Bin Ma, Kong-Aik Lee, Cheung-Chi Leung, Haizhou Li

Human Language Technology Department, Institute for Infocomm Research,  
A\*STAR, Singapore 138632

{dzhu, mabin, kalee, ccleung, hli}@i2r.a-star.edu.sg

## Abstract

We propose using the maximum-*a-posteriori* (MAP) estimation of subspace transform for speaker recognition. The linear transform is defined on the mean vectors of the Gaussian mixture model (GMM), where transform matrices and bias vectors are associated with separate regression classes so that both can be estimated with sufficient statistics given limited training data. The transform matrices are further defined as a linear combination of a set of basis transforms so that the weights are parameters to be estimated. We characterize the speakers with the transform parameters and model them using support vector machine (SVM). Experiments on the 2008 NIST SRE task illustrate the effectiveness of the method.

**Index Terms:** speaker recognition, subspace transform, maximum a posteriori

## 1. Introduction

This paper focuses on speaker recognition approaches based on support vector machine (SVM). The SVM-based approaches mainly differ in the definition of SVM features, e.g., the generalized linear discriminant sequence (GLDS) [1], the Gaussian mixture model (GMM) supervector [2], the bag of N-grams [3], and the maximum likelihood linear regression (MLLR) transform [4].

Among the SVM features, MLLR models the difference between a speaker-dependent and a speaker-independent model by sharing linear transforms across Gaussian mixtures. Multiple transforms are usually defined in order to sufficiently describe the speaker's characteristics [4]. However, the MLLR estimation may encounter numerical problems such as singular matrix inverse when the training data are sparse [5].

To address the sparse training data problem, one way is to estimate the linear regression with the maximum *a posteriori* (MAP) criterion [6][7][8]. MAP can yield a smooth estimation of parameters by assuming that the parameters belong to certain prior probabilities [9].

A second way to deal with the limited training data is to use the subspace concept e.g., eigenvoice [10], joint factor analysis (JFA) [11], extended maximum likelihood linear transform (EMLLT) [12], subspace on precisions and means (SPAM) [13], subspace GMM [14], cluster adaptive training (CAT) [15] and subspace feature MLLR [16]. The subspace can be defined either for the model parameters [10]-[14] or for the transform parameters [15][16]. This paper studies the subspace methods based on transform parameters.

We propose a method for estimating the subspace transform based on the MAP criterion. MAP estimation is used instead of ML estimation in order to eliminate possible numerical problems. The transform can be defined on either model space or feature space. In [15], MLLR is defined on Gaussian means and subspace transforms are estimated with the expectation-

maximum (EM) algorithm. On the other hand, in [16], feature-space MLLR is defined and subspace transforms are estimated with the line search algorithm because the Jacobian determinant makes the EM derivation intractable. Therefore, similar to [15], we define the transforms on Gaussian means to achieve a relatively simple solution. Considering that the estimation of bias vectors is more robust than that of transform matrices given limited amounts of training data, we previously studied a flexible definition of linear regression where transform matrices and bias vectors are associated with separate regression classes [17]. In this paper we define the transform in the same flexible form, and further define the transform matrices as a linear combination of a set of orthonormal basis transforms. The transform parameters, which can be used to characterize the speakers, are finally modeled by SVM.

Our method shares similarity with MLLR-SVM in the use of transform matrices, and has these two advantages: 1) MAP estimation eliminates possible numerical problems in ML estimation, and 2) subspace estimation is more robust than the full-matrix estimation in MLLR given limited training data. Our method also shares similarity with GMM-SVM where large number of bias vectors are defined, and has the advantage that bias vectors and transform matrices can be jointly estimated for speaker characterization.

The method is evaluated on the 2008 NIST SRE corpus. Results show the effectiveness of the subspace transform especially for short-duration speech. Good performance is obtained by setting different numbers of bias vectors and transform matrices.

## 2. Subspace Transform

Let's model the speaker-independent data consisting of  $D$ -dimensional feature vectors  $y_t$  with a GMM universal background model (UBM)

$$p(y_t|\mathcal{M}) = \sum_{m=1}^M c_m \mathcal{N}(y_t; \mu_m, \Sigma_m), \quad (1)$$

which is defined by a set of parameters  $\mathcal{M} = \{c_m, \mu_m, \Sigma_m; m = 1, \dots, M\}$ , where  $M$  is the number of Gaussian components,  $c_m$  are Gaussian mixture weights,  $\mathcal{N}(\cdot)$  is a Gaussian,  $\mu_m$  are mean vectors, and  $\Sigma_m$  are covariance matrices. Let's define a linear transform to map speaker-independent mean vectors  $\mu_m$  to speaker-dependent mean vectors  $\mu_{sm}$  as follows:

$$\mu_{sm} = A_{sk}\mu_m + b_{sl}, \quad (2)$$

where  $s$  denotes the speaker,  $A_{sk}$  is a nonsingular  $D \times D$  matrix,  $b_{sl}$  is a  $D$ -dimensional vector.

The regression classes  $k$  and  $l$  are associated with Gaussian components in the GMM-UBM by sharing the transform across

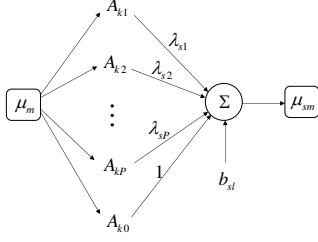


Figure 1: Subspace transform on Gaussian means.

mixture components. To this end, a centroid splitting algorithm with Euclidean distance measure is used to assign each component  $m$  to two separate classes:  $\mathcal{C}_k = \{m | k_m = k\}; k = 1, \dots, K$  and  $\mathcal{C}_l = \{m | l_m = l\}; l = 1, \dots, L$ . Typically we have  $K \leq M$  and  $L \leq M$ . The choice of  $K$  and  $L$  regulates the number of parameters. By allowing different combination of  $K$  and  $L$ , we are able to define a linear transform with the desired number of parameters.

When the speaker adaptation data are insufficient, the estimates of transform matrices  $A_{sk}$  may not be reliable. To address the problem, we define  $A_{sk}$  as a linear combination of a set of basis transforms as follows:

$$A_{sk} = A_{k0} + \sum_{p=1}^P \lambda_{skp} A_{kp}, \quad (3)$$

where the bases  $\{A_{kp}; p = 1, \dots, P\}$  define a subspace for transform matrices  $A_{sk}$ ,  $\lambda_{skp}$  are weights of the linear combination,  $A_{k0}$  denotes the speaker-independent transform matrix. Figure 1 illustrates the subspace transform on Gaussian means. Thus the speaker-dependent parameter set is composed of linear combination weights and bias vectors, denoted as  $\Theta_s = \{\lambda_{sk}, b_{sl}; k = 1, \dots, K; l = 1, \dots, L\}$ , where  $\lambda_{sk} = [\lambda_{sk1} \dots \lambda_{skP}]^T$ .

### 3. MAP Estimation

Given a speaker's feature set  $Y_s = \{y_{st}\}$ , the MAP estimation of  $\Theta_s$  is to maximize the posteriori probability as follows:

$$p(\Theta_s | Y_s, \mathcal{M}) = p(Y_s | \mathcal{M}, \Theta_s) p(\Theta_s), \quad (4)$$

where  $p(Y_s | \mathcal{M}, \Theta_s)$  is the likelihood of  $Y_s$  with respect to  $\mathcal{M}$  and  $\Theta_s$ , and  $p(\Theta_s)$  is the prior probability of  $\Theta_s$ . An iterative procedure, as described below, can be used to estimate  $\Theta_s$ , where  $\lambda_{sk}$  are estimated by fixing  $b_{sl}$  and  $b_{sl}$  are estimated by fixing  $\lambda_{sk}$ .

#### 3.1. Estimation of $\lambda_{sk}$

Let's define the prior probability of  $\lambda_{sk}$  as a Gaussian distribution as follows:

$$p(\lambda_{sk} | \Gamma_\lambda) = \mathcal{N}(\lambda_{sk}; \rho_k, \Psi_k / \varepsilon), \quad (5)$$

where  $\rho_k$  and  $\Psi_k$  are respectively the mean vectors and covariance matrices, while the parameter  $\varepsilon$  controls the broadness of the prior distribution.  $\Gamma_\lambda = \{\rho_k, \Psi_k\}$  is the set of hyperparameters.

The EM auxiliary function for the ML estimation of  $\Theta_s$  can be written as follows:

$$\mathcal{Q}_{ML} = C - \frac{1}{2} \sum_{m,t} \gamma_m(t) (y_{st} - \mu_{sm})^T \Sigma_m^{-1} (y_{st} - \mu_{sm}), \quad (6)$$

where the term  $C$  is irrelevant to the speaker parameter set  $\Theta_s$ , and  $\gamma_m(t)$  is the posteriori probability of the  $m$ -th Gaussian component given the feature  $y_{st}$  with respect to parameters  $\{c_m, \mu_{sm}, \Sigma_m\}$ . With the prior probability in Eq. (5), the EM auxiliary function for the MAP estimation of  $\lambda_{sk}$  can then be written as

$$\mathcal{Q}_{MAP}(\lambda_{sk}; \lambda'_{sk}) = \mathcal{Q}_{ML} - \frac{1}{2} \varepsilon (\lambda_{sk} - \rho_k)^T \Psi_k^{-1} (\lambda_{sk} - \rho_k). \quad (7)$$

Differentiating Eq. (7) with respect to  $\lambda_{sk}$  and equating to zero, we get

$$\lambda_{sk} = (G_{sk} + \varepsilon \Psi_k^{-1})^{-1} (e_{sk} + \varepsilon \Psi_k^{-1} \rho_k), \quad (8)$$

where

$$G_{sk} = \sum_{m \in \mathcal{C}_k, t} \gamma_m(t) U_m^T \Sigma_m^{-1} U_m, \\ e_{sk} = \sum_{m \in \mathcal{C}_k, t} \gamma_m(t) U_m^T \Sigma_m^{-1} (y_{st} - A_{k0} \mu_m - b_{slm}),$$

and  $U_m = [A_{k1} \mu_m \dots A_{kP} \mu_m]$ .

#### 3.2. Estimation of $b_{sl}$

Similar to  $\lambda_{sk}$ , we define the prior probability of  $b_{sl}$  as a Gaussian distribution as follows:

$$p(b_{sl} | \Gamma_b) = \mathcal{N}(b_{sl}; \nu_l, \Phi_l / \tau). \quad (9)$$

where  $\nu_l$  and  $\Phi_l$  are respectively mean vectors and covariance matrices, and the parameter  $\tau$  controls the broadness of the prior distribution.  $\Gamma_b = \{\nu_l, \Phi_l\}$  is the set of hyperparameters. The EM auxiliary function for the MAP estimation of  $b_{sl}$  can be written as

$$\mathcal{Q}_{MAP}(b_{sl}; b'_{sl}) = \mathcal{Q}_{ML} - \frac{1}{2} \tau (b_{sl} - \nu_l)^T \Phi_l^{-1} (b_{sl} - \nu_l). \quad (10)$$

Differentiating Eq. (10) with respect to  $b_{sl}$  and equating to zero, we get

$$b_{sl} = (H_{sl} + \tau \Phi_l^{-1})^{-1} (f_{sl} + \tau \Phi_l^{-1} \nu_l), \quad (11)$$

where

$$H_{sl} = \sum_{m \in \mathcal{C}_l, t} \gamma_m(t) \Sigma_m^{-1}, \\ f_{sl} = \sum_{m \in \mathcal{C}_l, t} \gamma_m(t) \Sigma_m^{-1} (y_{st} - A_{skm} \mu_m).$$

Assuming that both  $\Sigma_m$  and  $\Phi_l$  are diagonal matrices, i.e.  $\Sigma_m = \text{diag}\{\sigma_{m1}^2, \dots, \sigma_{mD}^2\}$  and  $\Phi_l = \text{diag}\{\phi_{l1}^2, \dots, \phi_{lD}^2\}$ , the estimation of  $b_{sl}$  can be simplified as follows:

$$b_{sli} = \frac{\sum_{m \in \mathcal{C}_l, t} \gamma_m(t) \frac{(y_{sti} - a_{skmi} \mu_m) + \frac{\tau \nu_{li}}{\phi_{li}^2}}{\sigma_{mi}^2}}{\sum_{m \in \mathcal{C}_l, t} \gamma_m(t) \frac{1}{\sigma_{mi}^2} + \frac{\tau}{\phi_{li}^2}}, \quad (12)$$

where the subscript  $i$  denotes the  $i$ -th element in vectors ( $b_{sli}$ ,  $y_{sti}$ ,  $\sigma_{mi}^2$ ,  $\nu_{li}$  and  $\phi_{li}^2$ ) or the  $i$ -th row in matrices ( $a_{skmi}$ ).

#### 3.3. Summary of estimation procedure

Given a speaker  $s$ ' data, we estimate  $\Theta_s$  with the following steps:

*Step 1:* Initialization of  $\lambda_{sk}$  and  $b_{sl}$ . Both vectors are initially set as 0 in our experiments.

*Step 2:* Estimation of  $\lambda_{sk}$  with Eq. (8).

*Step 3:* Estimation of  $b_{sl}$  with Eq. (12).

*Step 4:* Repeat Step 2 and Step 3 until a criterion is satisfied.

#### 4. Estimation of Basis Transforms

We estimate the basis transforms  $\{A_{kp}; k = 1, \dots, K, p = 0, \dots, P\}$  using the background data. For each speaker  $s$  in the background data, based on the speaker-independent GMM-UBM, a set of transforms  $\{A_{sk}; k = 1, \dots, K\}$  can be estimated by using MLLR [5]. The MLLR estimation among all speakers in the background data can be used as the zeroth transform matrices  $\{A_{k0}; k = 1, \dots, K\}$ .

It is useful to construct a set of orthonormal bases, i.e.  $tr(A_{ki}A_{kj}^T) = \delta(i, j)$ , where  $\delta(i, j)$  is the Kronecker delta function. To estimate the orthonormal bases, we compute the scatter matrix among the MLLR transforms as follows:

$$X_k = \sum_s vec(A_{sk})vec(A_{sk})^T, \quad (13)$$

where  $vec(A_{sk})$  denotes the concatenation of rows of  $A_{sk}$  into a supervector. Then SVD is used to compute the column vectors  $\{u_{kp}; p = 1, \dots, P\}$  corresponding to the top- $P$  singular values in  $X_k = U_k L_k V_k^T$ . Finally we get the bases as  $vec(A_{kp}) = u_{kp}$ .

#### 5. Estimation of Hyperparameters

The prior probabilities of  $\lambda_{sk}$  and  $b_{sl}$  are defined with hyperparameters consisting of  $\Gamma_\lambda = \{\rho_k, \Psi_k; k = 1, \dots, K\}$  and  $\Gamma_b = \{\nu_l, \Phi_l; l = 1, \dots, L\}$ . We estimate the hyperparameters on the background data. The idea is to estimate  $\lambda_{sk}$  and  $b_{sl}$  for each speaker in the background data, and to compute the distribution among these speakers as the prior probabilities. Therefore, the estimation includes the following two steps:

*Step 1:* Estimate  $\lambda_{sk}$  and  $b_{sl}$  for each speaker in the background data. As there is no prior probabilities in this stage, we use the ML estimation to update the parameters. ML is the special case of MAP where the prior probability is flat, i.e.  $\varepsilon \rightarrow 0$  in Eq. (5) and  $\tau \rightarrow 0$  in Eq. (9). Therefore, ML estimates of  $\lambda_{sk}$  and  $b_{sl}$  are respectively Eq. (8) and Eq. (12) in condition of  $\varepsilon = 0$  and  $\tau = 0$ .

*Step 2:* Estimate the hyperparameters  $\Gamma_\lambda$  and  $\Gamma_b$ . Given the values of  $\lambda_{sk}$  and  $b_{sl}$  among the speakers in the background data, we may compute the unbiased estimates of hyperparameters as  $\rho_k = E_s[\lambda_{sk}]$ ,  $\Psi_k = E_s[(\lambda_{sk} - \rho_k)(\lambda_{sk} - \rho_k)^T]$ ,  $\nu_l = E_s[b_{sl}]$ , and  $\Phi_l = E_s[(b_{sl} - \nu_l)(b_{sl} - \nu_l)^T]$ .

#### 6. SVM with Transform Parameters

Given the estimated  $\lambda_{sk}$ , we can compute the transform matrices  $A_{sk}$  with Eq. (3). Then we get a set of speaker-dependent transform parameters  $\Theta'_s = \{A_{sk}, b_{sl}; k = 1, \dots, K; l = 1, \dots, L\}$ , which we use for SVM modeling. For each speech utterance, the parameters are concatenated to form a supervector consisting of  $KD^2 + LD$  elements. An SVM is trained for each target speaker by taking the target speaker's training supervectors as positive samples, and the supervectors from the background data set as negative samples. Our experiments are implemented using the SVMtorch package with a linear inner-product kernel [18]. The supervectors are scaled to the same dynamic ranges using the rank normalization [4]. It replaces each value in the supervectors with its rank among the background data samples on a given dimension, and then scales the ranks to a value between 0 and 1. Rank normalization warps the distribution to be uniform for each dimension of the background vectors, which may result in better robustness for the SVM classifier. We perform nuisance attribute projection (NAP) on the

SVM supervector [19], and normalize the output SVM scores with Tnorm and Znorm [20].

#### 7. Experiments

We evaluate the method on the telephony English data in the 2008 NIST SRE corpus. According to the duration of training and testing speech utterances, the corpus is classified to five conditions which are denoted as: 10s-10s, 1c-10s, 8c-10s, 1c-1c, and 8c-1c. In the denotation, the two values represent training and testing duration respectively. "s" means a second and "c" means a conversation of approximately 2.5 minutes. The background data comprise the 2004 NIST SRE corpus which consists of 1816 utterances from 145 female speakers and 1268 utterances from 101 male speakers, and are used to train the GMM-UBM, basis transforms, and NAP transforms. The 1-conversation training data in the 2005 NIST SRE corpus are used for training the cohort models for Tnorm, and the 1-conversation training data in the 2004 NIST SRE corpus are used as cohort speakers in Znorm.

Each speech utterance is converted to a sequence of 39-dimensional feature vectors including 12 PLP coefficients,  $C_0$  and their first and second order derivatives, which are then filtered by a RASTA filter. An energy-based voice activity detection (VAD) process is then used to remove non-speech frames. Finally, feature vectors are normalized with Gaussianization.

Subspace transform (ST) is compared with two methods: 1) MLLR-SVM, in which the SVM supervector is composed of vectorized MLLR transform parameters that are estimated based on the GMM-UBM, and 2) GMM-SVM, in which the SVM supervector is composed of concatenated mean vectors in the speaker-dependent GMM which is MAP-adapted from the GMM-UBM. With the background data, we train two gender-dependent GMM-UBMs each consisting of 512 Gaussian components. In all methods, the NAP transform rank is set to be 40. In ST-SVM, we set  $P = 40$  and  $\varepsilon = \tau = 1$ .

We first study the performance of ST using the transform matrices  $A_{sk}$  only, which is defined by setting  $L = 0$ . The performance is compared with MLLR-SVM in which we define MLLR as  $K$  full matrices so that ST and MLLR have the same number of parameters. Without loss of generality, we compare the performance on male-speaker data. Table 1 shows the equal error rates (EERs) of the two methods. The value of  $K$  is changed from 1 to 8. For conditions involving short durations (10s-10s, 1c-10s and 8c-10s), ST-SVM outperforms MLLR-SVM, indicating that the short-duration speech is insufficient for MLLR estimation but is enough for ST estimation. In the mid-duration condition (1c-1c), ST cannot beat MLLR when  $K = 1, 2, 4$  and becomes better than MLLR when  $K = 8$ . In the long-duration condition (8c-1c), MLLR-SVM outperforms ST-SVM, showing that MLLR performs better than ST when there are sufficient data.

Next we study the performance of ST with both transform matrices  $A_{sk}$  and bias vectors  $b_{sl}$ . The performance is compared with GMM-SVM in which the SVM supervector is composed of mean vectors of 512 Gaussians. By setting  $L = 512$ , the SVM supervector in ST-SVM has  $KD^2$  more elements which come from the  $A_{sk}$  parameters. Table 2 shows EERs of the two methods on male-speaker data. When  $K$  equals 1 or 2, ST-SVM outperforms GMM-SVM in most conditions except that slight performance fluctuation appears in the condition of 8c-1c. Results also show that ST-SVM yields greater improvement in shorter duration conditions. When  $K$  is equal to or bigger than 4, the ST-SVM performance degrades especially in

Table 1: Equal error rates (in %) of MLLR-SVM and ST-SVM ( $L = 0$ ) on male-speaker data.

Method	K	10s-10s	1c-10s	8c-10s	1c-1c	8c-1c
MLLR -SVM	1	38.19	23.96	13.23	5.77	3.13
	2	36.80	23.83	12.70	5.16	2.90
	4	38.38	25.31	14.53	5.02	3.23
	8	41.16	27.27	15.65	6.19	2.78
ST-SVM ( $L = 0$ )	1	21.46	13.76	11.11	7.13	7.42
	2	21.72	12.84	9.52	6.24	6.43
	4	22.73	13.14	8.83	5.71	5.14
	8	23.74	11.90	8.47	4.80	4.84

short-duration conditions, indicating that the speech data becomes insufficient for estimating the more transform matrices together with the large number of bias vectors.

Table 2: Equal error rates (in %) of GMM-SVM and ST-SVM ( $L = 512$ ) on male-speaker data.

Method	K	10s-10s	1c-10s	8c-10s	1c-1c	8c-1c
GMM-SVM	-	22.98	9.68	3.70	1.93	1.18
ST-SVM ( $L = 512$ )	1	18.82	9.09	3.17	1.90	1.18
	2	18.40	8.84	2.68	1.82	1.22
	4	20.20	10.07	3.70	1.82	0.97
	8	23.48	11.55	5.82	1.82	1.29

As Table 2 shows that ST-SVM yields good performance on male-speaker data when  $L$  equals 512 and  $K$  equals 1 or 2, we extend the study to female-speaker data and to all-speaker data. The EERs are shown in Table 3 and Table 4. Results in the two tables confirm the effectiveness of ST-SVM especially for short-duration conditions.

Table 3: Equal error rates (in %) of GMM-SVM and ST-SVM ( $L = 512$ ) on female-speaker data.

Method	K	10s-10s	1c-10s	8c-10s	1c-1c	8c-1c
GMM-SVM	-	25.05	11.13	4.49	2.31	1.29
ST-SVM ( $L = 512$ )	1	21.67	10.37	4.42	2.28	1.10
	2	22.08	10.52	4.63	2.29	1.18

## 8. Conclusions

We present a speaker recognition method using MAP estimation of subspace transforms (ST). Transform matrices and bias vectors are associated with separate regression classes, and matrices are defined as a linear combination of basis transforms. Transform parameters are estimated with MAP and modeled with SVM. The method is evaluated on the 2008 NIST SRE task. ST with transform matrices outperforms MLLR in short-duration conditions, demonstrating that the subspace estimation is more robust than the full-matrix estimation given limited training data. ST with both transform matrices and bias vectors improves the performance of GMM-SVM when the number of bias vectors in ST is the same as the number of means in GMM, showing the effectiveness of combining different numbers of transform matrices and bias vectors. In future, we will study joint estimation of basis transforms and speaker transforms, and estimation of transforms on both model space and feature space.

Table 4: Equal error rates (in %) of GMM-SVM and ST-SVM ( $L = 512$ ) on all-speaker data.

Method	K	10s-10s	1c-10s	8c-10s	1c-1c	8c-1c
GMM-SVM	-	24.18	10.70	3.98	2.16	1.16
ST-SVM ( $L = 512$ )	1	20.76	9.88	3.95	2.14	1.11
	2	20.97	9.88	3.96	2.14	1.16

## 9. References

- [1] Campbell W.M., "Generalized linear discriminant sequence kernels for speaker recognition", Proc. ICASSP, 161-164, 2002.
- [2] Campbell W.M. *et al.*, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", Proc. ICASSP, 97-100, 2006.
- [3] Campbell W.M. *et al.*, "Phonetic speaker recognition with support vector machines", Proc. NIPS, 1377-1384, 2003.
- [4] Stolcke A. *et al.*, "Speaker recognition with session variability normalization based on MLLR adaptation transforms", IEEE Trans. Audio, Speech and Lang. Proc., 15(7):1987-1998, 2007.
- [5] Leggetter C.J. and Woodland P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, 9:171-185, 1995.
- [6] Chou W., "Maximum a posteriori linear regression with elliptically symmetric matrix variate priors", Proc. Eurospeech, 1-4, 1999.
- [7] Siohan O., Chesta C., and Lee C.-H., "Hidden Markov model adaptation using maximum a posteriori linear regression", Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, 147-150, 1999.
- [8] Lei X., Hamaker J. and He X., "Robust feature space adaptation for telephony speech recognition", Proc. ICSLP, 773-776, 2006
- [9] Gauvain J.-L. and Lee C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. Speech and Audio Proc., 2(2):291-298, 1994.
- [10] Kuhn R. *et al.*, "Rapid speaker adaptation in eigenvoice space", IEEE Trans. Speech and Audio Proc., 8(6):695-707, 2000.
- [11] Kenny P. *et al.*, "A study of interspeaker variability in speaker verification", IEEE Trans. Audio, Speech and Lang. Proc., 16(5):980-988, 2008.
- [12] Olsen P.A. and Gopinath R.A., "Modeling inverse covariance matrices by basis expansion", Proc. ICASSP, 945-948, 2002.
- [13] Axelrod S. *et al.*, "Subspace constrained Gaussian mixture models for speech recognition", IEEE Trans. Speech and Audio Proc., 13(6):1144-1160, 2005.
- [14] Povey D. *et al.*, "Subspace Gaussian mixture models for speech recognition", Proc. ICASSP, 4330-4333, 2010.
- [15] Gales M.J.F., "Cluster adaptation training of hidden Markov models", IEEE Trans. Speech and Audio Proc., 8(4):417-428, 2000.
- [16] Ghoshal A. *et al.*, "A novel estimation of feature-space MLLR for full-covariance models", Proc. ICASSP, 4310-4313, 2010.
- [17] Zhu D., Ma B. and Li H., "Joint MAP adaptation of feature transformation and Gaussian mixture model for speaker recognition", Proc. ICASSP, 4045-4048, 2009.
- [18] Collobert R. and Bengio S., "SVM-Torch: support vector machines for large-scale regression problems", Journal of Machine Learning Research, 2001.
- [19] Solomonoff A., Campbell W.M. and Boardman I., "Advances in channel compensation for SVM speaker recognition", Proc. ICASSP, 629-632, 2005.
- [20] Auckenthaler R., Carey M. and Lloyd-Thomas H., "Score normalization for text-independent speaker verification systems", Digital Signal Proc., 10:42-54, 2000.