TWO EXTENSIONS TO ENSEMBLE SPEAKER AND SPEAKING ENVIRONMENT MODELING FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Yu Tsao and Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
{yutsao, chl}@ece.gatech.edu

ABSTRACT

Recently an ensemble speaker and speaking environment modeling (ESSEM) approach to characterizing unknown testing environments was studied for robust speech recognition. Each environment is modeled by a super-vector consisting of the entire set of mean vectors from all Gaussian densities of a set of HMMs for a particular environment. The super-vector for a new testing environment is then obtained by an affine transformation on the ensemble super-vectors. In this paper, we propose a minimum classification error training procedure to obtain discriminative ensemble elements, and a super-vector clustering technique to achieve refined ensemble structures. We test these two extensions to ESSEM on Aurora2. In a per-utterance unsupervised adaptation mode we achieved an average WER of 4.99% from 0dB to 20dB conditions with these two extensions when compared with a 5.51% WER obtained with the ML-trained gender-dependent baseline. To our knowledge this represents the best result reported in the literature on the Aurora2 connected digit recognition task.

Index Terms—*environment modeling, noise robustness*

1. INTRODUCTION

For an automatic speech recognition (ASR) system, maintaining a robust performance over a wide range of unknown environments is a key design issue. Many techniques have been proposed to reduce mismatches between training and testing conditions and enhance ASR performance. The first category of approaches generates a new set of hidden Markov models (HMMs) for the testing environment by adapting the parameters of the original HMMs to the new environment. MAP [1] and MLLR [2] are two most prevailing methods used in most state-of-the-art ASR systems. The second group targets at reducing the differences between training and testing speech features according to signal conditioning or blind compensation. Spectral subtraction [3] and the ETSI advanced front-end [4] achieve very good robustness under noisy conditions. Finally some approaches, such as stochastic matching [5], jointly adapt model parameters and compensate for speech feature differences. Although these methods provide good performance improvement, they are not designed to handle multiple distortions, such as combined speaker variations, convolutive channels and additive noises.

In our previous study, an ensemble speaker and speaking environment modeling (ESSEM) [6] approach was proposed to characterizing unknown environments under the presence of either a single or multiple distortions. Here each environment is modeled by a super-vector consisting of the entire set of mean vectors from all Gaussian components of a set of HMMs for the particular environment. In the offline phase a large collection of ensemble super-vectors are built by simulating different combinations of multiple distortions. On the other hand in the online phase the super-vector for the unknown testing environment is obtained by converting the ensemble super-vectors with an affine transformation that is estimated with adaptation data from that environment. In [6] we proposed to use a cluster selection technique to optimally reduce the dimension of the environment super-vector. Based on the acoustic knowledge, we tested the performance of the cluster selecting method by implementing ESSEM on a gender-dependent system where the full set of environments were clustered into two groups, one for each gender.

In this paper we propose two ESSEM extensions, namely a general environment clustering procedure on the environments and a minimum classification error (MCE) training method [7] to obtain parameters of the environment space for discriminative modeling. We tested the extended ESSEM on the full evaluation set of the Aurora2 [8] task on both gender-independent (GI) and gender-dependent (GD) systems. For the averaged performance from 0dB to 20dB the best ESSEM result we achieved in the GI system was 5.39% WER, yielding a 16.56% WER reduction over the ML baseline result of 6.46% WER. The best result for the GD system was 4.99% WER which represents a 9.44% WER reduction over the ML-based GD baseline result of 5.51% WER.

2. ENSEMBLE SPEAKER AND SPEAKING ENVIRONMENT MODELING (ESSEM)

We first review the two stages in the ESSEM approach. In the offline phase we collect a wide range of speech data from different speaker and speaking environments, e.g., different speakers, noise types, SNR levels, and channel distortions. It is usually prohibitive to collect data from many different real world environments, so the Monte Carlo (MC) [9] technique can be used to artificially simulate these training sets. If there are P sets of training data collected, we can train P sets of HMMs. For each environment, the entire set of mean vectors of a set of HMMs is then concatenated into a super-vector $\mathbf{X}_p$,p=1,…,P. If there are M Gaussian mixture components in one set of HMMs, and every mean vector has D dimensions, the super-vector for the p-th environment is an R-dim (R=D×M) vector. These P super-vectors form an ensemble speaker and speaking (ESS) environment space $\Omega_E = Span\{\mathbf{X}_1, \ldots, \mathbf{X}_P\}$. Finally we concatenate them to form an ESS super-vector $\mathbf{Q} = [\mathbf{X}_1^T \ \mathbf{X}_2^T \ldots \mathbf{X}_P^T]^T$ of dimension (R×P). This ESS super-vector assumes *a priori* knowledge for the unknown testing environments.

In the online phase, 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:

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

Many optimal criteria can be used to estimate $\{\hat{\mathbf{A}},\hat{\mathbf{b}}\}$, while a maximum likelihood (ML) algorithm is the most popular one. For the ML algorithm, with a given segment of speech data from the testing environment $\mathbf{O}_{test}$, we have:

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

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

We can 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 sub-vectors, $\mathbf{X}_1, \mathbf{X}_{2\ldots}, \mathbf{X}_P$, such that 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.

The correlation across different mean vectors in each sub-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 $\mathbf{X}_{test}$ can be obtained by:

$$\mu_{m,test} = \sum_{p=1}^{P} (\mathbf{A}_{m,p}\,\mu_{m,p} + \mathbf{b}_{m,p}) \tag{4}$$

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. An extreme case is that only one environment is selected in the ESS super-vector, and ESSEM is equivalent to the conventional MLLR approach.

On the other hand a simpler matrix formulation can be used for the affine transformation in Eq. (3). For example, a matrix $\mathbf{A}_p = \omega_p \times \mathbf{I}$ (with ω_p a weighting coefficient and $\mathbf{I}$ is an identity matrix) neglecting the global bias vector $\mathbf{b}_{m,p}$ in Eq. (4), we have:

$$\mu_{m,test} = \sum_{p=1}^{P} \omega_p\,\mu_{m,p}\,. \tag{5}$$

The formulation in Eq. (5) is equivalent to that used in the interpolation-based environment modeling (IEM) [10] approach, and the weighting coefficients $\omega_p, p = 1,..., P$, in Eq. (5) are estimated based on the ML algorithm presented in Eq. (2).

3. TWO ESSEM EXTENSIONS

3.1. Tree Structure Environment Clustering

In order to enrich the variety of the ESS environment space, i.e., to have more complete *a priori* knowledge, we want to collect or artificially simulate speech data from many different conditions. However the dimension of ESS super-vector must be carefully limited to avoid a possible over-fitting when the amount of adaptation data is very limited (self adaptation or compensation). Moreover we want to fully employ the *a priori* knowledge in modeling unknown testing environments. In our previous study [6] we applied principle component analysis (PCA) to the ESS super-vector. We verified that PCA-imposed ESSEM achieves better accuracy than that with the full set of ESS super-vector. Here we propose tree structure clustering to reduce the super-vector dimension and provide better *a priori* knowledge for ESSEM.

The root of the tree is the entire set of training environments, and the tree is constructed by several layers, with each layer of environment clustering performed based on dissimilarity between each pair of environments. In the offline phase with the constructed tree structure of environment clustering the super-vectors belonging to the same cluster are concatenated to form a cluster-selected (CS) ESS super-vector $\mathbf{Q}_c$, c=1,…,C for C different clusters. If there are S super-vectors in the c-th cluster, the CS-based ESS super-vector is $\mathbf{Q}_c = [\mathbf{X}_{1,(c)}^{\mathrm{T}}\ \mathbf{X}_{2,(c)}^{\mathrm{T}} \ldots \mathbf{X}_{S,(c)}^{\mathrm{T}}]^{\mathrm{T}}$. A particular function R (.) is used to find the most representative super-vector $\mathbf{X}_{\mathrm{rep}}^{(c)}$ for the c-th cluster with $\mathbf{X}_{\mathrm{rep}}^{(c)} = R(\mathbf{Q}_c)$.

In the online phase, an additional best first process is performed to locate a cluster of environments $\mathbf{Q}_T$ where its representative super-vector $\mathbf{X}_{\mathrm{rep}}^{(T)}$ yields the highest likelihood with the given speech data from the unknown testing environment:

$$\mathbf{Q}_T = arg \max_{c} L(\mathbf{O}_{test} \mid R(\mathbf{Q}_c)) \quad c = 1,2...,C\,. \tag{6}$$

From the general formulation of ESSEM in Eq. (1), the super-vector for the testing environment $\mathbf{X}_{test}$ can be estimated by:

$$\mathbf{X}_{test} = \hat{\mathbf{A}}_T\,\mathbf{Q}_T + \hat{\mathbf{b}}_T\,, \tag{7}$$

where $\mathbf{Q}_T$ is the CS-based ESS super-vector, and $\{\hat{\mathbf{A}}_T,\hat{\mathbf{b}}_T\}$ is the affine transformation for the selected T-th cluster.

3.2. MCE Retraining of ESS Super-vector Parameters

We use the MCE training [7] to increase the average distance between pairs of components within the ESS super-vector. Two feasible procedures are presented here. First, if we consider each environment as a particular class, with training data $\mathbf{O}_{train} = \{\mathbf{O}_1,...,\mathbf{O}_P\}$ of totally I utterances for P different environments in the training set, we use the objective function:

$$l(\mathbf{Q}) = \frac{1}{I}\sum_{i=1}^{I} \frac{1}{1 + exp(-\gamma\, d(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}) + \theta)} \tag{8}$$

where $\mathbf{\Lambda}$ is for the parameter set other than means of HMMs, both γ and θ are control parameters for the sigmoid function, and the misclassification measure d (.) is defined as:

$$d(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}) = -\tilde{g}(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}, W_c) + \tilde{\mathrm{G}}(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}) \tag{9}$$

with

$$\tilde{\mathrm{G}}(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}) = \frac{1}{\eta} log\{\frac{1}{N}\sum_{n=1}^{N} exp[\,\eta \times \tilde{g}(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}, W_n)\,]\,\} \tag{10}$$

where η is a positive control parameter, and W_c and $\{W_1,...,W_N\}$ are the given correct transcription and the decoded N-best competing word sequence of the training utterance $\mathbf{O}_{train}^{i}$, respectively. We used a logarithm of the likelihood for the discriminant function of $\tilde{g}(\mathbf{O}_{train}^{i}, \mathbf{Q}, \mathbf{\Lambda}, W_c)$ in the implementation. The generalized probabilistic descent (GDP) algorithm [7] is used to update parameters in the ESS super-vector $\mathbf{Q}$ iteratively.

Second, we increase distance between components within each particular environment in the ESS super-vector. With the training data $\mathbf{O}_p$ of I_p utterances from the p-th environment, we have the following objective function:

$$l(\mathbf{X}_p) = \frac{1}{I_p}\sum_{i=1}^{I_p}\frac{1}{1+exp(-\gamma\, d(\mathbf{O}_p^i, \mathbf{X}_p, \Lambda)+\theta)}. \quad (11)$$

Again the misclassification measures in Eqs. (9) and (10) are used, and parameters within $\mathbf{X}_p$ are updated iteratively.

In our implementation, all parameters in the ESS super-vector are originally estimated with the ML criterion, and followed by the MCE training. Finally we have a MCE-refined ESS super-vector.

4. EXPERIMENTS AND RESULTS

We evaluated the ESSEM approach on the Aurora2 database. The multicondition training set is used both to train HMMs and to build the environment spaces. In this training set there are the same four types of noise as in test set A, at four SNR levels: from 5dB to 20dB, along with clean data. Therefore there are 17 different speaking environments. The training set is further divided into two gender-specific subsets, and now we have 34 (17×2) speaker and speaking environments. The complete test sets in Aurora2 are used for testing. There are totally 70 different testing environments with 1001 testing utterances in each environment. We test ESSEM in an unsupervised adaptation mode. Each testing utterance is first decoded into an *N*-best list, and then used as adaptation statistics for ESSEM. No end-pointing process was applied.

Here ESSEM with a simplified transformation presented in Eq. (5) is implemented and tested on both the GI and GD systems. In the GI system, a set of GI HMMs is trained on the multicondition training data, and 34 sets of environmental HMMs are trained corresponding to the 34 particular environments in the training set. We use a two-layered binary tree structure to cluster the 34 speaker and speaking environments into four groups. With this data-driven clustering scheme it is observed that in the first layer the 34 environments were exactly divided into two groups of two genders. This phenomenon exactly matches our intuition that genders assume the most discriminative power even under very noisy conditions. Then in the second layer another two groups of environments were defined roughly according to high/low SNR levels. In other words for the Aurora2 task the first layer of the tree structure clustering corresponds to speaker clustering, and the second layer is speaking environment clustering.

In the GD system, two sets of GD HMMs and 17 sets of environmental HMMs are trained corresponding to 17 different speaking environments for each gender. The data-driven clustering is used to further cluster speaking environments into two groups, again high/low SNR levels, for each gender. There is an additional set of HMMs for automatic gender identification (AGI) that determines a speaker's gender for every incoming testing utterance. It is noted that in the online phase two stages of cluster selection for the two-layered binary tree are sequentially performed based on Eq. (6) in the GI system. Contrarily there is only one layer of cluster selecting performed in the GD system because the gender identity is already determined by AGI. To maintain an adequate number of environments in every group, some environments, such as environments at middle SNR levels, are shared across different groups. Finally each cluster has 12 to 14 different environments.

4.1. Speaker and Speaking Environment Clustering for ESSEM

For the experiments in this section, each frame is characterized by 39 coefficients consisted of 13 MFCC parameters with their first and second order time derivatives. An utterance-level cepstral mean subtraction (CMS) was performed for normalization. All digits were modeled by 16-state whole word HMMs with each state characterized by 3 Gaussian components. There are 3 states for the background model and 1 state for the short pause model, with each state of background and pause characterized by 6 Gaussian mixture components.

4.1.1. Gender Independent System

We compared performance of ESSEM with different ESS super-vectors. Average word error rates from 0dB to 20dB across the three testing sets for four types of ESS super-vectors along with the baseline results are illustrated in Fig. 1. "Full" indicates the full ESS set was used; "PCA" is for using a PCA-imposed ESS super-vector; "CS (1)" and "CS(2)" denote CS-based ESS super-vector with a one layer (cluster number *C*=2) and two layers (cluster number *C*=4) tree structure environment clustering.

From Figure 1, it is clear that both "CS(1)" and "CS(2)" provide better performance over "Full" and "PCA". Moreover it is observed that the two-layered tree structure gives better performance than the one-layered tree.

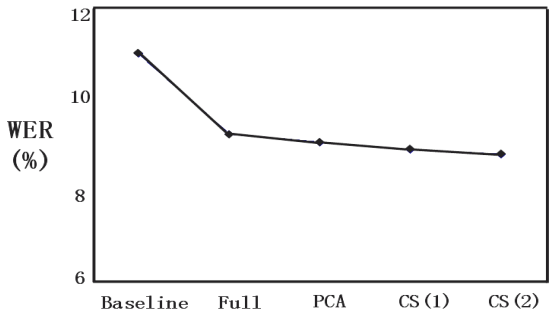

Fig. 1. Comparison of ESSEM with different ESS super-vectors

4.1.2. Gender Dependent System

We list results of ESSEM with the full set of ESS super-vectors and with a CS-based ESS super-vector in the GD system in Table 1. Because there is an AGI process beforehand, and the environments are already clustered into two groups for two genders, the original ESSEM does not need to do online cluster selection as presented in Eq. (6). These original results are listed [6] in the second row in Table 1 and denoted as "GD-ESSEM+Full". Next the results of using one-layered environment clustering are listed in the bottom row and denoted as "GD-ESSEM+CS(1)". We use the AGI process followed by one-layered environment clustering as presented in Eq. (6) to locate one set of HMMs to test recognition for the baseline results. This set of configuration has better recognition accuracy than that of the AGI-only process presented in [6]. We listed and denoted them as "GD-Baseline" in Table 1.

Similar results to the GI system were observed by comparing results of "GD-ESSEM+Full" and "GD-ESSEM+CS(1)" that in addition to two genders, a speaking environment clustering further enhances the performance of ESSEM.

Test conditions	Clean	0dB-20dB	-5dB
GD-Baseline	1.15	8.63	69.59
GD-ESSEM+Full	1.08	7.95	66.97
GD-ESSEM+CS(1)	1.07	7.89	66.84

Table 1. WER (in %) for ESSEM with 2 and 4 clusters.

4.2. MCE-training of Parameters in ESS Super-vector

Next we verify that using MCE-trained ESS super-vector provides a better accuracy than using the original ML-trained ESS super-vector. Here we presented results for the second procedure described in Section 3.2. We tested performance using a modified ETSI advanced front-end (AFE) suggested in [11] where the log-energy feature of each frame is replaced with the C0 coefficient. Every feature vector consists of 13 static plus their first and second order time derivatives. A complex back-end model topology suggested in [4] is used, where there are 20 mixtures per state for the digits and 36 mixtures per state for the silence and short pause.

4.2.1. Gender Independent System

The experimental results for the GI system are listed in Table 2. "GI-Baseline+ML" and "GI-Baseline+MCE" indicate that the recognition results using multicondition-trained HMMs with ML and MCE, respectively, while "GI-ESSEM+ML" and "GI-ESSEM+MCE" indicate recognition results for ESSEM with ML- and MCE-trained ESS super-vectors, respectively.

It can be seen from Table 2 that ML-trained ESSEM provides a performance improvement of 12.85% relative WER reduction (from 6.46% to 5.63% WER) when using such a more complex HMM topology. Next by comparing "GI-Baseline+ML" with "GI-Baseline+MCE", we observed that the improvement provided by using a MCE parameter refined multicondition HMM set is not significant (from 6.46% WER to 6.33% WER). Next, it can be easily observed that "GI-ESSEM+MCE" achieves clear improvements of 16.56% and 14.85% relative WER reductions over "GI-Baseline+ML" and "GI-Baseline+MCE", respectively. By comparing results of the "GI-ESSEM+ML" and "GI-ESSEM+MCE" tests, we verified that with a MCE-trained ESS super-vector, performance of ESSEM can be further enhanced.

	Set A	Set B	Set C	Overall
GI-Baseline+ML	5.92	6.69	7.11	6.46
GI-Baseline+MCE	5.85	6.18	7.60	6.33
GI-ESSEM+ML	5.12	6.07	5.78	5.63
GI-ESSEM+MCE	4.94	5.61	5.83	5.39

Table 2. Average word error rates (in %) from 0dB to 20dB.

4.2.2. Gender Dependent System

Finally we tested ESSEM on the GD system. We use the same procedure as described in Sec.4.1.2 for CS(1) to obtain the results for the "GD-Baseline+ML". Similar to the GI case, ESSEM with an MCE-trained ESS super-vector achieves a better performance than an ML-trained ESS one, so only results with the MCE-trained ESS super-vector, denoted as "GD-ESSEM+MCE", are listed in Table 3. We find that "GD-ESSEM+MCE" achieves a WER reduction of 9.44% over the "GD-Baseline+ML". (5.51% to 4.99%). This 4.99% WER is the best result we achieved using both tree structure clustering and MCE training for ESS super-vector.

	Set A	Set B	Set C	Overall
GD-Baseline+ML	5.11	5.38	6.56	5.51
GD-ESSEM+MCE	4.64	5.05	5.56	4.99

Table 3. Average word error rates (in %) from 0dB to 20dB.

5. CONCLUSION

We propose two ESSEM extensions with tree structure speaker and speaking environment clustering and MCE training to improve environment characterization and enhance ASR performance robustness. The framework is evaluated on the Aurora2 database. We first verified that clustering can be a better dimension reduction technique and can provide more suitable *a priori* knowledge to the testing environments than PCA. We then show that MCE-trained ESSEM is better than ML-trained cases. By integrating these two methods into the ESSEM framework, we achieves our best WERs of 5.39% and 4.99% WERs, corresponding to 16.56% and 9.44% relative WER reductions over the ML-trained baselines in the GI and GD systems, respectively.

6. ACKNOWLEDGEMENT

This work was supported by a Texas Instruments Leadership University (TILU) grant.

7. REFERENCES

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

[2] 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, pp.171-185, 1995.

[3] D. V. Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, vol. 3, pp.151-167, 1989.

[4] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *Proc. ICSLP'2002*, Denver, CO, 2002, pp. 17–20.

[5] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transaction on Speech and Audio Processing*, vol. 4, pp.190-202, May.1996.

[6] Y. Tsao and C.-H. Lee, "An ensemble modeling approach to joint characterization of speaker and speaking environments," in *Proc. Interspeech*, Aug. 2007.

[7] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Transaction on Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.

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

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

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

[11] J. Wu and Q. Huo, "Several HKU approaches for robust speech recognition and their evaluation on Aurora connected digit recognition tasks," in *Proc. Eurospeech03,* 2003.