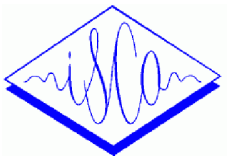

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Group Sparse Hidden Markov Models for Speech Recognition

Jen-Tzung Chien and Cheng-Chun Chiang

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC
jtchien@ieee.org

10.21437/Interspeech.2012-508

Abstract

This paper presents the group sparse hidden Markov models (GS-HMMs) where a sequence of acoustic features is driven by Markov chain and each feature vector is represented by two groups of basis vectors. The group of common bases represents the features across states within a HMM. The group of individual bases compensates the intra-state residual information. Importantly, the sparse prior for sensing weights is controlled by the Laplacian scale mixture (LSM) distribution which is obtained by multiplying Laplacian variable with an inverse Gamma variable. The scale mixture parameter in LSM makes the distribution even sparser. This parameter serves as an automatic relevance determination for selecting the relevant bases from two groups. The weights and two sets of bases in GS-HMMs are estimated via Bayesian learning. We apply this framework for acoustic modeling and show the robustness of GS-HMMs for speech recognition in presence of different noises types and SNRs.

Index Terms: Bayesian learning, group sparsity, hidden Markov model, speech recognition

1. Introduction

Sparse representation is a fascinating research topic in pattern recognition and machine learning which has been broadly developing for ubiquitous applications. This topic aims to find a sparse measurement based on a set of over-determined basis vectors and use a relatively small set of relevant bases to represent target data. The overfitting problem is alleviated. However, most of previous methods did not deal with this issue for sequential data. Recently, Bayesian sensing hidden Markov models (HMMs) [7] were proposed for large vocabulary continuous speech recognition (LVCSR) by incorporating the sparse Bayesian learning (SBL) into HMMs based on the relevance vector machine (RVM) [8]. Very competing LVCSR performance has been achieved [6]. However, RVM did not truly apply sparse prior for basis representation. This paper presents a group sparse representation of sequential data and adopts the *Laplacian scale mixture* (LSM) distribution [5] as sparse prior for SBL of acoustic models. The sensing weights and basis vectors are estimated according to the maximum *a posteriori* (MAP) principle. The automatic relevance determination (ARD) [7][8] is performed to select relevant bases for feature representation. We build the group sparse HMMs (GS-HMMs) for speech recognition where the common bases and individual bases [4] are estimated to represent the inter-state variations and the intra-state residual information, respectively.

2. Laplacian Sparse Representation

Considering a single feature vector $\mathbf{x} \in \mathcal{R}^D$ which is generated from a set of over-determined basis vectors $\Phi = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N]$ via $\mathbf{x} = \Phi\mathbf{w}$ with an $D \times N$ matrix Φ and a weight vector $\mathbf{w} \in \mathcal{R}^N$ where $N > D$. A typical solution to such an ill-posed problem is via a ℓ_1-regularized objective function with a regularization parameter ρ by solving $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{x} - \Phi\mathbf{w}\|_2^2 + \rho\|\mathbf{w}\|_1$. This objective function is comparable with MAP criterion based on a likelihood function $p(\mathbf{x}|\mathbf{w}) = \mathcal{N}(\mathbf{x}|\Phi\mathbf{w}, \sigma^2 = 1)$ and a Laplace prior $p(\mathbf{w}) = (\rho/2)\exp(-\rho\|\mathbf{w}\|_1)$ which is a sparse prior with scale ρ. Bayesian compressive sensing using Laplace prior was addressed in [1]. It is well-known that sparse distribution has a sharp peak with heavy tail and is robust to outliers.

2.1. Laplacian scale mixture (LSM) distribution

In [5], an LSM distribution was shown to be even sparser than Laplace distribution. The LSM distribution of the nth weight w_n is obtained by a transformation $w_n = \eta_n^{-1}u_n$ where u_n has a Laplace distribution $p(u_n) = \frac{1}{2}\exp(-|u_n|)$ with scale 1 and the multiplier η_n is positive and has a Gamma distribution $p(\eta_n) = (\beta^\alpha/\Gamma(\alpha))\eta_n^{\alpha-1}\exp(-\beta\eta_n)$ with hyperparameters α and β. Assuming u_n and η_n are independent, w_n turns out to be Laplacian given a scale mixture η_n, i.e. $p(w_n|\eta_n) = \frac{\eta_n}{2}\exp(-\eta_n|w_n|)$. LSM distribution is a marginal distribution $p(w_n)$ which is derived by marginalizing over continuous mixture parameter η_n

$$\int_0^\infty p(w_n|\eta_n)p(\eta_n)d\eta_n = \frac{\alpha\beta^\alpha}{2(\beta + |w_n|)^{\alpha+1}}. \quad (1)$$

2.2. Laplacian group sparse coding

We address the Laplacian group sparse coding for single observation $\mathbf{x}$. Since the logarithm of LSM distribution in (1) is not analytical, it is difficult to directly compute MAP estimate of $\mathbf{w}$. This issue could be tackled by computing the complete distribution $p(\mathrm{w}, \boldsymbol{\eta})$ and treating $\boldsymbol{\eta}$ as a latent variable. Finding MAP estimate is comparable to apply the Jensen's inequality or fulfil the EM algorithm where the lower bound $\mathcal{L}(q, \mathrm{w})$ (i.e. right-hand-side of (2)) on log posterior distribution

$$\log p(\mathrm{w}|\mathrm{x}) \geq \log p(\mathrm{x}|\mathrm{w}) + \int q(\boldsymbol{\eta}) \log \frac{p(\mathrm{w}, \boldsymbol{\eta})}{q(\boldsymbol{\eta})} d\boldsymbol{\eta} \quad (2)$$

is maximized. In E-step, we calculate the expectation function and find the variational distribution $q(\boldsymbol{\eta})$ that is closest to $p(\mathrm{w}, \boldsymbol{\eta})$ via $q^{(t+1)}(\boldsymbol{\eta}) = \arg\max_q \mathcal{L}(q, \mathrm{w}^{(t)})$. In M-step, the lower bound is further maximized to estimate the weights via $\mathrm{w}^{(t+1)} = \arg\max_{\mathrm{w}} \mathcal{L}(q^{(t+1)}, \mathrm{w})$. E-step is reduced to find $q^{(t+1)}(\boldsymbol{\eta}) = p(\boldsymbol{\eta}|\mathrm{w}^{(t)})$ which happens in case of equality in (2). M-step is equivalent to find new MAP estimate $\mathbf{w}^{(t+1)}$ by minimizing

$$\frac{1}{2}\|\mathbf{x} - \Phi\mathrm{w}\|_2^2 + \sum_{n=1}^{N} \langle \eta_n \rangle_{q^{(t+1)}} |w_n| \quad (3)$$

where $\langle \cdot \rangle_q$ denotes the expectation with respect to $q(\eta_n)$. In (3), the 2^{nd} term comes from the 2^{nd} term of lower bound in (2) which is seen as an expectation function of $\log p(w_n|\eta_n)$ with respect to $q(\eta_n)$. Notably, (3) is viewed as a *weighted* ℓ_1-regularized objective function. The expected scale mixture $\langle \eta_n \rangle_q$ acts as a weight factor. Since Gamma and Laplace distributions are *conjugate*, the posterior distribution is also Gamma. Here, $p(\eta_n)$ is a Gamma prior with hyperparameters α and β and $p(w_n|\eta_n)$ is a Laplace likelihood. The resulting posterior distribution $p(\eta_n|w_n)$ is Gamma with hyperparameters $\alpha + 1$ and $\beta + |w_n|$. The expected scale mixture $\langle \eta_n \rangle_{p(\eta_n|w_n^{(t)})}$ is calculated as the mean of Gamma distribution $p(\eta_n|w_n)$ which is given by $\frac{\alpha+1}{\beta+|w_n^{(t)}|}$. In sparse coding, this weight factor can be shared for a group of sensing weights or separate for individual weights [5]. Group sparse coding [2] can be performed. Assuming that weight w_n is associated with group Ω_g with N_g members, we calculate the expected scale mixture for Ω_g by

$$\langle \eta_g \rangle_{p(\eta_g|w_{\Omega_g}^{(t)})} \triangleq \eta_{(g)}^{(t+1)} = \frac{\alpha + N_g}{\beta + \sum_{n\in\Omega_g} |w_n^{(t)}|} . \quad (4)$$

This group sparse coding could not directly work for speech recognition because it only deals with single-pattern coding.

3. Group Sparse Hidden Markov Models

In this study, we construct GS-HMMs for representation of sequential data $X = \{\mathbf{x}_t\}_{t=1}^T$ and apply them for noisy speech recognition. Each feature vector $\mathbf{x}_t$ at frame t is generated from a Markov chain with state parameters consisting of common bases $\Phi_c = [\boldsymbol{\phi}_{c1}, \cdots, \boldsymbol{\phi}_{cN_c}]$ which are shared for different states within a HMM for context-dependent phone unit and also individual bases $\Phi_{si} = [\boldsymbol{\phi}_{si1}, \cdots, \boldsymbol{\phi}_{siN_s}]$ which are separate for individual state $s_t = i$. The corresponding weights are denoted by $\mathrm{w}_t = [\mathrm{w}_{ct}^T \ \mathrm{w}_{st}^T]^T = [w_{ct1}, \cdots, w_{ctN_c}, w_{st1}, \cdots, w_{stN_s}]^T$.

3.1. MAP parameter estimation

The group representation of speech features can be seen as a subspace approach [3]. The common bases span the *principal subspace* for representing the principal information within a HMM while the state-dependent individual bases span the *minor subspace* for catching the residual information due to an individual state. Two sets of sensing weights $\{\mathrm{w}_{ct}, \mathrm{w}_{st}\}$ should be estimated separately. Assuming that basis vectors $\{\Phi_c, \Phi_{si}\}$ are Gaussian distributed with zero mean and precision parameters $\{\gamma_c, \gamma_{si}\}$, GS-HMM parameters $\lambda = \{\mathrm{w}_{ct}, \mathrm{w}_{st}, \Phi_c, \Phi_{si}\}$ are estimated by maximizing the posterior likelihood $p(\lambda|X)$ through EM algorithm. In E-step, the expectation function $\mathcal{R}(\lambda|\lambda^{(k)}) = E\{-\log p(\lambda, S|X)|X, \lambda^{(k)}\}$ of the negative log posterior of new estimate λ given old estimate $\lambda^{(k)}$ at iteration k is computed over all states $S = \{s_t\}$ by

$$\begin{aligned} \sum_i \Bigg\{ \sum_{t=1}^{T} \xi_t^{(k)}(i) \Bigg[\frac{1}{2}\|\mathrm{x}_t - \Phi_c\mathrm{w}_{ct} - \Phi_{si}\mathrm{w}_{st}\|^2 \\ + \sum_{n=1}^{N_c} \eta_{cn}|w_{ctn}| + \sum_{p=1}^{N_s} \eta_{sp}|w_{ctp}| \Bigg] \\ + \frac{\gamma_{si}}{2} \sum_{p=1}^{N_s} \|\boldsymbol{\phi}_{sip}\|^2 \Bigg\} + \frac{\gamma_c}{2} \sum_{n=1}^{N_c} \|\boldsymbol{\phi}_{cn}\|^2 \end{aligned} \quad (5)$$

where $\xi_t^{(k)}(i) = p(s_t = i|X, \lambda^{(k)})$ is the posterior probability of being in state i at time t. In M-step, we minimize (5) with respect to four GS-HMM parameters.

3.2. Estimation of sensing weights

To estimate the sensing weight w_{ctn} for the nth common basis at frame t, we apply the coordinate descent method [9] and express the objective function $\mathcal{R}(w_{ctn}|\lambda^{(k)})$ based on all terms related to w_{ctn} as follows

$$\begin{aligned} \sum_i \Bigg\{ \sum_{t=1}^{T} \xi_t^{(k)}(i) \Bigg[\frac{1}{2}\|\mathrm{x}_t - \sum_{m\neq n}^{N_c} w_{ctm}\boldsymbol{\phi}_{cm} \\ - w_{ctn}\boldsymbol{\phi}_{cn} - \Phi_{si}\mathrm{w}_{st}\|^2 + \eta_{cn}|w_{ctn}| \Bigg] \Bigg\}. \end{aligned} \quad (6)$$

By expanding the first quadratic term, we rewrite (6) by

$$\sum_i \Bigg\{ \sum_{t=1}^{T} \xi_t^{(k)}(i) \Bigg[\sum_{m\neq n}^{N_c} w_{ctm} w_{ctn} (\boldsymbol{\phi}_{cm} \cdot \boldsymbol{\phi}_{cn}) + \sum_{p=1}^{N_s} w_{stp} w_{ctn} (\boldsymbol{\phi}_{sip} \cdot \boldsymbol{\phi}_{cn}) - w_{ctn} (\mathrm{x}_t \cdot \boldsymbol{\phi}_{cn}) + \frac{1}{2} w_{ctn}^2 \|\boldsymbol{\phi}_{cn}\|^2 + \eta_{cn} |w_{ctn}| \Bigg] \Bigg\} . \quad (7)$$

After taking differentiation of (7) with respect to w_{ctn} and setting it to zero, we derive

$$w_{ctn}^{(k+1)} = \frac{-b_{ct} \mp N_s \eta_{cn}}{N_t^{(k)} \|\boldsymbol{\phi}_{cn}\|^2} . \quad (8)$$

In (8), $N_t^{(k)} \triangleq \sum_i \xi_t^{(k)}(i)$, N_s is the number of states in a HMM and $b_{ct} \triangleq N_t^{(k)} \sum_{m\neq n}^{N_c} w_{ctm} (\boldsymbol{\phi}_{cm} \cdot \boldsymbol{\phi}_{cn}) + \sum_i \xi_t^{(k)}(i) \sum_{p=1}^{N_s} w_{stp} (\boldsymbol{\phi}_{sip} \cdot \boldsymbol{\phi}_{cn}) - N_t^{(k)} (\mathrm{x}_t \cdot \boldsymbol{\phi}_{cn})$ which is a constant when estimating w_{ctn}. The parameter $w_{ctn}^{(k+1)}$ is calculated by using observation x_t and current estimates w_{st}, $\{w_{ctm}\}_{m\neq n}$, Φ_c and Φ_{si} given with state occupation probability $\xi_t^{(k)}(i)$.

Similarly, the estimation of sensing weight w_{stp} of x_t for the pth individual basis is derived by maximizing the auxiliary function $\mathcal{R}(w_{stp}|\lambda^{(k)})$ with respect to w_{stp}. By expanding the quadratic term in $\mathcal{R}(w_{stp}|\lambda^{(k)})$ and taking differentiation of $\mathcal{R}(w_{stp}|\lambda^{(k)})$ with respect to w_{stp}, we find that MAP sensing weight $w_{stp}^{(k+1)}$ satisfies

$$w_{stp}^{(k+1)} = \frac{-b_{st} \mp N_s \eta_{sp}}{\sum_i \xi_t^{(k)}(i) \|\boldsymbol{\phi}_{sip}\|^2} \quad (9)$$

where $b_{st} \triangleq \sum_i \xi_t^{(k)}(i) \sum_{n=1}^{N_c} w_{ctn} (\boldsymbol{\phi}_{cn} \cdot \boldsymbol{\phi}_{sip}) + \sum_i \xi_t^{(k)}(i) \sum_{m\neq p}^{N_s} w_{stm} (\boldsymbol{\phi}_{sim} \cdot \boldsymbol{\phi}_{sip}) - \sum_i \xi_t^{(k)}(i) (\mathbf{x}_t \cdot \boldsymbol{\phi}_{sip})$. In (8) and (9), the sign in the numerator is determined by the sign of current estimate w_{ctn} or w_{stp}.

3.3. Estimation of basis vectors

The groups of basis vectors $\{\Phi_c, \Phi_{si}\}$ are also estimated by MAP principle. The basis vectors Φ_c and Φ_{si} span the subspaces to compensate inter-state and intra-state variations, respectively. The auxiliary function $\mathcal{R}(\boldsymbol{\phi}_{cn}|\lambda^{(k)})$ for the nth common basis $\boldsymbol{\phi}_{cn}$ is constructed and maximized to obtain

$$\underbrace{\sum_i \sum_{t=1}^{T} \xi_t^{(k)}(i) \left[\sum_{m\neq n}^{N_c} w_{ctm} w_{ctn} \boldsymbol{\phi}_{cm} + \sum_{p=1}^{N_s} w_{stp} w_{ctn} \boldsymbol{\phi}_{sip} - w_{ctn} \mathbf{x}_t \right]}_{\mathbf{b}_{cn}} + \sum_i \sum_{t=1}^{T} \xi_t^{(k)}(i) w_{ctn}^2 \boldsymbol{\phi}_{cn} + \gamma_c \frac{\boldsymbol{\phi}_{cn}}{\|\boldsymbol{\phi}_{cn}\|} = 0 . \quad (10)$$

MAP solution to the nth common basis is derived by

$$\boldsymbol{\phi}_{cn}^{(k+1)} = - \left[\sum_i \sum_{t=1}^{T} \xi_t^{(k)}(i) w_{ctn}^2 + \frac{\gamma_c}{\|\boldsymbol{\phi}_{cn}\|} \right]^{-1} \mathbf{b}_{cn} . \quad (11)$$

Following the same manner, MAP solution to the pth individual basis $\boldsymbol{\phi}_{sip}$ can be formulated as

$$\boldsymbol{\phi}_{sip}^{(k+1)} = - \left[\sum_{t=1}^{T} \xi_t^{(k)}(i) w_{stp}^2 + \frac{\gamma_{si}}{\|\boldsymbol{\phi}_{sip}\|} \right]^{-1} \mathbf{b}_{sip} \quad (12)$$

where $\mathbf{b}_{sip} \triangleq \sum_{t=1}^{T} \xi_t^{(k)}(i) \left[\sum_{n=1}^{N_c} w_{ctn} w_{stp} \boldsymbol{\phi}_{cn} + \sum_{m\neq p}^{N_s} w_{stm} w_{stp} \boldsymbol{\phi}_{sim} - w_{stp} \mathbf{x}_t + w_{stp}^2 \boldsymbol{\phi}_{sip} \right]$.

In our implementation, the state occupation probability $\xi_t^{(k)}(i)$ is determined by Viterbi decoding algorithm. GS-HMM parameters $\lambda^{(k+1)} = \{\Phi_c^{(k+1)}, \Phi_{si}^{(k+1)}\}$ are trained with a pool of speech frames and frame-based sensing weights $\{\mathrm{x}_t, \mathrm{w}_{ct}^{(k+1)}, \mathrm{w}_{st}^{(k+1)}\}$. In test session, the sensing weights are calculated for each test frame based on pre-trained basis vectors. During Viterbi decoding and searching, the likelihood function is based on the reconstruction error function calculated by *Laplacian sparse weights* and *group basis vectors*. The hyperparameters $\{\eta_{cn}, \eta_{sp}, \gamma_c, \gamma_{si}\}$ are selected from validation data with minimum reconstruction errors.

4. Experiments

4.1. Experimental setup

In the experiments, we carried out GS-HMMs for noisy speech recognition by using Aurora2 database consisting of English digits under different noise environments. Three test sets (sets A, B and C) with ten noise types and six signal-to-noise-ratio (SNR) conditions (-5, 0, 5, 10, 15 and 20 dBs) and a clean condition were used. Acoustic models in clean and multiconditional conditions were estimated for evaluation. There were 8440 clean training utterances. The multiconditional training was performed by using clean speech as well as noisy speech by adding four noise types (subway, babble, car and exhibition) of set A under SNR being 5, 10, 15 and 20 dBs. Noise materials were unseen in test data. In speech recognition system, we extracted 39-dimensional feature vectors containing 12 Mel-frequency cepstral coefficients and log energy, along with their first- and second-order derivatives. We specified 16 states for each word and three Gaussians for each state. The whole-word continuous-density HMMs were trained as the baseline system. Detailed experimental setup was addressed in [3]. The numbers of common and individual bases were set to be 30 and 25, respectively. In clean training condition (denoted by CT), the recognition rates were averaged over four noise types in set A for each SNR. In multiconditional training condition (denoted by MT), the recognition rates were averaged over ten noise types for each SNR.

4.2. Evaluation for sparsity of sensing weights

We first investigate the effect of sparsity of the estimated sensing weights in GS-HMMs. The sparsity is defined by $\text{sparsity}(\mathbf{w}) \triangleq \frac{\sqrt{N}-(\sum_n |w_n|)/\sqrt{\sum_n w_n^2}}{\sqrt{N}-1}$ which is between 0 and 1. Figures 1 and 2 display the histograms of sparsity of $\mathrm{w}_t = [\mathrm{w}_{ct}^T \ \mathrm{w}_{st}^T]^T$ for English digit 'o' under different HMM states and training conditions. The sparsity in transition states (1st and 16th states) is larger than that in middle state (7th state). The sparsity in MT condition is larger that that in CT condition. It is meaningful that the weights estimated in heterogeneous conditions are sparser than those in homogeneous conditions.

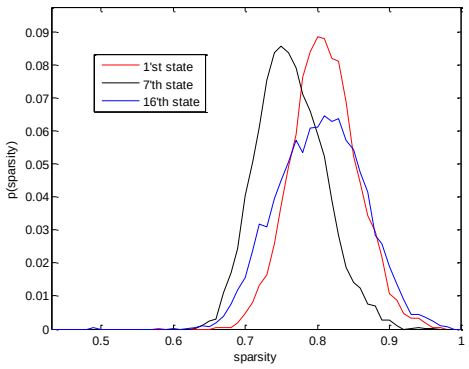

Figure 1: Histogram of the sparsity of weights under clean training condition with different states.

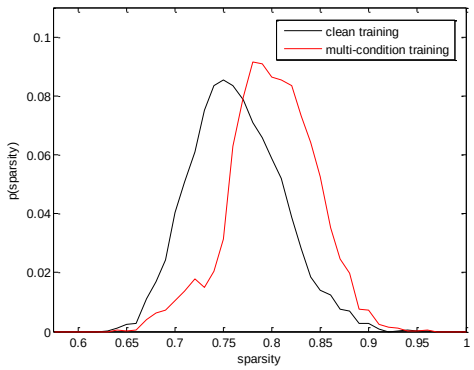

Figure 2: Histogram of the sparsity of weights under the 7th state of digit 'o' with different training conditions.

4.3. Evaluation for recognition results

Table 1 compares the recognition accuracies of baseline HMMs and GS-HMMs under different SNRs with CT and MT conditions. For both training conditions, the lower SNRs the test utterances are observed, the larger improvement the GS-HMMs obtain. The HMMs trained under MT condition attain higher recognition accuracy than those under CT condition. The averaged accuracy of noisy speech recognition is significantly increased from 62.9% to 71.0% in case of CT condition and is also improved from 81.3% to 86.2% in case of MT condition. Further improvement could be achieved by incorporating other noise robust schemes into GS-HMMs.

Table 1: Recognition accuracies (%) for HMMs and GS-HMMs with different training conditions and test SNRs

	-5dB	0dB	5dB	10dB	15dB	20dB	Clean
HMM(CT)	11.1	24.0	51.5	73.7	86.6	94.2	99.0
GS-HMM(CT)	24.2	36.0	62.2	83.9	94.2	97.7	99.0
HMM(MT)	27.5	63.0	88.3	95.6	97.6	98.3	98.9
GS-HMM(MT)	45.0	75.4	91.1	96.3	98.0	98.6	98.9

5. Conclusions

We have presented a new sparse Bayesian learning method for group basis representation of sequential data and successfully applied it for noisy speech recognition. The groups of common bases and individual bases and their sensing weights were estimated to represent the inter-state common information and the intra-state residual information, respectively. The Laplace distribution with a scale mixture parameter was adopted as the sparse prior for MAP estimation of frame-based sensing weights. The MAP estimation of HMM-dependent and state-dependent bases/dictionaries was derived according to EM procedure. The experiments on noisy speech recognition illustrated the robustness of GS-HMMs under different noise types, SNRs and training conditions.

6. References

[1] Babacan, S. D., Molina, R. and Katsaggelos, A. K., "Bayesian compressive sensing using Laplace priors", *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53-63, 2010.

[2] Bengio, S., Pereira, F., Singer, Y., and Strelow, D., "Group sparse coding", *Advances in Neural Information Processing Systems*, pp. 8289, 2009.

[3] Chien, J.-T. and Ting, C.-W., "Factor analyzed subspace modeling and selection", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 239-248, 2008.

[4] Chien, J.-T. and Hsieh, H.-L., "Bayesian group sparse learning for nonnegative matrix factorization" in *Proc. of Interspeech*, 2012.

[5] Garrigues, P. J. and Olshausen, B. A., "Group sparse coding with a Laplacian scale mixture prior", *Advances in Neural Information Processing Systems*, 2010.

[6] Saon, G. and Chien, J.-T., "Some properties of Bayesian sensing hidden Markov models", in *Proc. of IEEE ASRU Workshop*, pp. 65-70, 2011.

[7] Saon, G. and Chien, J.-T., "Bayesian sensing hidden Markov models", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 43-54, 2012.

[8] Tipping, M. E., "Sparse Bayesian learning and the relevance vector machine", *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001.

[9] Wu, T. T. and Lange, K., "Coordinate descent algorithm for lasso penalized regression", *Annals of Applied Statistics*, vol. 2, no. 1, pp. 224-244, 2008.