An Ensemble Modeling Approach to Joint Characterization of Speaker and Speaking Environments

Yu Tsao and Chin-Hui Lee

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, GA 30332-0250, USA
{yutsao, chl}@ece.gatech.edu

10.21437/Interspeech.2007-101

Abstract

We propose an ensemble modeling framework to jointly characterize speaker and speaking environments for robust speech recognition. We represent a particular environment by a super-vector formed by concatenating the entire set of mean vectors of the Gaussian mixture components in its corresponding hidden Markov model set. In the training phase we generate an ensemble speaker and speaking environment super-vector by concatenating all the super-vectors trained on data from many real or simulated environments. In the recognition phase the ensemble speaker and speaking environment super-vector is converted to the super-vector for the testing environment with an affine transformation that is estimated online with a maximum likelihood (ML) algorithm. We used a simplified formulation for the proposed approach and evaluated its performance on the Aurora 2 database. In an unsupervised adaptation mode, the proposed approach achieves 7.27% and 13.68% WER reductions, respectively, when tested in clean and averaged noisy conditions (from 0dB to 20dB) over the baseline performance on a gender dependent system. The results suggest that the proposed approach can well characterize environments under the presence of either single or multiple distortion sources.

Index Terms: environment modeling, noise robustness

1. Introduction

In most state-of-the-art automatic speech recognition (ASR) systems, hidden Markov model (HMM) is used as a fundamental tool to characterize speech patterns. Nonetheless one critical limitation of the HMM-base recognizers is that their performance may seriously degrade when there is an acoustic mismatch between training and testing environments. Many techniques have been proposed to reduce the degree of such mismatches. MAP [1] and MLLR [2] generate a new set of HMMs for testing environment by adapting parameters of the HMM set to the new environment. Stochastic matching [3] provides an effective way to estimate the compensation factor in a maximum likelihood self adaptation manner.

Recently, there is a growing interest in using training data from many different environments to reduce environment mismatches. It is well known that multicondition training achieves a better robustness than simple clean condition training [4]. Some techniques use diverse data to obtain a wide variety of environmental models in the training step. These environmental models are often referred to as prior information of the unknown testing environment. A tree-structured piecewise-linear transformation (PLT) [5] approach selects one most suitable HMM set from a pool of noisy speech HMM sets. Then the selected HMM set is further adapted using linear transformations. On the other hand SPLICE [6] estimates a correction vector that is a linear weighted sum of all correction vectors for the mean vectors of the Gaussian components in all of the environmental models to compensate for new testing environments. Another category of approaches represents each environment with a super-vector consisting of the entire set of mean vectors of all the Gaussian components in an HMM set. IEM [7] interpolates the super-vector for an unknown environment in the environment space. Eigenvoice [8] interpolates the super-vector for a testing speaker but using a speaker space generated from a large population of speakers.

In this paper, we proposed an *ensemble speaker and speaking environment modeling* (ESSEM) approach to simultaneously reduce the speaker and speaking distortions induced by mismatch conditions. In the ESSEM approach, every environment of interest is modeled by a super-vector consisting of the entire set of mean vectors from all Gaussian components in a set of HMMs. If we have available a large collection of super-vectors modeling environments for different speakers, adverse conditions, and signal-to-noise (SNR) levels, we can determine the super-vector for the unknown testing environment and use it to construct HMM set for ASR. To estimate the super-vector for the testing environment, the entire set of collected super-vectors is first concatenated to form an ensemble speaker and speaking (ESS) super-vector. Then the super-vector for the testing environment is estimated by converting the ESS super-vector with an affine transformation that is estimated online with adaptation data (adaptation) or testing data (self adaptation or compensation) from that environment.

We tested the ESSEM approach in an unsupervised adaptation mode on the Aurora 2 [9] connected digit recognition task. Because the Aurora 2 corpus provides a multicondition training set with training data from different speakers and speaking environments, it is particularly suited for evaluating the ESSEM approach. We tested performance on both gender independent (GI) and gender dependent (GD) systems. In a GI system, ESSEM achieves a 9.07% average word error rate (WER) among different adverse conditions in the three testing sets from 0dB to 20 dB, corresponding to a 17.62% (11.01% to 9.07%) WER reduction over the baseline performance. Meanwhile in a GD system, ESESM achieves an average 7.95% WER, corresponding to a 13.68% (9.21% to 7.95%) WER reduction over the baseline performance.

2. Ensemble speaker and speaking environment modeling (ESSEM)

2.1. ESSEM framework

There are two processes in the ESSEM approach. In the offline process we collect a wide range of training data from different speaker and speaking environments. Generally the

collection of training data of combinations of different speakers, adverse conditions, and noise levels may be too prohibitive, so the Monte Carlo (MC) [10] techniques can be used to simulate a wide range of speaker and speaking conditions. If there are P sets of training data collected or artificially simulated, we train P HMM sets for the P different speaker and speaking environments. Next the entire set of mean vectors of a HMM model set for the p-th environment is concatenated into a super-vector $\mathbf{X}_p$. If every mean vector is a D-dim vector, then the super-vector for the p-th environment is an R-dim ($R=D\times M$) vector, with M Gaussian mixture components in one HMM set. The space constructed by the P super-vectors, $\mathbf{\Pi}=\{\mathbf{X}_1\ \mathbf{X}_2 \ldots \mathbf{X}_P\}$, is referred to as an ensemble speaker and speaking (ESS) environment space in this paper. Finally we concatenate these P super-vectors to form an ensemble speaker and speaking (ESS) super-vector $\mathbf{Q}=[\mathbf{X}_1^{\mathrm{T}}\ \mathbf{X}_2^{\mathrm{T}} \ldots \mathbf{X}_P^{\mathrm{T}}]^{\mathrm{T}}$ of dimension ($R\times P$).

In the online process, we intend to estimate an R-dim super-vector $\mathbf{X}_{test}$ for an unknown testing environment by converting the ESS super-vector $\mathbf{Q}$ with a transformation matrix $\hat{\mathbf{A}}$ of dimension $R*(R\times P)$ and a compensation vector $\hat{\mathbf{b}}$ of dimension R:

$$\mathbf{X}_{test} = \hat{\mathbf{A}}\,\mathbf{Q} + \hat{\mathbf{b}}\ . \tag{1}$$

$\{\hat{\mathbf{A}},\hat{\mathbf{b}}\}$ is estimated based on a maximum likelihood (ML) algorithm with given a segment of feature vectors $\mathbf{O}_{test}$ corresponding to adaptation data for the testing environment:

$$\{\hat{\mathbf{A}},\hat{\mathbf{b}}\} = arg\ \max_{\{\tilde{\mathbf{A}},\tilde{\mathbf{b}}\}} L(\mathbf{O}_{test} \mid \tilde{\mathbf{A}}\,\mathbf{Q} + \tilde{\mathbf{b}}), \tag{2}$$

where L is the likelihood function and $\{\hat{\mathbf{A}},\hat{\mathbf{b}}\}$ is referred to the ensemble speaker and speaking (ESS) affine transformation.

If we decompose the matrix $\hat{\mathbf{A}}$ to P distinct $R*R$ matrices $\mathbf{A}_1, \mathbf{A}_2 \ldots, \mathbf{A}_P$, and partition the ESS super-vector $\mathbf{Q}$ into P vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_P$, Eq. (1) can be re-written as:

$$\mathbf{X}_{test} = \sum_{p=1}^{P} (\mathbf{A}_p\,\mathbf{X}_p + \mathbf{b}_p), \text{ with } \sum_{p=1}^{P} \mathbf{b}_p = \hat{\mathbf{b}}. \tag{3}$$

Therefore $\mathbf{X}_{test}$ can also be seen as a linear combination of transformed super-vectors for the P distinct environments.

2.2. Joint/individual characterization

Next we study the ESSEM approach for joint and individual characterization of environments with different distortion sources. In the offline process we classify the P training environments into several categories according to the distortion types, e.g., speaker variations, background noises, and channel distortions. The ESS super-vector can be re-arranged by $\mathbf{Q}=[\mathbf{Q}_s^{\mathrm{T}}\ \mathbf{Q}_n^{\mathrm{T}}\ \mathbf{Q}_h^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{Q}_s$, $\mathbf{Q}_n$, and $\mathbf{Q}_h$ are speaker , noise, and channel environment super-vectors, respectively, and Eq.(1) becomes:

$$\mathbf{X}_{test} = (\hat{\mathbf{A}}_s\,\mathbf{Q}_s + \hat{\mathbf{b}}_s) + (\hat{\mathbf{A}}_n\,\mathbf{Q}_n + \hat{\mathbf{b}}_n) + (\hat{\mathbf{A}}_h\,\mathbf{Q}_h + \hat{\mathbf{b}}_h). \tag{4}$$

$\{\hat{\mathbf{A}}_s,\hat{\mathbf{b}}_s\}$, $\{\hat{\mathbf{A}}_n,\hat{\mathbf{b}}_n\}$, and $\{\hat{\mathbf{A}}_h,\hat{\mathbf{b}}_h\}$ are for the speaker, noise, and channel environment affine transformations, respectively. From Eq.(4) we suggest that the ESS environment space can be generated according to the testing situations to enhance the ESSEM performance. Using a particular noise robustness case as an example, if we have a prior information about the identity of the testing speaker and the type of the communication channel, we simply collect or simulate data to cover a wide range of different noise types and SNR levels with fixed speaker and channel characteristics when building the ESS environment space. Then Eq.(4) becomes:

$$\mathbf{X}_{test} = \hat{\mathbf{A}}_n\,\mathbf{Q}_n + \hat{\mathbf{b}}_n\ . \tag{5}$$

From Eq.(5) ESSEM individually characterizes a testing environment under the presence of unknown noise distortions. The claim is actually supported by the successfulness of the IEM [7] approach for robust ASR. In a similar manner, ESSEM estimates $\{\hat{\mathbf{A}}_s,\hat{\mathbf{b}}_s\}$ for speaker adaptation and $\{\hat{\mathbf{A}}_h,\hat{\mathbf{b}}_h\}$ for channel factor compensation tasks.

3. Key issues in the ESSEM framework

3.1. Dimension reduction

In this section we describe some dimension reduction techniques to properly reduce the number of free parameters to be estimated when the amount of adaptation data is very limited or none (self adaptation or compensation). We provide two methods here in this paper. The first method is applying the PCA technique on the entire ESS environment space while keeping the K ($K\le P$) eigenvectors with the highest singular values to form a PCA-imposed environment space. Then the super-vector $\mathbf{X}_{test}$ is estimated by:

$$\mathbf{X}_{test} = \hat{\mathbf{A}}^{(PCA)}\,\mathbf{E}_{\mathbf{X}} + \hat{\mathbf{b}}^{(PCA)}, \tag{6}$$

where $\mathbf{E}_{\mathbf{X}}$ is the PCA-imposed ESS super-vector of dimension ($R\times K$), $\hat{\mathbf{A}}^{(PCA)}$ and $\hat{\mathbf{b}}^{(PCA)}$ are PCA-imposed transformation matrix of dimension $R*(R\times K)$ and PCA-imposed compensation vector of dimension R. Similar to the derivation from Eq.(1) to Eq.(3) , we represent Eq.(6) as:

$$\mathbf{X}_{test} = \sum_{k=1}^{K} (\mathbf{A}_k^{(PCA)}\,\mathbf{e}_{\mathbf{X}k} + \mathbf{b}_k^{(PCA)}), \text{ with } \sum_{k=1}^{K} \mathbf{b}_k^{(PCA)} = \hat{\mathbf{b}}^{(PCA)}, \tag{7}$$

where $\mathbf{e}_{\mathbf{X}k}$ is for the k-th principle eigenvector in the PCA-impose ESS environment space, and $\{\mathbf{A}_k,\mathbf{b}_k\}$ is for the k-th distinct PCA-imposed ESS affine transformation.

The second method is a cluster selecting (CS) technique. In the offline phase, we classify the ESS environments into several clusters based on the acoustic similarity between each pair of super-vectors in the ESS environment space. The entire set of super-vectors within the same cluster is then concatenated to form a cluster-based ESS super-vector, and we have $\{\mathbf{Q}_1, \mathbf{Q}_2 \ldots, \mathbf{Q}_C\}$, for C different clusters. In the online process, an additional step is performed to locate a cluster of environments that best matches to the unknown testing environment. Then the $\mathbf{X}_{test}$ can be estimated by:

$$\mathbf{X}_{test} = \hat{\mathbf{A}}_c\,\mathbf{Q}_c + \hat{\mathbf{b}}_c\ , \tag{8}$$

where $\mathbf{Q}_c$ is the CS environment super-vector, and $\{\hat{\mathbf{A}}_c,\hat{\mathbf{b}}_c\}$ is the affine transformation for the selected c-th cluster. It is noted that the number of eigenvectors used in the PCA method and number of environments within one cluster in the CS method directly affects the transformation complexity. These numbers should be optimally determined with respect to the quantity of available adaptation statistics.

3.2. Transformation formation

Finally we investigate using simpler transformation formations to further reduce computation complexity in this section. Four methods are discussed here. First the correlation across different mean vectors in an ESS super-vector can be ignored by setting each distinct matrix $\mathbf{A}_p$ in Eq.(3) to a block-diagonal formation. Then the m-th mean vector $\mu_{m,test}$ in the super-vector $\mathbf{X}_{test}$ is obtained by:

$$\mu_{m,test} = \sum_{p=1}^{P} (\mathbf{A}_{m,p}\, \mu_{m,p} + \mathbf{b}_{m,p}) \quad (9)$$

where $\mathbf{A}_{m,p}$ and $\mathbf{b}_{m,p}$ are matrix of dimension ($D \times D$) and compensation vector of dimension D to the m-th Gaussian mixture component for the p-th environment. It is noted here that if only one environment is used in the ESS super-vector, ESSEM turns into the conventional MLLR approach. Second we can use a global affine transformation or class-based affine transformations in Eq.(9) for many different mean vectors in the ESS super-vector. Third, a simpler matrix formulation can be used for the affine transformation. For example, a diagonal matrix can be used for each $\mathbf{A}_{m,p}$. When we use a simplified matrix $\mathbf{A}_p = \omega_p \times \mathbf{I}$ (ω_p is a weighting coefficient and $\mathbf{I}$ is an identity matrix) and a global bias vector $\mathbf{b}$ in Eq.(9), we have:

$$\mu_{m,test} = \sum_{p=1}^{P} \omega_p\, \mu_{m,p} + \mathbf{b}. \quad (10)$$

The formulation in Eq.(10) is equivalent to that used in the cluster weighting + bias (CWB) technique for speaker adaptation [11]. To further reduce complexity, we employ the PCA-imposed ESS super-vectors in Eq.(10):

$$\mu_{m,test} = \sum_{k=1}^{K} \omega_k\, e_{\mu_m,k} + \mathbf{b}, \quad (11)$$

where $e_{\mu_m,k}$ is for the mean vector for the m-th Gaussian mixture component in $\mathbf{e}_{\mathbf{X}_k}$, and ω_k is for the k-th weighting coefficient. If $\mathbf{b}$ in Eq.(11) is set to a constant for further simplification, we find that the simplified version of ESSEM resembles to the IEM and eigenvoice approaches. Finally, a TEM-typed [7] method may provide another way to efficiently estimate the ESS affine transformations. We first obtain another large collection of affine transformations in the offline processes with each transformation characterizing the correlation between a particular noisy super-vector and the ESS super-vector. In the online process, only a set of weighting coefficients is estimated to determine the new ESS affine transformation for the unknown testing environment.

4. Experimental setup and result analysis

We evaluated the ESSEM approach on the Aurora 2 database. The multicondition training set in Aurora 2 is used to train HMM sets and to build environment spaces. This training set involves the same four types of noise as in test set A, at four SNR levels: 20 dB, 15 dB, 10 dB, and 5 dB, along with clean data. Thus we have data for 17 (4×4+1) different speaking environments. To model speaker variation, we can use as many as possible training speakers in the training set to model diverse speaker environments. However the total number of different speaker and speaking environments may become too huge for ESSEM. *K*-means or tree-structure methods can be used to classify speakers into several groups, and each group stands for a particular speaker environment. Here we simply use two genders for two different speaker environments. The training set is divided into two gender-specific multicondition training sets, and we have data for 34 (17×2) different speaker and speaking environments.

For the front-end processing, the speech signals are characterized by 39 coefficients that consisted of 13 MFCC (C0 to C12) parameters plus their first and second order time derivatives. An utterance-level cepstral mean subtraction (CMS) was performed for normalization. The HMMs were trained by following the Aurora 2 standard specifications. A left-to-right topology is used to model whole digits, and each digit model has 16 active states, with each state characterized by 3 Gaussians mixture components. There are 3 states for the silence model and 1 state for the short pause model, with each state characterized by 6 Gaussian mixture components.

We evaluate ESSEM on both GI and GD systems. For the GI system a GI HMM set is trained on the multicondition training data, and 34 environmental HMM sets are obtained by adapting mean vectors from the GI HMM set to particular environments. Next we collected the entire set of mean vectors for each of these 34 HMM sets to generate an ESS super-vector. For the GD system two GD HMM sets were first trained. Then 17 environmental HMM sets for one GD HMM set are obtained by adapting mean vectors from that GD HMM set to particular environments. In another word, two sets of ESS super-vectors corresponding to the two GD HMM sets are prepared. An additional HMM set was prepared for automatic gender identification (AGI). In this HMM set, each gender is modeled with 16 active states with each state characterized by 88 Gaussian mixture components.

The complete test sets in Aurora 2 are used for testing. There are totally 70 different testing environments with 1001 testing utterances in each environment. We test ESSEM in a per-utterance unsupervised adaptation mode. Each testing utterance is first decoded into *N*-best list, and the *N*-best information is then used as adaptation statistics for ESSEM. In the preliminary experiments, we have tried different dimension reduction methods with various transformation formulations, and we found that the IEM approach with a simplified transformation formation in Eq. (11) achieves the best performance for this particular task. In the following experiments, we use IEM as a representative of the ESSEM-type methods. We evaluate the performance of recognition results in three conditions: 1) "Clean"- average WERs over the three test sets at clean condition; 2) "0dB-20dB"- average WERs among different noise types over the three test sets from 0dB to 20dB; 3) "-5dB"- average WERs among different noise types over the three test sets at -5dB.

4.1. Gender independent system

Table 1 listed results of ESSEM on the GI system. "GI-Baseline" indicates simply using a multicondition trained GI HMM set for ASR. Two types of ESSEM with different ESS super-vectors are implemented and evaluated. "GI-Full" is for ESSEM with the entire ESS super-vectors (*P*=34); "GI-PCA" is for ESSEM with PCA-imposed ESS super-vectors (*K*=17). The two types of ESSEM achieve very similar performance across "Clean", "0dB-20dB", and "-5dB" conditions, so only the results of "GI-PCA" are listed in Table 1. From Table 1 we observed that in "0dB-20dB" condition, "GI-PCA" achieves a 9.07% average WER, corresponding to a 17.62% (11.01% to 9.07%) WER reduction over "GI-Baseline". The improvement confirms that ESSEM well characterizes unknown testing environments and accordingly enhances robustness performance. Moreover, it is observed that in "Clean" condition, "GI-PCA" still achieves better results than "GI-Baseline". We assume the WER reductions are provided by ESSEM in reducing speaker variation distortion.

Table 1. *Word error rates (in %) in different test conditions for the ESSEM approach on a gender independent system.*

Test conditions	Clean	0dB-20dB	-5dB
GI-Baseline	1.68	11.01	73.08
GI-PCA	1.32	9.07	69.83

4.2. Gender dependent system

In this section, we use the results of ESSEM on the GD system to verify the claim in section 3.1 that the CS method can optimally reduce dimensionality for a particular task with very limited or even no adaptation data. For this set of experiments, every incoming testing utterance is used to: 1) do automatic gender identification; 2) select a more suitable GD HMM set and its corresponding ESS super-vector; 3) perform ESSEM in an unsupervised adaptation mode. The automatic gender identification is done by using the addition HMM set mentioned earlier. We listed results for this set of experiments in Table 2. Results for "GD-Baseline" are obtained by performing an automatic gender identification to select a more suitable GD HMM set followed by using the selected GD HMM set to do ASR. Similar to GI system, we evaluate ESSEM with different ESS super-vectors. "GD-Full" is ESSEM with all the ESS super-vectors (P=17), and "GD-PCA" is ESSEM with PCA-imposed ESS super-vectors (we set K=10 here). Again very similar trends as those from the GI system are observed from the GD system that "GD-Full" and "GD-PCA" achieve very similar performance, and here we only present the results of "GD-Full" in Table 2. From Table 2, it is observed that "GD-Full" achieves better performance over "GD-Baseline" across "-5dB", "0dB-20dB", and "Clean" conditions. Specially in "0dB-20dB" condition, "GD-Full" achieves a 7.95% WER, corresponding to a 13.68% (9.21 to 7.95%) WER reduction over "GD-Baseline".

Table 2. *Word error rates (in %) in different test conditions for the ESSEM approach on a gender dependent system.*

Test conditions	Clean	0dB-20dB	-5dB
GD-Baseline	1.10	9.21	72.59
GD-Full	1.02	7.95	66.97

When evaluating performance on the Aurora 2 database, it is usually more interested in the results for "0dB-20dB" condition. Therefore we plotted the results for "GI-Baseline", "GI-Full", "GI-PCA", "GD-Baseline", "GD-Full", and "GD-PCA" in "0dB-20dB" condition into Figure 1 for a clear comparison. Again we observe consistent improvements for ESSEM over the baseline performance with either full set or PCA-imposed ESS super-vectors on both GI and GD systems. Moreover although the "GD-Baseline" already achieves very good performance by reducing gender mismatch effects from the GI system, ESSEM can further improve the performance on GD system and provide clear WER reductions. Finally by comparing results for "GI-PCA" and "GD-Full", (both using ESS super-vectors of dimension 17) we verify that a CS method may be a better choice to the PCA technique for dimension reduction.

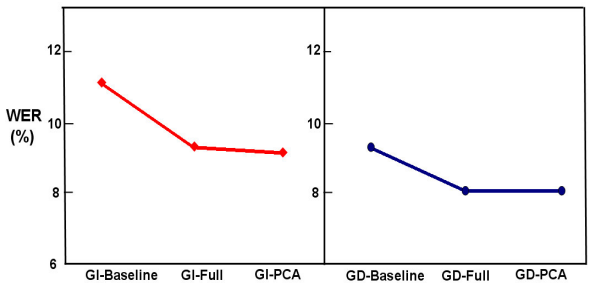

Figure 1: *Comparison of ESSEM on gender independent (left panel) and gender dependent (right panel) systems.*

5. Conclusions

In this paper, we propose an ensemble modeling approach to characterizing unknown speaker and speaking environments. The framework is evaluated on the Aurora 2 database. Significant improvements over the best multicondition training performance are observed for the proposed approach in an unsupervised adaptation mode. For a gender independent system, the ESSEM approach achieves a 9.07% WER using a PCA-imposed ESS super-vector, corresponding to a 17.62% average WER reduction over the baseline performance in average different noise types from SNR=0dB to 20dB for the three test sets in the Aurora 2 database. In a gender dependent system, the ESSEM approach achieves a 7.95% WER using the full set of ESS super-vector, corresponding to a 13.68% WER reduction. The results suggest that ESSEM is an effective way to perform ensemble speaker and speaking environment modeling even with very limited or no adaptation data.

6. Acknowledgements

This work was supported by a TI Leadership University grant.

7. References

[1] Gauvain, J.-L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. on Speech and Audio Proc., Vol. 2, no. 2 , pp.291-99, April 1994.

[2] Leggetter, C. and Woodland, P., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Compute. Speech Language, vol. 9, pp.171-185, 1995.

[3] Sankar, A. and Lee, C.-H., "A maximum-likelihood approach to stochastic matching for robust speech recognition", IEEE Trans. on Speech and Audio Proc., Vol. 4, pp.190-202, May.1996.

[4] Lippmann, R. P., Martin, E. A., and Paul, D. B., "Multi-style training for robust isolated-word speech recognition", in Proc. ICASSP, Dallas, TX, Apr. 1987.

[5] Zhang, Z. and Furui, S., "Piecewise-linear transformation-based HMM adaptation for noisy speech", Speech Commun., vol.42, pp.43–58, 2004.

[6] Deng, L., Droppo, J., and Acero, A., "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition", IEEE Trans. on Speech and Audio Proc., Vol. 11, pp.568-580, Nov.2003.

[7] Tsao, Y. and Lee, C.-H., "A vector space approach to environment modeling for robust speech recognition", Proc. ICSLP, pp.785-788, Sept. 2006.

[8] Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N., "Rapid speaker adaptation in Eigenvoice space", IEEE Trans. on Speech and Audio Proc., Vol. 8, pp.695-707, Nov. 2000.

[9] Hirsh, H. G. and Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. ISCA ITRW ASR 2000, Paris, 2000.

[10] Metropolis, N. and Ulam, S., "The Monte Carlo method", JASA, Vol. 44, pp.335-341, Sept. 1949.

[11] Erdoan, H., Gao, Y., and Picheny, M., "Rapid adaptation using penalized-likelihood methods", in Proc. ICASSP, Salt Lake City, USA, May 2001.