Regularized Feature-Based Maximum Likelihood Linear Regression for Speech Recognition

Mohamed Kamal Omar

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
mkomar@us.ibm.com

10.21437/Interspeech.2007-125

ABSTRACT

In many automatic speech recognition (ASR) applications, maximum likelihood linear regression (MLLR), and feature-based maximum likelihood linear regression (FMLLR) are used for speaker adaptation. This paper investigates a possible generalization of FMLLR which addresses the degradation in the performance of ASR systems due to small —possibly time-varying— perturbations of the training and the testing data. We formulate the problem as a regularized maximum likelihood linear regression problem. Based on this formulation, we describe a computationally efficient algorithm for estimating the linear regression parameters which maximize the sum of the log likelihood and the negative of a measure of the sensitivity of the estimated likelihood to these perturbations. This approach does not make any assumptions about the noise model during training and testing. We present several large vocabulary speech recognition experiments that show significant recognition accuracy improvement compared to using the speaker-adapted baseline models.

1. INTRODUCTION

In many ASR systems, adaptation techniques to compensate for speaker, channel, and environment effects are used. Examples include vocal tract length normalization [1], maximum likelihood linear regression (MLLR), and feature-based maximum likelihood linear regression (FMLLR) [2]. However even after using speaker adaptation algorithms, the residual error degrades the accuracy and the quality of the models estimated from the data. In this paper, we investigate the effect of small perturbations in the training and the testing data due to sources of variability not related to the word or the phoneme identity.

Many standard speaker adaptation techniques, like MLLR and FMLLR, maximize a log likelihood objective function. Stochastic perturbations in the log likelihood objective function can produce nonlinear variations of the resulting linear regression parameters estimator in ASR systems. For example, bagging under some constraints on the perturbations may cancel these effects as measured by either variance or mean-squared error and in other cases may amplify them [3]. In our work, we consider the possibility of generating an artificial local maximum of the log likelihood objective function by small perturbations in the training data. This may result in suboptimal estimation of the parameters, as they do not correspond to a maximum of the actual likelihood which we try to model. It is possible also that these small perturbations mask a local maximum of the actual likelihood function. To reduce the effect of these perturbations in the training and the testing data, we propose a variational approach which adds a penalty function or a regularization term to the log likelihood function which works as a measure of the sensitivity of the estimated log likelihood to small perturbations in the training and the testing data. This additional term reduces the effect of these perturbations, due to noise in the training data, on the value of the estimated linear regression parameters, as well as reducing the sensitivity of the estimated log likelihood of the testing data to perturbations in the testing data. The addition of this regularization term allows us to estimate robust utterance-specific affine transforms as well as speaker-specific transforms like in conventional FMLLR.

Our method can be formulated as an alternative way of selecting the prior density in Bayesian inference which reduces the effect of small perturbations of the data on the value of the conditional density. In this case, our work can be considered as a variant of MAP adaptation techniques [4].

We use our objective function to estimate affine transforms of the training and the testing data. This approach is related to feature-based maximum likelihood linear regression [2] and regression shrinkage and selection via the least absolute shrinkage and selection operator (LASSO) [5]. In the former, the maximum likelihood objective function is used to estimate the affine transforms of the features, and in the latter, the least square objective function is penalized by the l_1 norm of the transform coefficients.

In the next section, we formulate the problem and describe our objective criterion. In Section 3, the algorithm used in estimating the affine transform parameters to optimize our objective criterion is described. The experiments performed to evaluate the performance of our approach are described in Section 4. Finally, Section 5 contains a discussion of the results. In this paper, a subscript is used as an index of a component of a random vector, and a superscript is used as an index of a realization of the random vector. Capital letters are used to denote the random variables and the corresponding small letters to denote their realizations. Both vectors and matrices are in boldface to be distinguished from scalars.

2. PROBLEM FORMULATION

Motivated by the discussion of the previous section, we derive a measure of the sensitivity of the likelihood function to perturbations in the training and the testing data without making any assumptions about the sources of variability which cause these perturbations, and search for an estimate of the affine transform parameters which maximize a weighted sum of the log likelihood and the negative of a function monotonically proportional to this sensitivity measure. To do that, we need the following proposition.

Proposition 1 *Let $\mathbf{y} = f(\mathbf{x})$ be a continuously differentiable map of the random vector $\mathbf{x} \in \Re^n$ to $\mathbf{y} \in \Re^n$, and let $P_{\mathbf{Y}}(\mathbf{y}, \lambda)$ be its probability density function. The following relation is satisfied for every realization of the random vector $\mathbf{x}$ at which $f(.)$ is invertible with inverse $f^{-1}(.)$*

$$\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}^t}|_{\mathbf{y}=\mathbf{y}_i^t} = -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}_i^t} \mathbf{x}^{t^T} - \mathbf{J}_{if}^{t^{T^{-1}}} \quad \forall\, 0 \leq t < T, \tag{1}$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]^T$ is the parameters vector, K is the dimension of the parameters vector, $\mathbf{y}_i^t = f_i^t(\mathbf{x}^t)$ is the ith root of $f^{t^{-1}}(.)$ at $\mathbf{x}^t$, $\mathbf{x}^t$ is the tth realization of the random vector $\mathbf{x}$, $\mathbf{J}_{if}^t$ is the Jacobian matrix of the map $f^t(.)$ at the ith root $\mathbf{y}_i^t$, and T is the number of realizations of the random vector $\mathbf{x}$.

Proof: The relation between the probability density functions of $\mathbf{x}$, $P_{\mathbf{X}}(\mathbf{x})$, and the probability density functions of $\mathbf{y} = f(\mathbf{x})$, $P_{\mathbf{Y}}(\mathbf{y}, \lambda)$, is in general [6],

$$P_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{R} P_{\mathbf{Y}}(\mathbf{y}, \lambda)|_{\mathbf{y}=\mathbf{y}_i} \left| det(\mathbf{J}_{if}) \right|, \tag{2}$$

where $\mathbf{y}_i = f_i(\mathbf{x})$ is the ith root of $f^{-1}(.)$ at $\mathbf{x}$, R is the number of roots at $\mathbf{x}$, and $\mathbf{J}_{if}$ is the Jacobian matrix of the map $f(.)$ at the ith root $\mathbf{y}_i$.

Differentiating both sides with respect to $\mathbf{J}_{if}$, we get

$$\mathbf{0} = |det(\mathbf{J}_{if})| \frac{dP_{\mathbf{Y}}(\mathbf{y}, \lambda)}{d\mathbf{J}_{if}}|_{\mathbf{y}=\mathbf{y}_i} + P_{\mathbf{Y}}(\mathbf{y}, \lambda)|_{\mathbf{y}=\mathbf{y}_i} |det(\mathbf{J}_{if})| \mathbf{J}_{if}^{T^{-1}}. \tag{3}$$

Since $f^t(.)$ is invertible at $\mathbf{x}^t$, $|det(\mathbf{J}_{if})| \neq 0$ and therefore

$$\mathbf{0} = \frac{d \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{d\mathbf{J}_{if}}|_{\mathbf{y}=\mathbf{y}_i} + \mathbf{J}_{if}^{T^{-1}}. \tag{4}$$

But

$$\frac{d \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{d\mathbf{J}_{if}} = \sum_{j=1}^{n} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial y_j} \frac{\partial y_j}{\partial \mathbf{J}_{if}} + \sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}}, \tag{5}$$

and therefore

$$\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}}|_{\mathbf{y}=\mathbf{y}_i} = -\sum_{j=1}^{n} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial y_j} \frac{\partial y_j}{\partial \mathbf{J}_{if}}|_{y_j=y_{ij}} - \mathbf{J}_{if}^{T^{-1}}. \tag{6}$$

By writing the vector function $\mathbf{y}_i = f_i(\mathbf{x})$ at $\mathbf{x} = \mathbf{0}$ in terms of its vector Taylor series expansion

$$f_i(\mathbf{0}) = f_i(\mathbf{x}) - \mathbf{J}_{if}\mathbf{x} + O(\mathbf{x}^2), \tag{7}$$

where $O(\mathbf{x}^2)$ is the sum of terms of second order degree or higher in $\mathbf{x}$.

Taking the partial derivative of both sides with respect to $\mathbf{J}_{if}$, and substituting into Equation 6, we get

$$\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}}|_{\mathbf{y}=\mathbf{y}_i} = -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}_i} \mathbf{x}^T - \mathbf{J}_{if}^{T^{-1}}. \tag{8}$$

Since Equation 8 is valid for every realization of the random vector $\mathbf{x}$ at which $f(.)$ is invertible, then

$$\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_{if}^t}|_{\mathbf{y}=\mathbf{y}_i^t} = -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}_i^t} \mathbf{x}^{t^T} - \mathbf{J}_{if}^{t^{T^{-1}}} \quad \forall\, 0 \leq t < T. \tag{9}$$

Equation 9 proves the Proposition. For a nonlinear feature transformation, the Jacobian matrix of the transformation is a function of the values of the feature vectors. This makes the estimation of the Jacobian matrix at each realization of a high-dimensional input feature vector computationally expensive. A significant reduction in the computational complexity can be achieved by considering small perturbations in the training and the testing data. This is equivalent to stating that the map $f(.)$ is close to the identity map. This special case motivates using a variational approach to reduce the problem to estimating the local change in the values of the likelihood when the Jacobian matrix is very close to the identity matrix, i.e. $J_f \approx I$. This important special case is covered by the following lemma

Lemma: Let $\mathbf{x}$ be a random vector in $\Re^n$, and let $\mathbf{y} = f(\mathbf{x})$ be a continuously differentiable map of the random vector $\mathbf{x} \in \Re^n$ to $\mathbf{y} \in \Re^n$ such that $J_f \approx I$, and $P_{\mathbf{Y}}(\mathbf{y}, \lambda)$ be the probability density function of $\mathbf{y}$. The following relation is satisfied for every realization of the random vector $\mathbf{x}$

$$\sum_{k=1}^{K} \frac{\partial \log P_{\mathbf{Y}}(\mathbf{y}, \lambda)}{\partial \lambda_k} \frac{\partial \lambda_k}{\partial \mathbf{J}_f^t}|_{\mathbf{y}=\mathbf{y}^t} \approx -\frac{\partial \log P_{\mathbf{Y}}(\mathbf{y})}{\partial \mathbf{y}}|_{\mathbf{y}=\mathbf{y}^t} \mathbf{y}^{t^T} - \mathbf{I}, \tag{10}$$

where $\mathbf{y}^t$ is the tth realization of the random vector $\mathbf{y}$, $\mathbf{I}$ is the identity matrix. The proof of this lemma is straightforward by taking the limit of $\mathbf{x} \rightarrow \mathbf{y}$ in Proposition 1. The lemma proves that for every realization of the random vector $\mathbf{x}$, the local change in the value of the log likelihood due to the effect of applying the function $f^t(.)$ to $\mathbf{x}$ on the model parameters can be represented in terms of the gradient of the log likelihood function with respect to $\mathbf{y}$ and the value of $\mathbf{y}$.

In this work, we estimate utterance-specific and speaker-specific affine transforms of the feature vector such that

$$\mathbf{z} = \mathbf{A}\mathbf{y} + \mathbf{b}, \tag{11}$$

where $\mathbf{y}$ is the observed feature vector, $\mathbf{z}$ is the transformed feature vector, $\mathbf{A}$ is an $n \times n$ matrix, $\mathbf{b}$ is an $n \times 1$ vector, n is the dimension of the feature vector. The parameters of the affine transforms are estimated to maximize the weighted sum of the log likelihood

function and the negative of the sum of the squares of the l_2 norm of the vector representation of the matrix $\mathbf{D^t} = [d_{ij}^t]$ where

$$d_{ij}^t = \frac{\partial \log P_{\mathbf{Z}}(\mathbf{z})}{\partial z_i}|_{\mathbf{z}=\mathbf{z}^t} z_j^t + \delta_{ij}, \forall\, 0 \leq i < n, 0 \leq j < n, \tag{12}$$

where n is the length of the random vector $\mathbf{x}$,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

So our objective function is

$$O = \sum_{t=1}^{T} \log P_{\mathbf{Z}}(\mathbf{z}^t) + T \log |det(\mathbf{A})| + \zeta \sum_{t=1}^{T} \parallel vec[\mathbf{D}^t] \parallel_2^2, \tag{13}$$

where $\log P_{\mathbf{Z}}(\mathbf{z}^t)$ is an estimate of the log likelihood of the tth transformed feature vector, T is the number of realizations, $vec[\mathbf{D}^t]$ is the vector representation of the matrix $\mathbf{D}^t$, and ζ is a negative constant.

3. IMPLEMENTATION

In the previous section, we showed that by using a variational approach, the sensitivity of the model to the small perturbations in the training and the testing data can be reduced by adding a penalty term to the log likelihood objective function. This section therefore develops an algorithm which estimates the parameters of utterance-specific or speaker-specific affine transforms of the features to maximize the modified objective function in Equation 13. We use a gradient descent algorithm to estimate the values of the affine transforms.

The gradient of the objective function with respect to the parameters is

$$\begin{aligned} \frac{\partial O}{\partial \mathbf{A}} &= \sum_{t=1}^{T} \left[\left(\frac{\partial \log P_{\mathbf{Z}}(\mathbf{z}^t)}{\partial \mathbf{z}} + \sum_{i=1}^{n} \sum_{j=1}^{n} 2dij^t \frac{\partial d_{ij}^t}{\partial \mathbf{z}} \right) \mathbf{y}^{t^T} \right] \\ &\quad + T\mathbf{A}^{T^{-1}}, \\ \frac{\partial O}{\partial \mathbf{b}} &= \sum_{t=1}^{T} \left(\frac{\partial \log P_{\mathbf{Z}}(\mathbf{z}^t)}{\partial z} + \sum_{i=1}^{n} \sum_{j=1}^{n} 2dij^t \frac{\partial d_{ij}^t}{\partial z} \right) \end{aligned} \tag{14}$$

To avoid estimating the inverse of the transpose of an $n \times n$ matrix for each utterance in each iteration, we use the natural gradient to update the $\mathbf{A}$ matrix which is given by [7]

$$\begin{aligned} \mathbf{N_A} &= \sum_{t=1}^{T} \left[\left(\frac{\partial \log P_{\mathbf{Z}}(\mathbf{z}^t)}{\partial \mathbf{z}} + \sum_{i=1}^{n} \sum_{j=1}^{n} 2d_{ij}^t \frac{\partial d_{ij}^t}{\partial \mathbf{z}} \right) \mathbf{y}^{t^T} \right] \\ &\quad \mathbf{A}^T\mathbf{A} + T\mathbf{A}. \end{aligned} \tag{15}$$

Initially the values of the affine transform parameters , $\mathbf{W} = [\mathbf{A}, \mathbf{b}]$, are unknown, so we choose an initial value of $\mathbf{W}^0 = [\mathbf{I}, \mathbf{0}]$ and then we use the gradient descent algorithm to update the parameters to maximize the objective function. We iterate until a local maximum of the empirical objective function is found.

The update equations of the affine transform parameters are

$$\mathbf{A}^{i+1} = \mathbf{A}^i + \alpha^i \mathbf{N_A}, \tag{16}$$

$$\mathbf{b}^{i+1} = \mathbf{b}^i + \alpha^i \frac{\partial O}{\partial \mathbf{b}}, \tag{17}$$

where $\mathbf{A}^i$ is the $\mathbf{A}$ matrix after i iterations, $\mathbf{A}^{i+1}$ is the $\mathbf{A}$ matrix after $i+1$ iterations, and α^i is a step size that should be chosen small enough to guarantee convergence and large enough to reduce the number of iterations required to achieve convergence.

4. EXPERIMENTS

We apply the variational approach to regularized feature-based maximum likelihood linear regression (RFMLLR) estimation described in the previous section to two Arabic broadcast news systems. The main difference between the two systems is the phoneme set used. The phonetic transcription in Arabic requires the existence of certain diacritic symbols which are usually not found in text transcriptions, and correspond to short vowels. The unvowelized system models only the graphemes and does not model the diacritic symbols, while the vowelized system models the diacritic symbols and therefore models the short vowels in Arabic explicitly. For the two systems, each phoneme is represented by 3 HMM states with left-to-right topology with the exception of modeling short vowels with 2 states in the Arabic vowelized system. For the two systems, the raw features are 13-dimensional PLP features computed every 10 ms. from 25-ms. frames. The recognition features are computed from the raw features by splicing together nine frames of raw features (± 4 frames around the current frame), projecting the 117-dim. spliced features to 40 dimensions using a linear discriminant analysis (LDA) projection, and then applying maximum likelihood linear transformation (MLLT) to the 40-dim. projected features to reduce the mismatch between the statistics of the final features and the constraints of the diagonal-covariance Gaussian mixtures that model the HMM observation densities.

The decoding for both speaker-adapted Arabic broadcast news systems consists of two passes: the first speaker-independent pass output is used to adapt the models, and the second decoding pass uses the adapted models to generate the final output of the decoder. In the context of speaker-adaptive training to produce canonical acoustic models, we use vocal tract length normalization (VTLN), and feature-space MLLR. For the feature-based minimum phone error (FMPE) baseline, an FMPE transform is applied on top of the speaker-specific FMLLR transforms. We do also a single pass of MLLR adaptation, using a regression tree to generate transforms for different sets of mixture components. For both systems, we test two different setups. In the first setup, we use the regularized objective function to replace the log likelihood objective function in FMLLR speaker-specific adaptation. We call this setup RFMLLR. In the second setup, the regularized objective function is used to estimate utterance-specific affine transforms of the test data, on top of the FMLLR transforms for the FMLLR baseline system and on top of the FMPE transform for the FMPE baseline system, without updating the acoustic model. We call this setup URFMLLR.

For the unvowelized Arabic broadcast news system, the acoustic model consists of 5032 context-dependent states and 400K diagonal-covariance Gaussian mixtures. The language model is a 617K vocabulary interpolated back-off 4-gram language model. The vowelized acoustic model consists of 4006 context-dependent states and 400K diagonal-covariance Gaussian mixtures. We report results of all systems on both the Arabic DARPA 2004 Rich Transcription (RT04) evaluation data and the 2005 Arabic broadcast news tune test set (BNAT05).

As shown in Table 1, the WER results for the FMLLR unvowelized system improved by $\sim 2.0\%$ relative for the RT04 eval-

System	RT04	BNAT05
FMLLR Baseline	17.2	18.9
FRMLLR	16.9	18.7
UFRMLLR	16.7	17.9

Table 1. Word error rates (%) on the Arabic RT04 and BNAT05 evaluation data using the FMLLR unvowelized system

System	RT04	BNAT05
FMPE Baseline	14.4	15.4
RFMLLR+FMPE	14.2	15.1
FMPE+URFMLLR	14.1	14.8

Table 2. Word error rates (%) on the Arabic RT04 and BNAT05 evaluation data using the FMPE unvowelized system

uation data and for the BNAT05 evaluation data compared to the FMLLR baseline by using the regularized estimation of the HMM model and the speaker-specific feature-based linear regression parameters. An improvement in WER results of $\sim 3.0\%$ on the RT04 evaluation data and of $\sim 5.3\%$ on the BNAT05 evaluation data is achieved by using the regularized objective function to train utterance-specific affine transforms of the test data on top of the FMLLR transforms.

As shown in Table 2, the WER results for the FMPE unvowelized system improved by $\sim 2.0\%$ relative for the RT04 evaluation data and for the BNAT05 evaluation data compared to the FMPE baseline system by using the regularized estimation of the speaker-specific feature-based linear regression parameters. An improvement in WER results of $\sim 2.0\%$ on the RT04 evaluation data and of $\sim 3.9\%$ on the BNAT05 evaluation data is achieved by using the regularized objective function to train utterance-specific affine transforms of the test data on top of the FMPE transform.

As shown in Table 3, the WER results for the FMLLR vowelized system improved by $\sim 2.0\%$ relative for the RT04 evaluation data and for the BNAT05 evaluation data compared to the FMLLR baseline by using the regularized estimation of the HMM model and the speaker-specific feature-based linear regression parameters. An improvement in WER results of $\sim 3.4\%$ on the RT04 evaluation data and of $\sim 3.0\%$ on the BNAT05 evaluation data is achieved by using the regularized objective function to train utterance-specific affine transforms of the test data on top of the FMLLR transforms.

As shown in Table 4, the WER results for the FMPE vowelized system improved by $\sim 1.4\%$ relative for the RT04 evaluation data and for the BNAT05 evaluation data compared to the FMPE baseline system by using the regularized estimation of the speaker-specific feature-based linear regression parameters instead. An improvement in WER results of $\sim 2.3\%$ on the RT04 evaluation data and of $\sim 4.2\%$ on the BNAT05 evaluation data is achieved by using the regularized objective function to train utterance-specific affine transforms of the test data on top of the FMPE transform.

System	RT04	BNAT05
FMLLR Baseline	14.9	16.4
RFMLLR	14.6	16.1
URFMLLR	14.4	15.9

Table 3. Word error rates (%) on the Arabic RT04 and BNAT05 evaluation data using the vowelized FMLLR system

System	RT04	BNAT05
FMPE Baseline	12.9	14.4
RFMLLR+FMPE	12.8	14.2
FMPE+URFMLLR	12.6	13.8

Table 4. Word error rates (%) on the Arabic RT04 and BNAT05 evaluation data using the FMPE vowelized system

5. RESULTS AND DISCUSSION

This work proposes a variational approach to regularized feature-based maximum likelihood linear regression for speaker adaptation of speech recognition systems. This approach calculates a measure of the sensitivity of the model to small variations in the training and testing data. The negative of this measure is added to the log likelihood objective function. The proposed regularized objective function is used by a gradient descent algorithm to estimate the parameters of the affine transforms.

The experiments show small improvement from using the regularized objective function instead of the maximum likelihood objective function in conventional speaker-specific FMLLR training. The experiments show also consistent improvement in recognition accuracy achieved by adding utterance-specific affine transforms trained using the regularized objective function on top of FMLLR or FMPE transforms compared to the corresponding baseline system.

6. REFERENCES

[1] Puming Zhan, and Martin Westphal, "Speaker Normalization Based On Frequency Warping," *IEEE Proceedings of ICASSP*, Munich, Germany, 1997.

[2] Mark J. F. Gales, *Maximum Likelihood Linear Transformation For HMM-Based Speech Recognition,* Technical Report, Cambridge University Engineering Department, May 1997.

[3] Jerome H. Friedman, and Peter Hall, "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference*, 2007, pp. 669–683.

[4] O. Siohan, C. Chesta and C. Lee, "Hidden Markov Model Adaptation Using Maximum A Posteriori Linear Regression," *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

[5] R. Tibshirani, "Regression Shrinkage and Selection via The Lasso," *Journal of Royal Statistical Society Bulletin*, Vol. 58, No. 1, 1996, pp. 267–288.

[6] Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill, New York, 1991.

[7] Shun-ichi Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation*, Vol. 10, No. 2, 1998, pp. 251–276.