LASSO ENVIRONMENT MODEL COMBINATION FOR ROBUST SPEECH RECOGNITION

Xiong Xiao[1], *Jinyu Li*[2], *Eng Siong Chng*[1,3], *Haizhou Li*[1,3,4]

[1]Temasek Lab@NTU, Nanyang Technological University, Singapore
[2]Microsoft Corporation, USA
[3]School of Computer Engineering, Nanyang Technological University, Singapore
[4]Department of Human Language Technology, Institute for Infocomm Research, Singapore

xiaoxiong@ntu.edu.sg, jinyli@microsoft.com, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

In this paper, we propose a novel acoustic model adaptation method for noise robust speech recognition. Model combination is a common way to adapt acoustic models to a target test environment. For example, the mean supervectors of the adapted model is obtained as a linear combination of mean supervectors of many pre-trained environment-dependent acoustic models. Usually, the combination weights are estimated using a maximum likelihood (ML) criterion and the weights are nonzero for all the mean supervectors. We propose to estimate the weights by using Lasso (least absolute shrinkage and selection operator) which imposes an L_1 regularization term in the weight estimation problem to shrink some weights to exactly zero. Our study shows that Lasso usually shrinks to zero the weights of those mean supervectors not relevant to the test environment. By removing some nonrelevant supervectors, the obtained mean supervectors are found to be more robust against noise distortions. Experimental results on Aurora-2 task show that the Lasso-based mean combination consistently outperforms ML-based combination.

Index Terms— noise robust speech recognition, model adaptation, L_1 regularization, Lasso regression, model combination.

1. INTRODUCTION

The performance of automatic speech recognition (ASR) degrades significantly when the training and test environment conditions are different, e.g. recognizing noisy speech using clean trained acoustic model. The key to improve the robustness of ASR is to reduce the mismatch between the training and test conditions. Common approaches include feature compensation/normalization methods [1, 2] which make the clean and noisy features similar to each other and model adaptation methods (e.g. [3, 4]) which adapt the clean acoustic model towards the noisy test condition.

Another common practice in noise robust ASR is to use multi-style training (MST) [5] which uses speech data from many environments to train a single acoustic model. Since this MST model needs to cover different kinds of environments, it doesn't perform very well in each individual test environment. One solution is to build a set of acoustic models, each modeling one specific environment. During recognition, all these models are combined together, usually with the maximum likelihood (ML) criterion, to construct a target model to recognize the current test utterance. Methods belonging to this category include eigenvoice [6] which builds the target model by linearly combining basis mean supervectors.

The problem of ML model combination is that usually all combination weights are not zero, i.e., every environment-dependent model contributes to the final model. This is obviously not optimal if the test environment is exactly the same as one of the training environment. There is also such a scenario that the testing environment can be well approximated by interpolating only several training environments. Including unrelated models into the construction brings unnecessary distortion to the target model.

In this paper, we propose an environment model combination method based on Lasso (least absolute shrinkage and selection operator) [7] which uses an L_1 regularization term to regularize the combination weights. The L_1 regularization term shrinks some weights to exactly zero. In this way, the target model can be combined from the related environment models, without the distortion brought from unrelated environment models. This is demonstrated by experiments on Aurora-2 test set A, which has the matched testing environment as training. Moreover, experiments on all the Aurora-2 test sets show that the Lasso-based model combination is consistently better than the ML-based model combination.

The organization of this paper is as follows. In section 2, the model combination approach to robust speech recognition is described and both the ML and Lasso estimation of the combination weights are presented. In section 3, the estimated weights are investigated and recognition performance is evaluated on Aurora-2 task. Finally, we conclude in section 4.

2. LASSO-BASED MODEL COMBINATION

2.1. Model Adaptation by Model Combination

Assume there are N acoustic models, each representing a specific environment condition. Furthermore, assume that these environment models are adapted from a single seed model (e.g. a model trained in multi-style) by adapting the mean vectors while keeping the covariance matrices unchanged. During recognition, to adapt the acoustic model to the test condition, the mean supervector of the adapted model is obtained as a linear combination of the mean supervectors of the environment-dependent models:

$$\mathbf{s}' = \sum_{i=1}^{N} w_i \mathbf{s}_i \qquad (1)$$

where $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\}$ is the set of environment-dependent mean supervectors, w_i is the weight of $\mathbf{s}_i$, and $\mathbf{s}_i$ is obtained by concatenating the mean vectors of the i^{th} acoustic model. Each supervector has MD dimensions where M is the number of mixtures in each acoustic model and D is the dimension of feature vectors. The adapted mean supervector $\mathbf{s}'$ is used to construct the mean vectors of the adapted model.

2.2. Maximum Likelihood Estimation of Weights

The combination weights in (1) can be found by maximizing the likelihood of the test utterance on the adapted acoustic model:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \log p(\mathbf{X}|\mathbf{s}', \lambda, \mathcal{L}_\mathbf{X}) \tag{2}$$

where $\mathbf{w} = [w_1, ..., w_N]^T$ is the weight vector, $\mathbf{X}$ and $\mathcal{L}_\mathbf{X}$ are the observation feature vectors and transcriptions of the current test utterance, respectively, and λ denotes acoustic model parameters other than mean vectors. In unsupervised adaptation, $\mathcal{L}_\mathbf{X}$ can be obtained by a first pass decoding using an initial acoustic model.

Expectation-maximization (EM) algorithm is used to find the solution of $\mathbf{w}$ iteratively. The auxiliary function is defined as

$$Q(\mathbf{w}; \hat{\mathbf{w}}) = -\frac{1}{2}\sum_{m,t}\gamma_m(t)(\mathbf{x}_t - \mu_m)^{\mathrm{T}}\Sigma_m^{-1}(\mathbf{x}_t - \mu_m) \tag{3}$$

where $\hat{\mathbf{w}}$ is the previous weight estimate, $\mathbf{x}_t$ is the feature vector of frame t, and $\gamma_m(t)$ is the posterior of the m^{th} mixture at frame t given $\hat{\mathbf{w}}$, $\mathbf{X}$, and $\mathcal{L}_\mathbf{X}$. T is the number of frames in $\mathbf{X}$, and μ_m and Σ_m are the mean and diagonal covariance matrix of mixture m in the adapted model. Terms not related to $\mathbf{w}$ are not included in (3).

The mean μ_m is represented as $\mu_m = \sum_{i=1}^N w_i \mathbf{s}_{i,m} = \mathbf{S}_m\mathbf{w}$, where $\mathbf{s}_{i,m}$ is the subvector for mixture m in supervector $\mathbf{s}_i$ and $\mathbf{S}_m = [\mathbf{s}_{1,m}, ..., \mathbf{s}_{N,m}]$ is a $D \times N$ matrix. After maximizing the auxiliary function, the solution of $\mathbf{w}$ is

$$\hat{\mathbf{w}} = \left[\sum_{m,t}\gamma_m(t)\mathbf{S}_m^T\Sigma_M^{-1}\mathbf{S}_m\right]^{-1}\sum_{m,t}\gamma_m(t)\mathbf{S}_m^T\Sigma_M^{-1}\mathbf{x}_t \tag{4}$$

Note that the solution in (4) is the same as the solution of eigenvoice weights in [6]. The only difference is that in mean combination the regressors are the mean supervectors, while in eigenvoice the regressors are a set of basis supervectors which are obtained from the mean supervectors through principal component analysis (PCA).

2.3. Lasso Estimation of Weights

Lasso [7] imposes an L_1 regularization term on linear regression problems. Due to the L_1 term, some of the weights will shrink to exactly zero. This property may be desirable in mean combination as it may exclude those mean supervectors not related to the current test environment. In this way, the adapted model may fit better to the test environment as there is less noise in mean combination.

With L_1 regularization, the auxiliary function in (3) becomes

$$\begin{aligned} Q(\mathbf{w}; \hat{\mathbf{w}}) &= -\frac{1}{2}\sum_{m,t}\gamma_m(t)(\mathbf{x}_t - \mathbf{S}_m\mathbf{w})^T\Sigma_m^{-1}(\mathbf{x}_t - \mathbf{S}_m\mathbf{w}) \\ &\quad -T\alpha\sum_{i=1}^N |w_i| \end{aligned} \tag{5}$$

where α is a tuning parameter that controls the weight of the L_1 constraint and $|w_i|$ denotes the absolute value of w_i. As different test utterances have different number of frames T and the first term in (5) is summed over T, we also multiply the Lasso regularization term with T such that the same α produces similar degree of weight shrinkage for utterances of different lengths.

Due to the L_1 constraint, it is difficult to maximize (5) directly w.r.t. all the weights. Instead, we optimize the weights one by one iteratively. Let's define

$$\mathbf{w}_{-i} = [w_1, w_2, ..., w_{i-1}, 0, w_{i+1}, ..., w_N]^T \tag{6}$$

$$\mathbf{z}_t^i = \mathbf{x}_t - \mathbf{S}_m\mathbf{w}_{-i} \tag{7}$$

where $\mathbf{w}_{-i}$ is the weight vector with the i^{th} element reset to 0. Then we maximize (5) w.r.t. to w_i and treat other weights as constants

$$\begin{aligned} \widehat{w_i} &= \arg\max_{w_i}\Bigg\{ -\frac{1}{2}\sum_{m,t}\gamma_m(t)(\mathbf{z}_t^i - w_i\mathbf{s}_{i,m})^T\Sigma_m^{-1} \\ &\qquad (\mathbf{z}_t^i - w_i\mathbf{s}_{i,m}) - T\alpha|w_i|\Bigg\} \end{aligned} \tag{8}$$

Take the derivative w.r.t. w_i and make it equal to 0, we get:

$$\begin{aligned} 0 &= \sum_{m,t}\gamma_m(t)\mathbf{s}_{i,m}^T\Sigma_m^{-1}(\mathbf{z}_t^i - w_i\mathbf{s}_{i,m}) - T\alpha\mathrm{sign}(w_i) \\ &= c - dw_i - T\alpha\mathrm{sign}(w_i) \end{aligned} \tag{9}$$

where $\mathrm{sign}(w_i)$ returns the sign of w_i and

$$\begin{aligned} c &= \sum_{m,t}\gamma_m(t)\mathbf{s}_{i,m}^T\Sigma_m^{-1}\mathbf{z}_t^i \\ &= \sum_m \mathbf{s}_{i,m}^T\Sigma_m^{-1}(\overline{\mathbf{x}_m} - \gamma_m\mathbf{S}_m\mathbf{w}_{-i}) \end{aligned} \tag{10}$$

$$d = \sum_m \gamma_m\mathbf{s}_{i,m}^T\Sigma_m^{-1}\mathbf{s}_{i,m} \tag{11}$$

where $\gamma_m = \sum_t \gamma_m(t)$ and $\overline{\mathbf{x}_m} = \sum_t \gamma_m(t)\mathbf{x}_t$. The final solution for w_i is then [8]

$$\widehat{w_i} = \left(\left|\frac{c}{d}\right| - \frac{T\alpha}{d}\right)_+ \mathrm{sign}\left(\frac{c}{d}\right) \tag{12}$$

where $(x)_+ = \max(x, 0)$.

The weights are initialized as the ML estimate in (4). Then the weights are re-estimated by using the Lasso estimation in (12) one by one. After all the weights are optimized by Lasso, the transcription of the test utterance is updated and the optimization process is repeated. In this study, we only run 2 iterations of Lasso estimation for each utterance as most of the improvement is obtained within 2 iterations. Due to the shrinking nature of Lasso estimation, the weights turn to be scaled toward 0. This problem is solved by simply renormalizing the weights' sum to 1 after Lasso estimation.

3. EXPERIMENTS

3.1. Experimental Settings

The proposed Lasso mean combination is evaluated on the Aurora-2 benchmark task [9]. A baseline system is trained using the standard multi-condition training scheme with standard simple back-end configuration. The training data includes 17 environment conditions, including clean condition and 4 noise types with each noise type further divided into 4 SNR levels, i.e. 20dB, 15dB, 10dB, and 5dB. In addition, each of the 17 conditions is further divided into male and female parts, hence the training data is divided into 34 homogenous sets. The baseline acoustic model is adapted to the 34 homogenous sets by using maximum likelihood linear regression (MLLR) mean transforms [3] to generate environment-dependent models, from which 34 mean supervectors are extracted to form the mean supervector set $\mathcal{S}$. We experiment with two sets of supervectors, one set is obtained using global MLLR transforms and the other is obtained using regression class-based MLLR transforms (4 transforms per condition). In all model combination experiments, the initial transcription of a test utterance is obtained by a first pass decoding using the baseline model.

The features are 39 MFCCs defined by the Aurora-2 task, and c0 is used instead of log energy. All features are normalized by utterance-dependent mean and variance normalization (MVN).

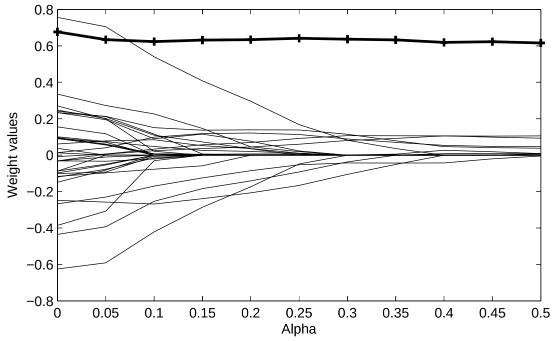
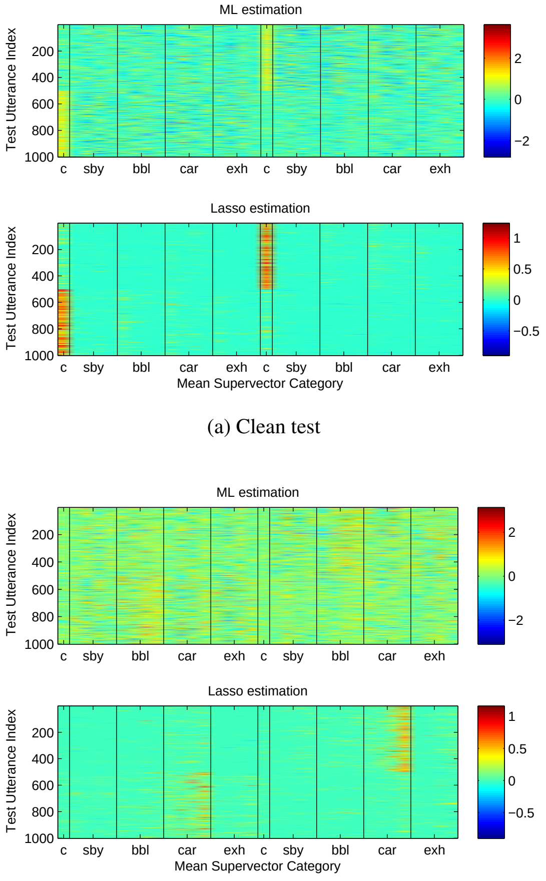


Fig. 1: Illustration of weight shrinkage in Lasso mean combination using utterance "MAH_O789A" (male speaker, subway noise, SNR=15dB). There are 34 lines in the figure, each representing the weight evolution of one mean supervector with increasing α. The bold line shows the weight of the correct mean supervector.

3.2. Investigation of Weights

Let's first investigate the shrinkage of the weights in Lasso estimation. Fig. 1 shows the evolution of weights for one test utterance using the supervectors obtained with global MLLR transforms. Note that the environment condition of the utterance matches exactly one of the training conditions. The tuning parameter α starts from 0 and gradually increases to 0.5, the larger the α, the stronger the L_1 constraint. From the figure, it is observed when $\alpha = 0$ (equivalent to ML estimate), all the weights are non-zero. As α increases, the weights generally shrink towards zero due to the L_1 constraint. However, the weight of the mean supervector that corresponds to the true environment and gender of the test utterance (shown in bold in the figure) does not shrink significantly. After the weights become stable at high α, the correct mean supervector has a large weight, while most other mean supervectors have zero weights. In addition, our investigation shows that those supervectors with non-zero weights are usually related to the true environment. This shows that the L_1 constraint helps to choose the correct mean supervector.

Next, let's investigate the weights for two environment conditions, clean and 5dB car noise, both included in the training conditions. In the clean test condition as shown in Fig. 2(a), the first 500 utterances are spoken by male speakers and the second 501 utterances by female speakers. Each row of the weight matrices corresponds to one test utterance and contains 34 weights of mean supervectors . The weights in the left half of each row are for the supervectors of female speakers and those in the second half are for male speakers. In each gender category, the 17 weights are further divided into 5 categories (marked by the black vertical lines), i.e. clean (c), subway noise (sby), babble noise (bbl), car noise (car), and exhibition noise (exh). Furthermore, in each noisy category, there are four weights corresponding to (from left to right) 20dB, 15dB, 10dB, and 5dB. By examining the weights in Fig. 2(a), it is observed that while the weights obtained by ML are quite random, the weights obtained by Lasso are sparse and consistent with the true environment conditions most of the time. Specifically, the Lasso weights of the first 500 utterances (from male speakers) usually have large positive values for the supervector "c" in the center of the figure (corresponding to male clean supervector). Similar observation is also true for the second half of the utterances. Fig. 2(b) shows the same type of investigation for 5dB car noise test condition. It is observed that the ML weights are again quite random, and the Lasso weights concentrate around the true supervectors. The investigation of clean and 5dB car noise test conditions shows that the Lasso estimation selects better mean supervectors than the ML estimation for target model construction when the test condition is exactly the same as one of the training conditions.



(a) Clean test

(b) Car noise, SNR = 5dB

Fig. 2: Comparison of weights obtained by ML and Lasso.

When the test condition does not match any of the training condition, both ML and Lasso estimation will combine supervectors of multiple conditions to predict the target supervector. However, the weights using Lasso estimation are much more sparse than the ML weights. We will examine whether the sparseness in the weight vector is good for robust speech recognition in the next section.

3.3. Experimental Results and Discussions

We first investigate the effect of α on speech recognition performance. Fig. 3 shows the average word accuracy obtained by Lasso mean combination with different α. Note that $\alpha = 0$ corresponds to the pure ML estimation. The figure shows that the recognition accuracy is quite stable around $\alpha = 0.2$. Hence, we will use $\alpha = 0.2$ for the rest of the paper. At $\alpha = 0.2$, about 65% to 80% of the weights on average are exactly zero.

Table 1 shows the results obtained by ML and Lasso mean combination in test set A. The baseline refers to the results of standard multi-condition training scheme. As the environment conditions in test set A is the same as those in the training data, there is one true

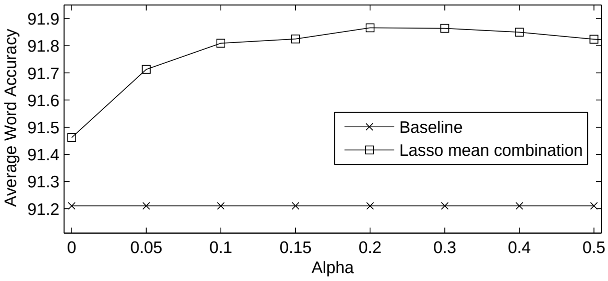


Fig. 3: Recognition word accuracy obtained with Lasso mean combination with different α. The results are averaged over test cases from 0dB to 20dB and over all test sets. The supervector set is obtained with global MLLR transforms.

Table 1: Recognition accuracy averaged over test set A. Oracle uses correct mean supervectors in testing. Oracle, ML, and Lasso use the supervector set generated by global MLLR mean transforms, while Oracle4, ML4, and Lasso4 use the supervector set from 4 class-based transforms. Avg. denotes results averaged from 20dB to 0dB.

SNR	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	Avg.
Baseline	98.3	98.5	97.6	96.1	90.6	73.6	39.4	91.3
Oracle	98.4	98.7	98.0	96.5	91.2	75.5	41.7	92.0
ML	98.4	98.7	97.9	96.5	91.2	74.4	36.4	91.7
Lasso	98.4	98.7	98.0	96.5	91.6	75.9	40.5	92.1
Oracle4	98.8	98.8	98.3	96.9	91.8	77.1	43.5	92.6
ML4	98.5	98.7	98.0	96.6	91.2	75.1	36.4	91.9
Lasso4	98.7	98.8	98.1	96.7	92.0	76.4	41.1	92.4

mean supervector for each test condition. We use the true supervector to generate a model for each test utterance and call the obtained results “Oracle” results. For -5dB and 0dB where there is no true supervector, 5dB supervectors are used. From Table 1, we have two major observations. First, both ML and Lasso produce better results than the baseline. Second, Lasso outperforms ML and achieves similar performance as the “Oracle” results. These results show that the accurate supervector selection by Lasso as shown in Fig. 2 helps to improve the robustness of the adapted model.

Table 2 shows the performance of mean combination on test set B where the test environment conditions are not observed during training. This time there is no “Oracle” result. From the table, it is observed that Lasso still produces better overall performance than ML. This shows that the sparseness in the weight vectors obtained by Lasso is beneficial for robust speech recognition even when the test condition is not seen during training. Finally, the performance of Lasso and ML averaged over all test cases are shown in Table 3. The results show that the Lasso mean combination performs better than the ML mean combination in almost all SNR levels.

Table 2: Recognition accuracy averaged over test set B.

SNR	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	Avg.
Baseline	98.3	98.6	97.8	96.2	90.5	73.7	38.6	91.4
ML	98.4	98.7	98.0	96.5	90.8	73.1	35.6	91.4
Lasso	98.4	98.7	98.0	96.5	91.1	74.3	39.2	91.7
ML4	98.5	98.7	98.0	96.6	91.0	73.6	36.1	91.6
Lasso4	98.7	98.8	98.1	96.7	91.2	74.6	39.5	91.9

Table 3: Recognition accuracy averaged over test sets A, B, and C.

SNR	Clean	20dB	15dB	10dB	5dB	0dB	-5dB	Avg.
Baseline	98.3	98.5	97.6	96.0	90.4	73.5	38.9	91.2
ML	98.5	98.6	97.9	96.3	90.9	73.6	35.9	91.5
Lasso	98.5	98.7	98.0	96.4	91.3	75.0	39.6	91.9
ML4	98.5	98.7	98.0	96.4	91.0	74.2	36.2	91.7
Lasso4	98.7	98.8	98.1	96.6	91.5	75.4	40.2	92.1

4. CONCLUSIONS

In this paper, we adapt the acoustic model towards test environment condition by linearly combining multiple pre-trained environment-dependent models. As the proposed Lasso mean combination with L_1 regularization shrinks the weights of unrelated mean supervectors to exactly zero, a better combined acoustic model could be obtained. Experimental results on Aurora-2 task verified the advantage of the Lasso mean combination over the conventional ML mean combination. For test conditions that are observed in the training data, Lasso mean combination produces similar results as that obtained using true the mean supervectors. For all test conditions, Lasso mean combination generally outperforms the ML mean combination.

5. REFERENCES

[1] L. Deng, J. Droppo, and A. Acero, “Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise,” *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[2] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, “Histogram equalization of speech representation for robust speech recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[3] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[4] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, “A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions,” *Computer Speech and Language*, vol. 23, no. 3, pp. 389–405, Jul. 2009.

[5] R. Lippmann, E. Martin, and D. Paul, “Multi-style training for robust isolated-word speech recognition,” Dallas, TX, Apr. 1987, vol. 12, pp. 705–708.

[6] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, Nov 2000.

[7] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[8] J. Li, M. Yuan, and C.-H. Lee, “Lasso model adaptation for automatic speech recognition,” in *Workshop on Learning Architectures, Representations, and Optimization for Speech and Visual Information Processing*, Bellevue, Washington, USA, Jul. 2011.

[9] D. Pearce and H.-G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recogntion systems under noisy conditions,” in *Proc. ICSLP ’00*, Beijing, China, Oct. 2000, vol. 4, pp. 29–32.