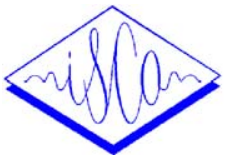
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A STUDY OF MINIMUM CLASSIFICATION ERROR TRAINING FOR SEGMENTAL SWITCHING LINEAR GAUSSIAN HIDDEN MARKOV MODELS

Jian Wu, Donglai Zhu and Qiang Huo

Department of Computer Science and Information Systems
The University of Hong Kong, Pokfulam Road, Hong Kong, China
(Email: jwu@csis.hku.hk, dlzhu@csis.hku.hk, qhuo@csis.hku.hk)

10.21437/Interspeech.2004-725

ABSTRACT

In our previous works, a Switching Linear Gaussian Hidden Markov Model (SLGHMM) and its segmental derivative, SSLGHMM, were proposed to cast the problem of modelling a noisy speech utterance by a well-designed dynamic Bayesian network. We presented parameter learning procedures for both models with maximum likelihood (ML) criterion. The effectiveness of such models was confirmed by evaluation experiments on Aurora2 database. In this paper, we present a study of minimum classification error (MCE) training for SSLGHMM and discuss its relation to our earlier proposals based on stochastic vector mapping. An important implementation issue of SSLGHMM, namely the specification of switching states for a given utterance, is also studied. New evaluation results on Aurora3 database show that MCE-trained SSLGHMMs achieve a relative error reduction of 21% over a baseline system based on ML-trained continuous density HMMs (CDHMMs).

1. INTRODUCTION

A Switching Linear Gaussian Hidden Markov Model (SLGHMM), as shown in Fig. 1(a), was proposed in [5] to compensate for the nonstationary distortion that may exist in a speech utterance to be recognized. It is a hybrid Dynamic Bayesian Network (DBN) with two coupled streams of dynamic models. One stream is a Continuous Density HMM (CDHMM) to model the generic linguistic information of clean speech $X = \{x_t\}$. Another stream is a Switching Linear Gaussian model to model the nonstationary distortion mechanism with a set of parallel linear Gaussian dynamic streams, $B^{(k)} = \{b_t^{(k)}\}$ (each representing a possible additive stationary distortion in feature vector space), and a discrete-state Markov chain, $Q = \{q_t\}$ (controlling the choice of the distortion source at each time step). An SLGHMM with such a mechanism, is thus able to model approximately the distribution of speech, $Y = \{y_t\}$ corrupted by switching-condition distortions. In [7], a variational approach has been proposed to solve the approximate maximum likelihood (ML) parameter learning and probabilistic inference problems for SLGHMMs. Unfortunately, it is not computationally feasible for ASR applications that require prompt response. Therefore, a Segmental SLGHMM (SSLGHMM hereinafter), as illustrated in Fig. 1(b), was proposed in [5].

In an SSLGHMM, several assumptions are made to simplify the model. Each switch state q_t is assumed to be independent

This research was supported by grants from the RGC of the Hong Kong SAR (Project Numbers HKU7022/00E and HKU7039/02E).

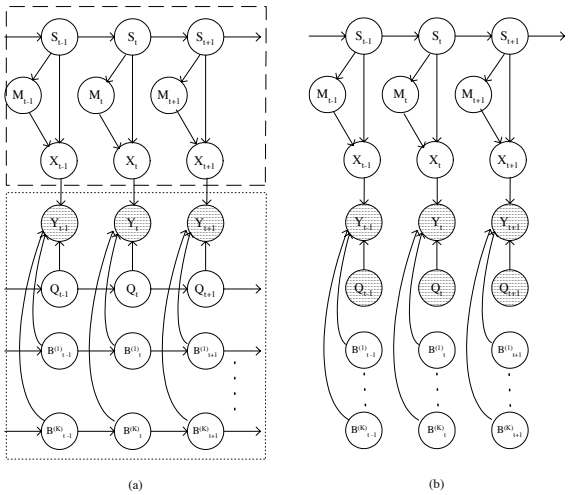


Fig. 1. Directed acyclic graph specifying conditional independence relations for (a) SLGHMM and (b) SSLGHMM

of all switch states at other time and they are treated as observations of this DBN. The values of q_t are assigned by an appropriate pre-segmentation procedure. For a particular stream k, all of $b_t^{(k)}$'s are assumed to follow an i.i.d. (identical independently distributed) distribution $\mathcal{N}(b_t^{(k)}; r^{(k)}, \Xi^{(k)})$. It is assumed that $y_t = x_t + b_t^{(k)} + u_t$ given $q_t = k$, where u_t is a zero-mean Gaussian noise with diagonal covariance matrix Ω. It is further assumed that each CDHMM $\Lambda = \{a_{ij}, \pi_i, c_{sm}, \mu_{sm}, \Sigma_{sm}, i, j, s = 1, ..., S, m = 1, ..., M\}$ of the recognizer is fixed but unknown. It consists of S states with transition probability a_{ij} from state i to state j. Each state has M Gaussian components with D-dimensional mean vectors μ_{sm} and diagonal covariance matrices Σ_{sm}. c_{sm} denotes the weight of m-th Gaussian component in the s-th state. Then, the joint likelihood of observations and hidden variables given a set of particular parameters of SSLGHMM, Γ, is

$$\begin{aligned} & p(\mathbf{Y}, \mathbf{X}, \mathbf{B}, \mathbf{S}, \mathbf{M}, \mathbf{Q}|\Gamma) \\ &= \pi_{s_1} \prod_{t=2}^{T} a_{s_t, s_{t-1}} \cdot \prod_{t=1}^{T} c_{s_t m_t} \mathcal{N}(x_t; \mu_{s_t m_t}, \Sigma_{s_t m_t}) \\ & \cdot \prod_{t=1}^{T} \mathcal{N}(y_t; x_t + b_t^{(q_t)}, \Omega) \cdot \mathcal{N}(b_t^{(q_t)}; r^{(q_t)}, \Xi^{(q_t)}) \,, \end{aligned} \quad (1)$$

from which the marginal distribution of Y given parameters Γ can

be easily derived as follows:

$$p(\mathbf{Y}|\Gamma) = \sum_{J} A_J^* \prod_{t=1}^{T} \sum_{m=1}^{M} c_{s_t m} \mathcal{N}(y_t; \mu_{s_t m} + r^{(q_t)}, \Sigma_{s_t m} + \Xi^{(q_t)} + \Omega) \ , \quad (2)$$

where A_J^* is the product of transition probabilities given the state sequence J of the CDHMMs for the word sequence W associated with the observation $\mathbf{Y}$. In [5], a procedure for ML estimation of SSLGHMM parameters was presented. Considering the fact of that many HMM-based ASR systems using minimum classification error (MCE) training (e.g.[2]) can offer a further error reduction in comparison with that based on ML training, it would be interesting to verify whether an even better performance can be achieved by an ASR system with MCE-trained SSLGHMM parameters.

In the following, we first present a general formulation for MCE training of SSLGHMM parameters and discuss its relation to our previous proposals based on stochastic vector mapping. Then we discuss an important implementation issue, namely how to specify the values of switching states for a given utterance. Finally, experimental results on Aurora3 database are used to demonstrate the effectiveness of the MCE training for SSLGHMM.

2. MCE TRAINING OF SSLGHMM PARAMETERS

2.1. General Formulation

As mentioned above, in our proposed Segmental Switching Linear Gaussian HMMs, both the CDHMM parameters and the parameters of the switching linear Gaussian models are unknown. In order to find their optimal values in terms of achieving the minimum empirical classification error on training set $\mathcal{Y} = \{Y_i\}_{i=1}^{I}$, the following objective function can be defined and used:

$$\ell(\Gamma) = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{1 + exp(-\alpha d(Y_i; \Gamma) + \beta)} \ , \quad (3)$$

where α and β are two control parameters. In the above equation, $d(\cdot)$ is a misclassification measure defined as

$$d(Y; \Gamma) = -\tilde{g}(Y; \Gamma, W_c) + \tilde{G}(Y; \Gamma) \ , \quad (4)$$

with

$$\tilde{G}(Y; \Gamma) = \frac{1}{\eta} log\{\frac{1}{N} \sum_{n=1}^{N} exp[\eta \cdot \tilde{g}(Y; \Gamma, W_n)]\},$$

where η is a positive control parameter, W_c and $\{W_n\}$ are the correct word sequence and the N-best competing word sequences of the training utterance Y, respectively. Similar to the MCE training of CDHMM parameters, the discriminant function of $\tilde{g}(Y; \Gamma, W)$, for instance, is defined as the logarithm of the likelihood function $p(\mathbf{Y}|\Gamma)$ shown in Eq.(2), given the word sequence W.

Given the objective function in Eq.(3), a sequential gradient descent algorithm, known as GPD algorithm, is usually used (as in our experiments) to update the parameters iteratively. Of course, other optimization algorithms may also be used to minimize the above objective function. No matter which approach is used, the partial derivative of the discriminant function with respect to the parameters to be estimated, $\partial\tilde{g}/\partial\Gamma$, need be calculated utterance by utterance. In the following, we present the formulae related to these partial derivatives.

2.2. Partial Derivatives of The Discriminant Function

In order to maintain the constraints for both the CDHMM parameters and the parameters of the switching linear models, the following parameter transformations, which are similar to that described in [2], are applied to the relevant parameters during updating (For simplicity, we only list the formulae related to $\mu_{sm}(= [\mu_{sm1}, ..., \mu_{smD}]^{T_r})$, $\Sigma_{sm}(= diag[\sigma_{sm1}^2, ..., \sigma_{smD}^2])$, $r_k(= [r_{k1}, ..., r_{kD}]^{T_r})$ and $\Xi_k(= diag[\tau_{k1}^2, ..., \tau_{kD}^2])$, while $\Omega(= diag[\kappa_1^2, ..., \kappa_D^2])$ is set as a fixed matrix):

$$\tilde{\mu}_{smd} = \frac{\mu_{smd}}{\sigma_{smd}} \ , \quad (5)$$

$$\tilde{\sigma}_{smd} = \log \sigma_{smd} \ , \quad (6)$$

$$\tilde{r}_{kd} = \frac{r_{kd}}{\tau_{kd}} \ , \quad (7)$$

$$\tilde{\tau}_{kd} = \log \tau_{kd} \ . \quad (8)$$

Therefore, given a training utterance $Y = \{y_1, y_2, ..., y_T\}$, the current value of the partial derivatives of discriminant function with respect to the above parameters are as follows:

$$\frac{\partial\tilde{g}}{\partial\tilde{\mu}_{smd}} = \sum_{t=1}^{T} \tilde{\zeta}_t(s, m) \left(\frac{y_{td} - r_{q_t d} - \mu_{smd}}{\sqrt{\sigma_{smd}^2 + \tau_{q_t d}^2 + \kappa_d^2}} \right) \ , \quad (9)$$

$$\frac{\partial\tilde{g}}{\partial\tilde{\sigma}_{smd}} = \sum_{t=1}^{T} \tilde{\zeta}_t(s, m) \left[\frac{(y_{td} - r_{q_t d} - \mu_{smd})^2}{\sigma_{smd}^2 + \tau_{q_t d}^2 + \kappa_d^2} - 1 \right] \cdot \frac{\sigma_{smd}^2}{\sigma_{smd}^2 + \tau_{q_t d}^2 + \kappa_d^2} \ , \quad (10)$$

$$\frac{\partial\tilde{g}}{\partial\tilde{r}_{kd}} = \sum_{t=1}^{T} \sum_{s,m} \tilde{\zeta}_t(s, m) \left(\frac{y_{td} - r_{q_t d} - \mu_{smd}}{\sqrt{\sigma_{smd}^2 + \tau_{q_t d}^2 + \kappa_d^2}} \right) \cdot \delta(q_t - k) \ , \quad (11)$$

$$\frac{\partial\tilde{g}}{\partial\tilde{\tau}_{kd}} = \sum_{t=1}^{T} \sum_{s,m} \tilde{\zeta}_t(s, m) \left[\frac{(y_{td} - r_{q_t d} - \mu_{smd})^2}{\sigma_{smd}^2 + \tau_{q_t d}^2 + \kappa_d^2} - 1 \right] \cdot \frac{\tau_{q_t d}^2}{\sigma_{smd}^2 + \tau_{q_t d}^2 + \kappa_d^2} \cdot \delta(q_t - k) \ , \quad (12)$$

where $\tilde{\zeta}_t(s, m)$ denotes the occupation probability of Gaussian component m in state s at time t, whose formula can be found in [7]; and $\delta(\cdot)$ denotes the Kronecker delta function.

Using the above equations, the transformed parameters can be updated iteratively. After each update of the transformed parameters, the original parameters will be obtained by applying the inverse transform of Eqs.(5)-(8).

2.3. Relation to Our Previously Published Approaches

In order to improve the performance of ASR in a mismatched situation, usually either a feature compensation module is embedded as part of the front-end processing which aims at transforming the feature vector of noisy speech to that of pseudo clean speech, or a model compensation procedure is used to derive a more accurate model to model the noisy speech. In both cases, an exact knowledge of the mismatch mechanism is often unavailable, thus the compensation results can not be fully trusted. Based on this reasoning, several research directions were pursued to consider the uncertainty during recognition, one of which is to consider the uncertainty in feature space during recognition (e.g. [4]).

Generally speaking, the feature compensation can be regarded as a process of feature transformation $\mathcal{F}_\Theta(\cdot)$ with parameter Θ, which estimates the pseudo clean speech feature vector $\hat{x} = \mathcal{F}_\Theta(y)$ from the noisy speech feature vector y. In [4], we consider the uncertainty of the transformation parameters Θ by treating them as if they were random. Their prior uncertainty is modelled by a joint *a priori* probability density function (pdf) $p(\Theta|\varphi_\Theta)$, with $\Theta \in \Omega_\Theta$, where Ω_Θ denotes admissible region of possible Θ, and φ_Θ is the set of unknown parameters of the prior pdf. Given a noisy speech utterance with a feature vector sequence Y, we define the following discriminant function of Y for word sequence W to take into account the uncertainty of the above transformation parameters Θ:

$$\tilde{g}(Y;\Lambda,\varphi_\Theta,W) \triangleq \log p(\hat{X}|\Lambda,W) = \log \sum_J p(\hat{X},J|\Lambda,W)$$
$$= \log \sum_J A_J^* \prod_{t=1}^{T} p(\hat{x}_t|s_t,\Lambda,\varphi_\Theta,W), \quad (13)$$

where J denotes a possible state sequence, and

$$p(\hat{x}_t|s_t,\Lambda,\varphi_\Theta,W) = \int_{\Omega_\Theta} p(\hat{x}_t|s_t,\Lambda,\Theta,W)p(\Theta|\varphi_\Theta)d\Theta$$

is the marginal state observation pdf of $\hat{x}_t$. The recognizer will choose the word sequence $\hat{W}$ as the recognition result that produces the maximum value of the above discriminant function.

In our previous work [3], we adopted a specific stochastic vector mapping function for the feature compensation. Consequently, the feature vector of pseudo clean speech is estimated as follows:

$$\hat{x} \triangleq \mathcal{F}_\Theta(y) = y + b_k \ , \quad (14)$$

where

$$k = \arg\max_{j=1}^{K} p(j|y) = \arg\max_{j=1}^{K} \frac{p(j)p(y|j)}{\sum_{j'=1}^{K} p(j')p(y|j')} \ , \quad (15)$$

and $\Theta = \{b_k\}_{k=1}^K$ is the set of mapping function parameters (also referred to as *correction vectors*) associated with the K Gaussian components $p(y|k)$'s. Moreover, each correction vector b_k is assumed to follow a normal pdf $\mathcal{N}(b_k; r_k, \Xi_k)$ with mean vector r_k and diagonal covariance matrices Ξ_k. With these assumptions, a specific discriminant function can be derived as

$$\tilde{g}(Y;\Lambda,\varphi_\Theta,W) = \log \sum_J A_J^* \prod_{t=1}^{T} \sum_{m=1}^{M} c_{s_t m} \cdot \quad (16)$$
$$\mathcal{N}(y_t; \mu_{s_t m} - r_{k_t}, \Sigma_{s_t m} + \Xi_{k_t}) \ ,$$

where k_t denotes the index of correction vector chosen at time t, and $\varphi_\Theta = \{r_k, \Xi_k\}$.

By comparing Eq.(16) and Eq.(2), it is observed that if Ω in Eq.(2) is fixed, the MCE learning of Segmental SLGHMM is essentially equivalent to considering the uncertainty of the particular feature compensation module in designing a new discriminant function for speech recognition. This explains why we refer readers of [5, 6] to [4] for the formulation of MCE training of SSLGHMM.

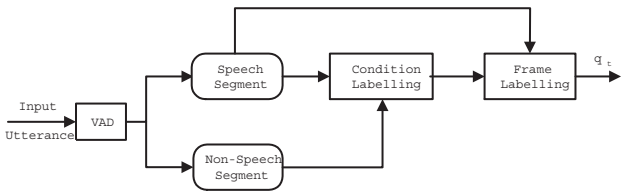


Fig. 2. Process of switching state segmentation in SSLGHMMs

3. SWITCHING STATE SEGMENTATION

As mentioned above, one of most important modifications of Segmental SLGHMM is to explicitly assign a proper value for the switching state q_t at each time step by an extra environment identification process, instead of treating it as a hidden variable in general SLGHMM. The intuition of such a modification comes from the observation of the success of using environmental partition techniques in "Broadcast News" tasks (e.g., [1]), where speech frames are segmented and then clustered into several acoustic conditions before training and recognition. It is demonstrated by experiments that the condition dependent model will help improve the discriminability among the basic speech units to be recognized, even though sometimes the environment detection is far from perfect. Indeed, the general framework of SLGHMM was proposed to automatically learn the underlying switching structure of environment purely from noisy observation. But unfortunately, it is computationally inefficient for the exact inference process as explained in [7]. Segmental SLGHMM trades it off with an extra "flawed" detection process, which is separated from the learning of the other model parameters.

3.1. Process of Switching State Segmentation

We have developed several possible solutions for switching state segmentation of SSLGHMMs in noise-robust speech recognition. Their details and comparison will be reported elsewhere. In the following, we will only examine one of them, and demonstrate its effectiveness by an evaluation on Aurora3 database.

As shown in Fig. 2, the value of switching state, q_t, is determined in two stages: *condition labelling* and *frame labelling*. The *condition labelling* process makes use of the results of voice activity detection (VAD) to segment the input signal into speech and non-speech segments. Then the closest one out of E conditions is chosen based on the whole input utterance and two sets of Gaussian Mixture Models (GMMs), $\Lambda^S = \{\Lambda_e^S\}$ and $\Lambda^N = \{\Lambda_e^N\}$, designed for both speech and non-speech segments, respectively:

$$e^* = \arg\max_e p(Y|e) = \arg\max_e \prod_{t=1}^{T} p(y_t|\Lambda_e^{VAD(t)}) \ , \quad (17)$$

where VAD(t) represents the result of VAD at time t (either "S" or "N"). In this study, non-speech segments are assumed to be irrelevant to the condition labelling of each utterance, thus only the likelihood of speech frames are counted in the above equation. After the condition labelling for the whole utterance, the *frame labelling* process will determine the value of switching state at each time t with the corresponding condition dependent GMM: $\Lambda_{e^*}^S$, which is similar to the procedure described by Eq. (15): $q_t = \arg\max_q p(q|y_t, \Lambda_{e^*}^S)$.

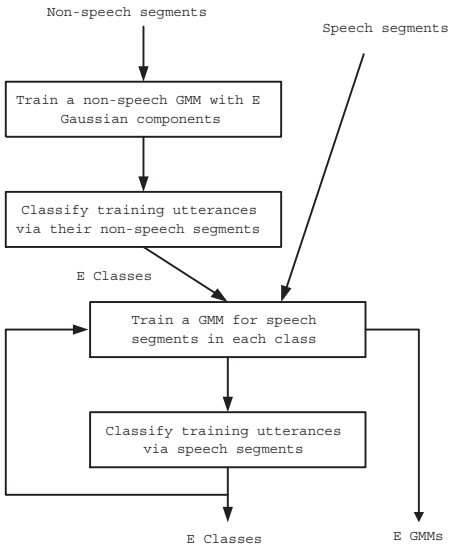


Fig. 3. Training of Λ^S (E GMMs in the figure) for condition labelling of a speech segment.

3.2. Training of GMMs for Switching State Segmentation

In this study, since we assume that non-speech segments have no contribution to the condition labelling and the environmental condition is unchanged within the same utterance, we use a procedure as illustrated in Fig. 3 for training Λ^S. A GMM with E Gaussian components is first created by using non-speech segments of all training utterances. Each training utterance is then classified to one of E classes that has the maximum likelihood of non-speech segments for the associated Gaussian component. Given this initial labelling of training utterances, a GMM Λ_e^S is trained for each condition class with all the speech segments associated with that class. Given $\{\Lambda_e^S\}$, each training utterance can be re-classified into the corresponding condition with the maximum likelihood of its speech segments against Λ_e^S. The above two steps can be iterated until no change of utterance labelling results or a maximum number of iterations is reached.

4. EXPERIMENTAL RESULTS

This section presents the application of MCE trained Segmental SLGHMM to the automatic recognition of connected digit string recorded in noisy environments such as in a moving vehicle. The database used to verify our algorithm is the Aurora3 database (the so-called "well-matched" condition, WM), which contains versions of four European languages: Finnish, Spanish, German and Danish. The front-ends used in this evaluation are detailed in our previous paper [6]. In the baseline systems, which are shown as "ML-HMM" and "MCE-HMM" in Table 1, a CDHMM with 16 emitting states, each having 3 Gaussian mixture components with diagonal covariance matrices, is trained for each digit under either ML or MCE criterion. Similarly, the results of other two systems based on SSLGHMMs are shown as "ML-SSLGHMM" and "MCE-SSLGHMM", in which model parameters are trained by ML and MCE criteria, respectively. The formulae of ML estimation for SSLGHMM parameters can be found in [5]. For SSLGHMMs, the exactly same model structure as that in CDHMM-based baseline system is used in its CDHMM portion, while the total number of linear Gaussian streams are 256, i.e., the number of conditions, E, as illustrated in Fig. 3 is 8 and speech frames of each condition are further classified into 32 classes.

Table 1. Aurora3 Word Error Rate (in %) on WM Condition

	Fin.	Span.	Ger.	Dan.	Ave.
ML-HMM	3.95	3.39	4.87	6.02	4.56
ML-SSLGHMM	3.36	3.17	4.49	5.73	4.19
MCE-HMM	2.82	2.84	4.69	5.27	3.91
MCE-SSLGHMM	2.30	2.73	4.19	5.15	3.59

From the results, it is observed that, as expected, the MCE trained models, no matter whether HMM or SSLGHMM is used, can achieve a better performance than their ML trained counterparts by about 14.3% relative error reduction. Furthermore, with the help of an advanced structure like SSLGHMM, the error reduction over the conventional CDHMM is more than 8%, no matter which criterion is used in parameter learning. Our MCE-trained SSLGHMMs achieve a relative error reduction of 21% over the baseline system based on ML-trained CDHMMs.

5. SUMMARY

In this paper, a formulation of MCE training for SSLGHMM is presented and its relation to our previously published works is clarified. New evaluation results on Aurora3 database are reported that confirm our findings on Aurora2 task. SSLGHMM is a promising new model for improving the accuracy and robustness of the current generation CDHMM-based ASR systems. We are studying the effectiveness of SSLGHMM in a subword-based large vocabulary ASR application and will report those results in future.

6. REFERENCES

[1] R. Bakis, S. Chen, P. Gopalakrishan and et al. "Transcription of broadcast news – system robustness issues and adaptation techniques", *Proc. ICASSP-1997*, pp.711-714. 1997.

[2] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. on SAP*, Vol.5, No.3, pp.257-265, 1997.

[3] J. Wu and Q. Huo, "An environment compensated minimum classification error training approach and its evaluation on Aurora2 database," *Proc. ICSLP-2002*, pp.I-453-456, 2002.

[4] J. Wu and Q. Huo, "Modelling uncertainty in stochastic vector mapping with minimum classification error training for robust speech recognition," *Proc. ICASSP-2003*, Hong Kong.

[5] J. Wu and Q. Huo, "A switching linear Gaussian hidden Markov model and its application to nonstationary noise compensation for robust speech recognition", *Proc. Eurospeech-2003*, 2003.

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

[7] J. Wu, "Discriminative speaker adaptation and environmental robustness in automatic speech recognition", Ph.D thesis, the Department of Computer Science and Information Systems, The University of Hong Kong, Jan. 2004.