A Robust Compensation Strategy for Extraneous Acoustic Variations in Spontaneous Speech Recognition

Hui Jiang, *Member, IEEE,* and Li Deng, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a robust compensation strategy to deal effectively with extraneous acoustic variations for spontaneous speech recognition. This strategy extends speaker adaptive training, and uses hidden Markov models (HMM) parameter transformations to normalize the extraneous variations in the training data according to a set of predefined *conditions*. A "compact" model and the associated prior probability density functions (PDFs) of transformation parameters are estimated using the maximum likelihood criterion. In the testing phase, the generic model and the prior PDFs are used to search for the unknown word sequence based on Bayesian prediction classification (BPC). The proposed strategy is evaluated in the switchboard task, and is used to deal with three types of extraneous variations and mismatch in conversational speech recognition: pronunciation variations, inter-speaker variability, and telephone handset mismatch. Experimental results show that moderate word error rate reduction is achieved in comparison with a well-trained baseline HMM system under identical experimental conditions.**

***Index Terms*—Bayesian predictive classification (BPC), extraneous variation, generic (or compact) model, prior PDF, speaker-adaptive training (SAT).**

I. Introduction

In the past few decades, statistical models, such as hidden Markov models (HMM), have achieved significant success in automatic speech recognition (ASR); see some recent review papers in [15]. In the conventional statistical paradigm of ASR, statistical models are usually estimated based on a large amount of training data. Then the estimated models are used to recognize unknown utterances. The training data are usually collected under as many different conditions as possible for the purpose of properly representing all possible incoming speech data in the future use. Even though the data collection conditions may greatly differ due to a wide range of factors, the conventional paradigm treats all training data collected in different conditions in an identical manner by simply pooling them together. The model parameters are then determined from the pooled data set via parameter estimation techniques, e.g., maximum likelihood (ML) or discriminant training. The variations contained in the data come from many different sources, and some of these sources are germane to the recognition problem or task while the others are extraneous. An apparent shortcoming of the above training paradigm is that the large amount of pooled training data not only include the pertinent variability (such as phonetic distinction), but also involve many other extraneous variations which are irrelevant to our modeling or recognition purpose and should therefore be compensated for. In this paper, we call those variations existing in the data which are not directly related to our modeling or recognition purpose as *extraneous variations*. Obviously, extraneous variation varies from problem to problem. For instance, in a typical case of speech recognition, it is important to model the phonetically relevant variation sources. All other variabilities are considered to be extraneous, including those arising from speaker, transducer, telephone channel, speaking style, speaking rate, pronunciation change, etc. On the other hand, for the speaker recognition problem, speaker variations become pertinent while other variations are extraneous. The extraneous variations have several realization levels. In this paper we consider the extraneous variations at the acoustic level only. All other issues related to phonetic or higher levels will be beyond the scope of this paper.

In conventional implementation of speech recognizers, one does not have an explicit mechanism to compensate the extraneous variations in the training procedure. In particular, when we recognize spontaneous speech, where many types of extraneous variations abound, the performance of speech recognition can be significantly affected. In the training phase, due to the extraneous variations, the training data may diverge substantially from what is assumed in the model. This would make the estimated models diverge from the desired behavior. In the testing phase, the deviation due to the extraneous variations can also be viewed as a special kind of mismatch between the models and the testing data. In this paper, we describe a robust strategy to deal with the extraneous acoustic variations in the training phase in order to achieve a "generic" (or compact) model which better reflects the pertinent variations in the speech recognition tasks.

Recently, researchers began to realize the importance of compensating the extraneous variations in the training phase in order to improve the modeling capability of the models. In [1], the "speaker-adaptive training" (SAT) by Anastasakos et al. is one of important steps along the direction of the robust training strategy. In SAT, some linear regression transformations are

Manuscript received December 21, 1999; revised September 24, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafid A. Sukkar.

H. Jiang was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1. He is currently with the Dialogue Systems Research, Multimedia Communication Research Laboratory, Bell Labs, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: hui@research.bell-labs.com).

L. Deng is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1. He is currently with Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com).

Publisher Item Identifier S 1063-6676(02)00298-5.

used to normalize the inter-speaker variations in the speech data to construct a "compact" model. In [1], an iterative algorithm has been proposed to estimate the parameters of both transformation and compact models in a sequential mode based on the ML criterion. The work reported in [8] shows another way to normalize "irrelevant variability" in the training phase for the purpose of learning a model structure (HMM state tying). Further, in [5], [6], another interesting robust training method was proposed for the same purpose as aimed by the work reported in this paper. In that work, several clusters are pre-defined and a canonical model is estimated for each cluster based on the ML criterion in the training phase. During testing, an interpolated model of all cluster-specific canonical models is constructed for each separate test utterance. The interpolating weights are estimated on-line from each current utterance.

In this paper, we propose and evaluate a new robust training strategy to compensate and normalize the extraneous variations, with a solid theoretical foundation and with practical effectiveness. It differs from the previous work discussed above in its novel use of the *distribution* of the transformation parameters in a Bayesian framework. Briefly, we label each utterance in the training set with one of pre-defined *conditions*, depending on the nature of the extraneous variation to be compensated, such as speaker *id*, speaking style, pronunciation, transducer, transmission channel, etc. The data from different conditions are first normalized by using some appropriate transformations before they are pooled to estimate a "generic" (or compact) model. Meanwhile, a prior distribution of transformation parameters is also automatically estimated from the data to represent the knowledge of all possible transformations used across the various "conditions" in the training phase. In this way, the extraneous variation is adequately compensated for and the generic model will converge properly to represent the pertinent variations in question. In the testing/decoding phase, based on the generic model and the prior distribution of transformation parameters, we use a new search algorithm to decode the unknown input utterance according to Bayesian predictive classification (BPC) [9], [10].

In order to obtain the "generic" acoustic models which can adequately describe phonetically relevant variation sources, the proposed strategy is used to normalize and/or compensate several types of major extraneous variations in spontaneous speech recognition. Throughout this paper, we take the switchboard corpus as our evaluation data set. There are several interesting aspects in this corpus for the evaluation of our new robust training strategy. Firstly, in the switchboard task, the pronunciation variation in conversational speech is shown to be one major extraneous variation source hampering speech recognition. Thus, we can justifiably define the "condition" which characterizes the pronunciation variation, and in this case the proposed strategy can be employed to compensate for the pronunciation variation. Secondly, like SAT, we utilize our robust training strategy to normalize the inter-speaker differences that also clearly exhibit themselves in the switchboard corpus. Here, the "condition" is defined based on the speaker id. Thirdly, the robust training strategy is also used to normalize the extraneous variation related to the mismatches caused by different telephone handsets in the switchboard corpus. Here, the information about the telephone number of each participant in both conversation sides in the switchboard corpus is used to define the "condition". To facilitate the implementation, we choose a very simple transformation, i.e., piecewise linear functions, to normalize and/or compensate all of these three types of extraneous acoustic variations in the conversational telephony speech data of the switchboard corpus. Experimental results show that the proposed method has achieved some moderate improvement in recognition performance, i.e., nearly 1% absolute reduction in word error rate (WER) for each type of extraneous variations, over a well-trained baseline HMM system.

The remainder of the paper is organized as follows. First, the basic ideas underlying the proposed robust training strategy are presented in Section II. Next, the robust training and decoding algorithms are presented in detail in Sections III and IV, respectively. In Section V, the experiments on the switchboard task are reported and the results are discussed. Finally, the paper is concluded with a summary of our findings in Section VI.

II. Overview of the New Strategy

Following the idea originally presented in [1], suppose we have a generic (or compact) mixture Gaussian HMM $\lambda_c = (\pi, A, \theta)$ for each speech unit W we desire to model, where π is the initial state distribution, $A = \{a_{ij} \,|\, 1 \le i,j \le N\}$ is the transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{w_{ik}, m_{ik}, r_{ik}\}_{k=1,2,\ldots,K}$ for each state i, where K denotes the number of Gaussian mixture components in each state. The state observation probability denisity functions (PDF) is assumed to be a mixture of multivariate Gaussian PDFs with diagonal precision matrices

$$p_i(x) = \sum_{k=1}^{K} w_{ik} \cdot \prod_{d=1}^{D} \sqrt{\frac{r_{ikd}}{2\pi}} \exp\Big[-\frac{r_{ikd}}{2}(x_d - m_{ikd})^2\Big] \quad (1)$$

where D denotes the dimension of feature vectors. We denote all training data for W as $X = \{X^{(r)} \,|\, r = 1, 2, \ldots, R\}$, where $X^{(r)}$ stands for those data collected under condition r and we have a total of R different conditions. The condition is defined according to the extraneous variations to be normalized, which will be explained for the specific examples in detail in the following sections. Then, we aim to choose some proper transformations to normalize/compensate the extraneous variations in speech signals. In other words, we need to choose a set of transformations $\{T_\eta^{(r)}(\cdot) \,|\, r = 1, 2, \ldots, R\}$, for the generic model λ_c. Each transformation $T_\eta^{(r)}(\cdot)$, which is parameterized by η, corresponds to a specific condition r so that for each condition r the transformed model $T_\eta^{(r)}(\lambda_c)$ gives a better description for the data $X^{(r)}$ which are collected under this condition r $(r = 1, 2, \ldots, R)$. The same SAT algorithm in [1] could be used to estimate the compact model λ_c and the corresponding transformations $T_\eta^{(r)}(\cdot)$ according to the ML criterion. However, in the testing phase, it would not be appropriate to use the compact model λ_c to evaluate the testing data directly because λ_c would not match the original data due to the involved transformations. Furthermore, we do not know which transformation should be used for each testing utterance because we

have no idea of which *condition* the test data come from. In this paper, the idea of Bayesian prediction is proposed to solve this problem. The specific transformation parameters η for different "conditions" are viewed as some sampling outcomes from a prior PDF for the transformation parameters, denoted as $\rho(\eta)$. In the training stage, the prior $\rho(\eta)$ is simultaneously estimated to represent the knowledge of all transformations possibly used in the training stage. In the testing phase, the BPC algorithm helps to make an optimal decision given the information supplied by the prior $\rho(\eta)$.

Before we derive the robust training and testing algorithms, we have to carefully determine the functional form for the transformation $T_\eta^{(r)}(\cdot)$ and that for the prior PDF of transformation parameters $\rho(\eta)$. Firstly, as for the transformation $T_\eta^{(r)}(\cdot)$, the requirements are 1) the transformation is sufficiently powerful to normalize the acoustic difference caused by extraneous variations and 2) The transformation form is simple enough so that Bayesian prediction is tractable in the decoding phase. One possible choice is the piecewise linear transformation. In this work, as a first step, we choose the simplest transformation form, namely the bias vector plus the mean vector of an HMM

$$m'_{ikd} = m_{ikd} + \beta_d \quad (d = 1, 2, \ldots, D) \tag{2}$$

where $\vec{\beta} = \{\beta_1, \beta_2, \ldots, \beta_d\}$ denotes the transformation parameters. We assume that all other HMM parameters remain unchanged. In principle, each transformation could be related or tied to any different segments of speech signal. In this paper, we suppose that each transformation is HMM state-dependent, i.e., we use different transformations for different HMM states and the transformations of various states are tied based on the triphone state-tying in the entire HMM set. Secondly, as for the prior PDF of transformation parameters, i.e., $\rho(\vec{\beta})$ in this case, in order to have a simple form in the decoding stage, we choose the following prior PDF based on the concept of natural conjugate prior[3]

$$\rho(\vec{\beta}) = \prod_{d=1}^{D} \sqrt{\frac{\tau_d}{2\pi}} \exp\Big[-\frac{\tau_d}{2}(\beta_d - \mu_d)^2\Big] \tag{3}$$

where $\{\mu_d, \tau_d \,|\, d = 1, 2, \ldots, D\}$ are the hyperparameters.[1]

III. Robust Training Algorithm

In this section, we integrate the above robust training ideas into the conventional acoustic modeling method used in a large vocabulary speech recognition system, i.e., triphone model state tying based on the phonetic decision tree [20]. Our robust training strategy consists of the following steps.

1) Define a set of "conditions" according to the specific extraneous variations to be normalized or compensated. Specifically, we define a total of R different "conditions," and each is indexed by r $(1 \le r \le R)$.
2) Build a baseline system based on the conventional HMM approach.
3) Align all speech utterances in the whole training set to obtain the Viterbi segmentation for each utterance at the HMM's state level; Then, label each frame of feature vectors with one of the "condition" r $(1 \le r \le R)$ where the feature vector belongs.
4) Tying states of all triphone models and parameter estimation: build a single decision-tree for each state of phone models based on all data belonging to its corresponding triphone states. For each tied state of the tri-phone models (i.e., every leaf node in the decision tree).
   a) According to the above alignment results, pool all labeled data together, $X = \{X^{(r)} \,|\, r = 1, 2, \ldots, R\}$, where $X^{(r)}$ denotes all data labeled with condition r. Use the state distribution in the current leaf node as the initial estimate of the compact model λ_c for this tied state. Here, λ_c is a mixture Gaussian model, i.e., $\lambda_c = \{w_k, m_k, r_k \,|\, 1 \le k \le K\}$.
   b) Given the current λ_c, estimate all R transformations $\{T_\eta^{(r)}(\cdot) \,|\, r = 1, 2, \ldots, R\}$ for each condition r based on the data $X^{(r)} = \{x_t^{(r)} \,|\, 1 \le t \le T^{(r)}\}$ (See Appendix for derivation):
   For each dimension $d = 1, 2, \ldots, D$ (use $\beta^{(r)}[d] = 0$ as initialization)

$$\beta^{(r)}[d] = \frac{\sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \xi_t^{(r)}(k) \cdot r_{kd} \cdot \Big(x_{td}^{(r)} - m_{kd}\Big)}{\sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \xi_t^{(r)}(k) \cdot r_{kd}} \tag{4}$$

where $\xi_t^{(r)}(k)$ denotes the probability of $x_t^{(r)}$ residing in the mixture component $l_t = k$, i.e.,

$$\begin{aligned} \xi_t^{(r)}(k) &= \Pr\Big(l_t = k \,\Big|\, x_t^{(r)},\ \vec{\beta}^{(r)}\Big) \\ &= \frac{w_k \cdot \mathcal{N}\Big(x_t^{(r)} \,\Big|\, m_k + \vec{\beta}^{(r)}, r_k\Big)}{\sum_{k=1}^{K} w_k \cdot \mathcal{N}\Big(x_t^{(r)} \,\Big|\, m_k + \vec{\beta}^{(r)}, r_k\Big)}. \end{aligned} \tag{5}$$

   c) Re-estimate the compact model λ_c (See Appendix for derivation):
   For $1 \le k \le K$ and $1 \le d \le D$

$$m_{kd} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k) \cdot \Big(x_{td}^{(r)} - \beta^{(r)}[d]\Big)}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k)} \tag{6}$$

$$r_{kd} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k)}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k) \cdot \Big(x_{td}^{(r)} - m_{kd} - \beta^{(r)}[d]\Big)^2} \tag{7}$$

$$w_k = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k)}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \xi_t^{(r)}(k)}. \tag{8}$$

[1] It is also possible to use a finite mixture form for the prior distribution as in [11] to supply a more accurate description of the prior information. In the work described in this paper, we have implemented the simpler Gaussian form only.

d) Goto step (4b) unless some convergence conditions are met.

e) Estimate the hyperparameters $\{\mu_d, \tau_d \,|\, 1 \le d \le D\}$ of the prior PDF $\rho(\vec{\beta})$ for the current tied state: For $1 \le d \le D$

$$\mu_d = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T^{(r)}}\sum_{k=1}^{K}\xi_t^{(r)}(k)\cdot\beta^{(r)}[d]}{\sum_{r=1}^{R}\sum_{t=1}^{T^{(r)}}\sum_{k=1}^{K}\xi_t^{(r)}(k)} = \frac{\sum_{r=1}^{R}T^{(r)}\cdot\beta^{(r)}[d]}{\sum_{r=1}^{R}T^{(r)}} \tag{9}$$

$$\tau_d = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T^{(r)}}\sum_{k=1}^{K}\xi_t^{(r)}(k)}{\sum_{r=1}^{R}\sum_{t=1}^{T^{(r)}}\sum_{k=1}^{K}\xi_t^{(r)}(k)\cdot\left(\beta^{(r)}[d]-\mu_d\right)^2} = \frac{\sum_{r=1}^{R}T^{(r)}}{\sum_{r=1}^{R}T^{(r)}\cdot\left(\beta^{(r)}[d]-\mu_d\right)^2} \tag{10}$$

where $\{\mu_d, \tau_d \,|\, 1 \le d \le D\}$ are tied for all HMM states related to the current leaf node.

IV. Robust Decoding Based on Bayesian Predictive Classification

According to [10], the BPC decision rule makes a speech recognizer minimize the overall recognition error when the expectation is taken with respect to the uncertainty described by the prior PDF. Assume that the functional form of the parameter transformation is exactly known and all available information about the transformation parameters is completely contained in the prior PDF $\rho(\eta)$, given unknown observation X, then such an optimal recognition result $\hat{W}$ can be expressed as

$$\begin{aligned}\hat{W} &= \arg\max_{W}\Pr(W)\cdot\int_{\eta}\Pr(X\,|\,T_\eta(\lambda_c),W)\cdot\rho(\eta)\,d\eta \\ &= \arg\max_{W}\Pr(W)\cdot\int_{\eta}\sum_{s,l}\Pr(X,s,l,\,|\,T_\eta(\lambda_c),W) \\ &\quad\cdot\rho(\eta)\,d\eta \\ &\approx \arg\max_{W}\Pr(W)\cdot\max_{s,l}\int_{\eta}\Pr(X,s,l,\,|\,T_\eta(\lambda_c),W) \\ &\quad\cdot\rho(\eta)\,d\eta\end{aligned} \tag{11}$$

where s, l denote the HMM state sequence and the Gaussian component label sequence, respectively. In this equation, the Viterbi approximation [10] has been adopted to make the integral tractable. Based on the work in [10], where we have prior PDFs of all HMM parameters and the integral is taken with respect to the HMM parameters, this paper introduces new transformation-based structure constraint into the BPC method. Therefore, (11) can be thought as a kind of "constraint-based" BPC.

In this improved BPC, the optimal HMM parameters for each testing utterance are assumed to follow some constraints, which are established by applying transformations into a known "generic" HMM λ_c. The transformations are known exactly except for a small set of parameters η treated as random variables. It is further assumed that our prior knowledge about the transformation parameters η is contained in a prior PDF $\rho(\eta)$. Under these assumptions, the optimal decision rule will be the "constraint-based" BPC shown in (11). When the number of the transformation parameters is much fewer than that of HMM ones, the "constraint-based" BPC makes it easier to determine the prior PDF. An additional contribution of this work is that we significantly simplified the estimation of the prior PDF for the transformation parameters by incorporating the SAT in our training stage, as shown in step (4e) of Section III.

After we adopt the linear transformation as shown in (2) for $T_\eta(\cdot)$ where each transformation is associated with a HMM state, we now present a frame-synchronous search algorithm to implement the above "constraint-based" BPC rule. The search algorithm has been modified from the general VBPC algorithm presented in [10]. According to (11), the value of the integral depends on the path in the HMM. This makes it difficult to derive a recursive algorithm to compute an accurate value of the integral. The solution to this difficulty we have adopted is to incorporate the calculation of the integral in the Viterbi search. For each time frame, we compute the integration over the transformation parameters for all active hypothesized partial paths. Then, for each node in the search network, we merge all incoming partial paths by selecting the one with the largest integral value. The selected path is propagated and the integral is recomputed according to the extended partial path. The search procedure is repeated until the end of the utterance. In this way, we are able to achieve a Viterbi approximation of the integral.

Given a test utterance $X = (x_1, x_2, \ldots, x_T)$, the generic model λ_c, and the prior PDF $\rho(\vec{\beta})$ shown in (3) (with the hyperparameters estimated from (9) and (10)), the recursive search procedure for accomplishing the computation in (11) is described as follows.

1) Initialization

$$\delta_1(i) = \pi_i\cdot\tilde{b}_i(x_1) \quad 1 \le i \le N \tag{12}$$

$$\psi_1(i) = 0 \quad 1 \le i \le N \tag{13}$$

where (for $t = 1$)

$$\begin{aligned}\tilde{b}_i(x_t) &= \arg\max_{1\le k\le K}\int w_{ik}\cdot\mathcal{N}(x_t\,|\,m_{ik}+\vec{\beta}_i, r_{ik}) \\ &\quad\cdot\rho(\vec{\beta}_i)\,d\vec{\beta}_i \\ &= \arg\max_{1\le k\le K} w_{ik}\cdot\prod_{d=1}^{D}\sqrt{\frac{\tau_d^{(i)}r_{ikd}}{2\pi\left(\tau_d^{(i)}+r_{ikd}\right)}} \\ &\quad\times\exp\left[-\frac{\tau_d^{(i)}r_{ikd}}{2\left(\tau_d^{(i)}+r_{ikd}\right)}\left(x_{td}-m_{ikd}-\mu_d^{(i)}\right)^2\right]\end{aligned}$$

where $\vec{\beta}_i$ denotes the transformation parameters related to state i. Here, $\delta_t(i)$ denotes the partial predictive value based on the optimal partial path arriving at state i at the time instant t. The corresponding best partial path is represented by a chain of points starting from $\psi_t(i)$.

2) Recursion: for $2 \le t \le T$, $1 \le j \le N$, do

1) path-merging in state j

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}] \tag{14}$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) \cdot a_{ij}]. \tag{15}$$

2) Update the partial predictive value:
If (it is the first time to involve state j in the computation of $\delta_t(j)$)[2], **then**

$$\delta_t(j) = \delta_t(j) \times \tilde{b}_j(x_t) \tag{16}$$

else

$$\delta_t(j) = \delta_t(j) \times \frac{\tilde{b}_j\left(x_{j_1}, x_{j_2}, \ldots, x_{j_{L_j}}\right)}{\tilde{b}_j\left(x_{j_1}, x_{j_2}, \ldots, x_{j_{(L_j-1)}}\right)} \tag{17}$$

where L_j is the accumulated number of feature vectors belonging to state j based on the optimal partial path up to the time instant t; x_{j_i} denotes the ith vector in the state j; and $\tilde{b}_j(x_{j_1}, x_{j_2}, \ldots, x_{j_{L_j}})$ denotes the contribution of data $\{x_{j_1}, x_{j_2}, \ldots, x_{j_{L_j}}\}$, residing in state j, to the partial predictive value $\delta_t(j)$

$$\tilde{b}_j(x_{j_1}, x_{j_2}, \ldots, x_{j_n}) = \int p\left(x_{j_1}, x_{j_2}, \cdots, x_{j_n} \,|\, m_{jk} + \vec{\beta}_j, r_{jk}\right) \cdot \rho(\vec{\beta}_j)\, d\vec{\beta}_j. \tag{18}$$

3) Termination

$$\tilde{p}(X \,|\, W) \approx \max_i \ \delta_T(i) \tag{19}$$

$$s_T^* = \arg\max_i \delta_T(i). \tag{20}$$

4) Path (state sequence) backtracking

$$s_t^* = \psi_{t+1}(s_{t+1}^*) \quad t = T-1, T-2, \ldots, 1. \tag{21}$$

In (18), $\tilde{b}_j(x_{j_1}, x_{j_2}, \ldots, x_{j_n})$ can be approximated based on the "closest" mixture component label sequence corresponding to the data $\{x_{j_1}, x_{j_2}, \ldots, x_{j_n}\}$

$$\begin{aligned}\tilde{b}_j(x_{j_1}, x_{j_2}, \ldots, x_{j_n}) &\approx \prod_{k=1}^{K} w_{jk}^{L'_k} \cdot \tilde{f}_{jk}\left(x_{l_1^k}, \ldots, x_{l_{L'_k}^k}\right) \\ &= \prod_{k=1}^{K} w_{jk}^{L'_k} \cdot \prod_{d=1}^{D} \tilde{f}_{jkd}\left(x_{l_1^k d}, \ldots, x_{l_{L'_k}^k d}\right)\end{aligned} \tag{22}$$

[2]Including all states tied to state j.

where $\{x_{j_1}, x_{j_2}, \ldots, x_{j_n}\}$ denote feature vectors belonging to state j in X, among which $l_1^k \ldots l_{L'_k}^k$ denote labels of the vectors "closest" to the mixture component k of state j. Then, we have

$$\begin{aligned}&\tilde{f}_{jkd}(x_{1d}, x_{2d}, \ldots, x_{vd}) \\ &= \sqrt{\left(\frac{r_{jkd}}{2\pi}\right)^v \frac{\tau_d^{(j)}}{r_{jkd} + \tau_d^{(j)}}} \exp\left[-\frac{v r_{jkd}}{2}\left(\overline{x_v^2} - \overline{x}_v^2\right)\right] \\ &\quad \times \exp\left[-\frac{v r_{jkd} \tau_d^{(j)}}{2\left(v r_{jkd} + \tau_d^{(j)}\right)}\left(\overline{x}_v - \mu_d^{(j)}\right)^2\right]\end{aligned}$$

where

$$\overline{x}_v = \frac{1}{v}\sum_{i=1}^{v}(x_{id} - m_{jkd}) \tag{23}$$

and

$$\overline{x_v^2} = \frac{1}{v}\sum_{i=1}^{v}(x_{id} - m_{jkd})^2. \tag{24}$$

In the above VBPC, we search for a single best path to compute the integral-based Bayesian prediction instead of calculating the integral over all possible paths, as shown in (11) and (22). As in [10], we have found that the VBPC generally leads to a rather good approximation because the contribution from the best path almost always dominates the entire Bayesian prediction.

V. Experiments: Switchboard Task

In order to examine the viability of the proposed robust training strategy, we apply it to several types of extraneous acoustic variations in spontaneous speech recognition. In this paper, we choose a fast evaluation set of the switchboard corpus used in Workshop 1996 (WS96) at Johns Hopkins University, Baltimore, MD, which is approximately 10 h in duration. It is called "10-h" set hereafter. In the following recognition experiments, we first build a baseline system from the "10-h" training data according to the conventional training method. Then, starting from the baseline system, the new training method is used to deal separately with three different types of the extraneous variations in the switchboard task, i.e., the pronunciation variation, speaker difference and handset mismatch. The experimental setup and comparative results will be reported in this section, together with some discussions on the experimental results and on the computation complexity of the algorithm used.

A. The 10-h Baseline System

We use HTK v2.2 to implement our baseline system. In the baseline system, we use the 39-dimension feature vector which is composed of 12 MFCCs with log-energy, and delta and acceleration coefficients. The cepstral mean normalization is performed in the utterance level for both training and testing data. The acoustic models are three-state five-mixture-per-state word-internal triphone HMMs. The standard phonetic decision-tree method is used for state-tying. After the tying, the total number of all distinct tied-states is reduced to approximately

TABLE I
PERFORMANCE (IN %) COMPARISON WITH THE 10-h BASELINE SYSTEM WHEN THE ROBUST METHOD IS USED TO DEAL WITH PRONUNCIATION VARIATIONS

	Sub	**Del**	**Ins**	**WER**
10-hr baseline	43.94	17.92	3.49	65.39
RobustPron-I	41.94	18.58	4.31	64.84
RobustPron-II	42.61	18.17	3.90	64.68

2000. The dictionary consists of all words (about 6474 words) occurring in the "10-h" training set. Multiple pronunciations are used for some words. The language model is the back-off bigram model trained only on the transcriptions of all utterances in the "10-h" set. The test set consists of 200 utterances (a total of 1948 words) randomly selected from the evaluation test set in WS96, which is disjointed from the "10-h" set.

The recognition performance of the baseline system with these 200 test utterances is shown in the first line of the Table I, i.e., with 65.39% word error rate (WER). The performance is close to the best baseline results which were reported under identical conditions in WS96.

B. Dealing With Pronunciation Variations

According to [2] and [17], in the switchboard task, pronunciation variations in conversational speech is one major extraneous variation source hampering speech recognition performance. How to cope with pronunciation variations in conversational speech recognition has been studied by many researchers, e.g., [2]. It is straightforward to incorporate multiple pronunciations in the search network for some words. However, this strategy also increases the perplexity of the search network and makes it more confusable. In this section, we attempt to deal with the pronunciation variations by using our robust training strategy. In principle, the speech data which come from the same word may be treated as from different "conditions" if the word is pronounced differently. That is, the set of all "conditions" can be defined by all distinct pronunciations of all words in the vocabulary. In this way, the above robust training approach can be directly used to normalize the acoustic variations caused by pronunciation differences.

One most important implementation issues here is how to define the condition and partition the data into different conditions. It is crucial to have a good tradeoff between the number of conditions and the amount of data used for each condition. In order to obtain reliable estimation for the transformation parameters for each condition, it is important to ensure that there is enough data for each condition. In this work, we use the baseline recognition system and a phoneme recognizer to automatically determine pronunciation of every word in the training data. We have explored the following two methods in defining a *condition* for pronunciation variations.

1) Each utterance in the training set is force-aligned at the word level based on its transcription by using the baseline system to obtain the segmentation information for every single word. Then, phone recognition based on free-phone looping is performed on acoustic phone models in the baseline system for each word according to the above alignment boundary. The phoneme recognition results are viewed as the pronunciation of this word. However, this method usually causes too many pronunciations for each word. Thus, we use a very simple distance measure between two pronunciations, e.g., number of different phoneme, to cluster all different pronunciations of each word into four classes or fewer. In training stage (4a), all data from the same word and the same pronunciation class are treated as from the same condition. This method is denoted as *RobustPhon-I* hereafter.
2) Phoneme recognition is directly performed for each utterance in the training set to obtain the phoneme sequence for the sentence by using the baseline system, where phoneme HMMs are used without any language model. In training stage (3) described in Section III, each feature vector is labeled with the recognized phoneme where the vector belongs. When doing decision-tree state-tying in training stage (4a), all data in this state which corresponds to the same recognized phoneme is treated as from the same condition. This method is denoted as *RobustPhon-II* hereafter.

The *RobustPhon-I* and *RobustPhon-II* methods are implemented under the same experimental conditions as that of the baseline system. From the comparative experimental results in Table I, where **Sub**, **Del**, and **Ins** denote the substitute, deletion, and insertion error rate, respectively, we observe that the robust training method gives close to 1% reduction in word error rate (WER) over the baseline system. We also note that, for the 10-h data set, the *RobustPhon-II* achieves somewhat better results because *RobustPhon-I* usually causes too many conditions and in turn too few training data for the some conditions.

In both *RobustPhon-I* and *RobustPhon-II*, we can identify several factors which influence the final performance. The first factor is the poor phoneme recognition results when transcribing the switchboard data by using the baseline system. The high error rate in phoneme recognition causes the conditions related to pronunciation variations not to be well defined. The second factor is that the functional form (2) for transformation may not be powerful enough to normalize the acoustic variations caused by the pronunciation changes. Furthermore, the acoustic variation is only one aspect of pronunciation variations. In particular, phonetic reduction has been found to be one major cause of variability for spontaneous speech, which requires new dynamic modeling methodologies beyond the conventional HMMs that we have used in this work [4].

C. Dealing With Inter-Speaker Differences

The inter-speaker difference is another major source of extraneous variations for any speaker-independent speech recognition system. In this section, we report experimental results using our robust training strategy to compensate/normalize the inter-speaker difference in the switchboard data. Here, each "condition" in training stage (1) is related to each speaker in the training set. Then, in training stage (3), we label every feature vector in the entire training set with the speaker who utters the current sentence. In stage (4a), for every tied state, all speech data which come from the same speaker is considered under

TABLE II
PERFORMANCE (IN %) COMPARISON WITH THE 10-h BASELINE SYSTEM WHEN THE ROBUST METHOD IS USED TO DEAL WITH SPEAKER DIFFERENCE

	Sub	Del	Ins	WER
10-hr baseline	43.94	17.92	3.49	65.39
RobustSpk	43.28	17.57	3.33	64.18

TABLE III
PERFORMANCE (IN %) COMPARISON WITH THE 10-h BASELINE SYSTEM WHEN THE ROBUST METHOD IS USED TO DEAL WITH HANDSET MISMATCH

	Sub	Del	Ins	WER
10-hr baseline	43.94	17.92	3.49	65.39
RobustHandset	42.94	18.20	3.44	64.58

the same condition. This method is denoted as *RobustSpk* hereafter. Recognition performance comparison of *RobustSpk* with the baseline system is shown in the Table II. From the results, we note that when the robust training strategy is used to normalize the inter-speaker difference, 1.2% WER reduction over the baseline system has been achieved. We also see that the performance improvement here is somewhat larger than both *RobustPhon-I* and *RobustPhon-II*. One possible reason is that the condition here is well-defined because the condition is decided solely by the speaker id and is independent of the performance of the baseline system.

D. Dealing With Handset Mismatches

The switchboard data consist of recorded telephone conversations among a set of registered participants. A participant would initiate a conversation by calling an automaton that would find another participant to receive the call. The automaton would note the telephone numbers used by both participants. We generally assume that when the phone numbers are the same, the handsets are also the same, though there may be exceptions. In this section, based on the information of telephone numbers, the robust training strategy presented in Sections III and IV is similarly used to normalize the acoustic variations caused by handset mismatch. Here, each condition is related to one telephone number recorded in the training set. In training stage (4a), all training data from the same telephone number are considered to be under the same condition. This method is denoted as *RobustHandset* hereafter. The comparative results in Table III show that the robust training method *RobustHandset* achieves nearly 1% WER reduction over the baseline system.

E. Discussion

Although we have observed some moderate WER reduction for the switchboard task from all above promising experimental results, the performance improvement is smaller than what we had expected. One possible reason is that the switchboard task is an extremely difficult one, and the data contains many other types of variabilities which have not been addressed in this work.

One important issue here is the computational complexity of the new robust training approach introduced in this paper. Compared with the conventional training method and SAT, the robust training algorithm here does not significantly increase the computational complexity. However, as discussed in [10], the decoding algorithm based on BPC demands much more computation or memory overload than the normal Viterbi search algorithm. As shown in [10] and [12], this usually does not cause any problem for small-vocabulary tasks. For the switchboard task, where the search network is constructed from several thousand words, a fast implementation version of the VBPC search algorithm usually requires a memory greater than 1000 megabytes. Although the fast version of the VBPC search has a similar running speed as the normal Viterbi search, memory requirement is not affordable in most current machines. Thus, it is very important to have a good programming design to achieve a good tradeoff between the speed and memory. Even so, in most cases, in order to have an acceptable speed of response, heavy pruning is necessary in the search algorithm.

From the experimental results reported in this section, we have observed over 1% WER reduction (absolute) separately for each type of extraneous variations. It will be interesting to see whether we can have additive improvements when the method is used to jointly deal with all three types of variations. However, we will face a serious problem of "sparse data" when jointly normalizing three types of variations. For example, we usually have the total number of "conditions" from several tens to several hundreds in Sections V-B, C, and D. If we jointly deal with three types of variations, the total number of "conditions" will increase up to around one million. The training data will not be enough for most "conditions" unless we have a good method to tie some "conditions" together. This issue is a subject of future research.

VI. CONCLUSION

In this paper, we have proposed a robust training strategy to deal with extraneous acoustic variations in building robust speech recognition systems. In this strategy, we first define some "conditions" for the training data according to the extraneous variations to be compensated. Then, the data under different conditions are normalized prior to pooling them together to estimate the "compact" model. The corresponding decoding algorithm based on the BPC is also presented in the paper. The new approach can be used to deal with any type of extraneous variations in the speech recognition problem. This paper provides some examples of using this approach to deal with three types of extraneous variations: the pronunciation variation, inter-speaker difference, and handset mismatches, in the switchboard task for spontaneous speech recognition. The experimental results have provided evidence that the proposed robust training strategy is effective to deal with some extraneous acoustic variations in speech recognition. For all three types of extraneous variations we have examined, moderate performance improvements over a well-trained baseline system have been achieved. From our experiments, we also note that the performance gain of the new approach depends on several factors including 1) whether the functional form of the transformation we have chosen is powerful enough for the extraneous variations and 2) whether we have properly defined the "condition" so that we have an adequate amount of training data for each condition.

In this paper, we have only investigated a very simple functional form of the transformation: a piecewise shift transformation on the HMM mean vectors. We plan to extend this work to other more powerful transformation forms, such as the affine transformation involving all HMM parameters. Also the parameters of the prior PDF has been estimated from the training data based on the ML criterion. The method of moment in [7] is an alternative, over the ML method, to estimating the priors. This may be superior in performance. Moreover, it will be interesting and informative to experimentally compare the proposed method with other well-known techniques in terms of compensating acoustic variations, such as cluster adaptive training (CAT) in [6], bias removal in [14] and [18] and stochastic matching in [13] and [19]. Finally, our current robust training strategy can be further developed for new dynamic models of spontaneous speech intended to incorporate phonetic reduction (target undershoot) as well as pronunciation variations. In particular, phonetic reduction has been found to be one major cause of variability for spontaneous speech which requires dynamic modeling methodologies beyond the conventional HMMs [4].

APPENDIX
DERIVATION OF (4)–(8) IN THE ROBUST TRAINING ALGORITHM

In this Appendix, we provide the derivation of (4)–(8) in robust training algorithm presented in Section III.

Assume that we adopt the functional form (2) for the transformation and we have the generic model $\lambda_c = \{w_k, m_k, r_k \,|\, 1 \le k \le K\}$ for one tied state in a leaf node of the decision tree. Following , [1] and [19] given the data $X = \{X^{(r)} \,|\, r = 1, 2, \ldots, R\}$, where $X^{(r)} = \{x_t^{(r)} \,|\, 1 \le t \le T^{(r)}\}$ denotes all data labeled with the condition r and $x_t^{(r)}$ denotes a single feature vector belonging to condition r, the EM algorithm is used to estimate the generic model and the transformation parameters sequentially. Here the mixture component label k is the *missing data* while using the EM algorithm.

Firstly, given the generic model λ_c, we estimate the transformation parameter $\vec{\beta}^{(r)}$ for each condition r based on $X^{(r)}$.

E-Step:

$$\begin{aligned} &Q\left(\vec{\beta}^{(r)} \,|\, \vec{\beta}_0^{(r)}\right) \\ &\quad = E_k\left[\ln f\left(x_t^{(r)}, l_t \,|\, \vec{\beta}^{(r)}\right) \,|\, X^{(r)}, \lambda_c, \vec{\beta}_0^{(r)}\right] \\ &\quad = \sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \sum_{d=1}^{D} = \left[-\frac{r_{kd}}{2}\left(x_t^{(r)} - m_{kd} - \vec{\beta}^{(r)}[d]\right)^2\right] \\ &\qquad \cdot \xi_t^{(r)}(k) + \text{const} \end{aligned} \tag{25}$$

where $\xi_t^{(r)}(k)$ denotes the probability of $x_t^{(r)}$ residing in mixture component $l_t = k$, i.e.,

$$\begin{aligned} \xi_t^{(r)}(k) &= \Pr\left(l_t = k \,\middle|\, x_t^{(r)},\ \vec{\beta}_0^{(r)}\right) \\ &= \frac{w_k \cdot \mathcal{N}\left(x_t^{(r)} \,\middle|\, m_k + \vec{\beta}_0^{(r)}, r_k\right)}{\sum_{k=1}^{K} w_k \cdot \mathcal{N}\left(x_t^{(r)} \,\middle|\, m_k + \vec{\beta}_0^{(r)}, r_k\right)}. \end{aligned} \tag{26}$$

M-Step:

$$\begin{aligned} &\frac{\partial Q\left(\vec{\beta}^{(r)} \,\middle|\, \vec{\beta}_0^{(r)}\right)}{\partial \vec{\beta}^{(r)}[d]} \\ &= \sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \left[-r_{kd} \cdot \left(x_t^{(r)} m_{kd} - \vec{\beta}^{(r)}[d]\right) \cdot \xi_t^{(r)}(k)\right] = 0. \end{aligned} \tag{27}$$

Thus,

$$\beta^{(r)}[d] = \frac{\sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \xi_t^{(r)}(k) \cdot r_{kd} \cdot \left(x_{td}^{(r)} - m_{kd}\right)}{\sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \xi_t^{(r)}(k) \cdot r_{kd}}. \tag{28}$$

Secondly, given the generic model λ_c^0 and transformation $\{\vec{\beta}^{(r)} \,|\, 1 \le r \le R\}$, we re-estimate the generic model $\lambda_c = \{w_k, m_k, r_k \,|\, 1 \le k \le K\}$ based on the data X.

E-Step:

$$\begin{aligned} &Q\left(\lambda_c \,|\, \lambda_c^0\right) \\ &\quad = E_k\left[\ln f(X, l_t \,|\, \lambda_c) \,\middle|\, \lambda_c^0,\ X^{(1)}, \ldots, X^{(R)}, \vec{\beta}^{(1)}, \ldots, \vec{\beta}^{(R)}\right] \\ &\quad = \sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \left\{\sum_{d=1}^{D} \left[-\frac{r_{kd}}{2} \cdot \left(x_t^{(r)} - m_{kd} - \vec{\beta}^{(r)}[d]\right)^2 \right.\right. \\ &\qquad \left.\left. + \frac{1}{2} \ln r_{kd}\right] + \ln w_k\right\} \cdot \xi_t^{(r)}(k) + \text{const}. \end{aligned} \tag{29}$$

M-Step:

$$\begin{aligned} \frac{\partial Q\left(\lambda_c \,|\, \lambda_c^0\right)}{\partial m_{kd}} &= \sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \left[-r_{kd} \cdot \xi_t^{(r)}(k) \right. \\ &\qquad \left. \cdot \left(x_t^{(r)} - m_{kd} - \vec{\beta}^{(r)}[d]\right)\right] = 0. \end{aligned} \tag{30}$$

Thus,

$$m_{kd} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k) \cdot \left(x_{td}^{(r)} - \beta^{(r)}[d]\right)}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k)} \tag{31}$$

$$\begin{aligned} \frac{\partial Q\left(\lambda_c \,\middle|\, \lambda_c^0\right)}{\partial r_{kd}} &= \sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \left[-\frac{1}{2}\left(x_t^{(r)} - m_{kd} - \vec{\beta}^{(r)}[d]\right)^2 \right. \\ &\qquad \left. + \frac{1}{2r_{kd}}\right] \cdot \xi_t^{(r)}(k) = 0. \end{aligned} \tag{32}$$

Thus,

$$r_{kd} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k)}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k) \cdot \left(x_{td}^{(r)} - m_{kd} - \beta^{(r)}[d]\right)^2} \tag{33}$$

$$\frac{\partial Q\left(\lambda_c \,\middle|\, \lambda_c^0\right)}{\partial w_k} - \Lambda = \sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \frac{1}{w_k} \cdot \xi_t^{(r)}(k) - \Lambda = 0 \tag{34}$$

where Lagrange multiplier is used. Based on the constraint $\sum_k w_k = 1$, we have

$$w_k = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \xi_t^{(r)}(k)}{\sum_{r=1}^{R} \sum_{t=1}^{T^{(r)}} \sum_{k=1}^{K} \xi_t^{(r)}(k)}. \tag{35}$$

ACKNOWLEDGMENT

The authors would like to thank Dr. Q. Huo of the Department of Computer and Information Systems, the University of Hong Kong, for his useful suggestions and comments on this work and many discussions which greatly enhance the paper. They also thank I. Stokes-Rees for the help in building the baseline system, and thank Dr. C.-H. Lee of Bell labs for his helpful comments on the work.

REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwarts, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 1137–1140.

[2] B. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modeling for conversational speech recognition: A status report from WS97," *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, pp. 26–33, Nov. 1997.

[3] M. H. DeGroot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[4] L. Deng and J. Ma, "A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 1499–1502.

[5] M. J. F. Gales, "Cluster adaptive training for speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 1783–1786.

[6] ——, "Cluster adaptive training of hidden markov models," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 417–428, July 2000.

[7] Q. Huo, C. Chan, and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 334–345, Sept. 1995.

[8] Q. Huo and B. Ma, "Irrelevant variability normalization in learning structure from data: A case study on decision-tree based HMM state tying," in *Proc. ICASSP'99*, May 1999, pp. 577–580.

[9] Q. Huo and C.-H. Lee, "Robust speech recognition based on adaptive classification and decision strategies," *Speech Commun.*, to be published.

[10] H. Jiang, K. Hirose, and Q. Huo, "Robust speech recognition based on Bayesian prediction approach," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 426–440, July 1999.

[11] ——, "Improving Viterbi Bayesian predictive classification via sequential Bayesian leaning in robust speech recognition," *Speech Commun.*, vol. 28, no. 4, pp. 313–326, Aug. 1999.

[12] ——, "A minimax search algorithm for CDHMM based robust continuous speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, Nov. 1998, pp. 389–392.

[13] H. Jiang, F. Soong, and C.-H. Lee, "Hierarchical stochastic feature matching for robust speech recognition," in *Proc. ICASSP'01*, Salt Lake City, UT, May 2001.

[14] C. Lawrence and M. Rahim, "Integrated bias removal techniques for robust speech recognition," *Comput. Speech Lang.*, vol. 13, pp. 283–298, 1999.

[15] C.-H. Lee, F.-K. Soong, and K.-K. Paliwal, Eds., *Automatic Speech and Speaker Recognition: Advanced Topics*. Norwell, MA: Kluwer, 1996.

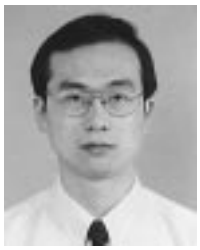
[16] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[17] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *Proc. Int. Conf. Spoken Language Processing*, Dec. 1999, pp. 1847–1850.

[18] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, p. 19, Jan. 1996.

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

[20] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, 1994, pp. 307–312.



Hui Jiang (M'00) received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China (USTC) and the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in September 1998, all in electrical engineering.

From 1992 to 1994, he worked on large-vocabulary Chinese speech recognition at USTC. From January 1995 to September 1998, he was with the Department of Information and Communication Engineering, University of Tokyo, where he mainly worked on robust speech recognition. From October 1998 to April 1999, he was a Researcher with the University of Tokyo. From April 1999 to June 2000, he was a Postdoctoral Fellow with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. Since June 2000, he has been with Dialogue Systems Research, Multimedia Communication Research Lab, Bell Labs, Lucent Technologies Inc., Murray Hill, NJ. His current research interests include all issues related to speech recognition and understanding, especially robust speech recognition, utterance verification, adaptive modeling of speech, spoken language systems, and speaker recognition/verification.

Li Deng (S'83–M'86–SM'91) received the B.S. degree in biophysics from the University of Science and Technology of China in 1982, the M.S. degree from the University of Wisconsin-Madison in electrical engineering in 1984, and the Ph.D. degree in electrical engineering from the University of Wisconsin, Madison, in 1986.

He worked on large-vocabulary automatic speech recognition for telecommunications from 1986 to 1989. In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as Assistant Professor; he became Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and, from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In December 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher. His research interests include acoustic-phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, statistical methods and machine learning, nonlinear signal processing and system theory, spoken language systems, and human–computer interaction.