TR-IT-0246

# Towards Robust Automatic Speech Recognition: A Bayesian Perspective

Qiang Huo

1997.12

**Abstract**

In this report, we start with a revisit to the statistical formulation of the automatic speech recognition (ASR) problem, identify the factors which might influence the performance of the conventional *plug-in MAP* decision rule for ASR. We summarize our recent research efforts on a class of robust speech recognition problem in which mismatches between training and testing conditions exist but an accurate knowledge of the mismatch mechanism is unknown. The only available information is the test data along with a set of pre-trained speech models and the decision parameters. We focus on two types of Bayesian techniques, namely on-line Bayesian adaptation of hidden Markov model parameters and Bayesian predictive classification approach. We conclude the report with a brief mention of our ongoing research efforts towards a robust and intelligent spoken dialogue system.

# Contents


1 **Introduction** 1

2 **Robust Speech Recognition Problem And Approaches** 2

3 **Bayesian Approaches to Robust Speech Recognition** 4
- 3.1 Dynamic System Design Strategy: On-line Bayesian Adaptation . . . . . 5
  - 3.1.1 Background . . . . . 5
  - 3.1.2 Bayesian Approaches to On-line Adaptation . . . . . 5
  - 3.1.3 Relation to Other Adaptation Approaches . . . . . 7
- 3.2 Robust Decision Strategy: Bayesian Predictive Classification . . . . . 9
  - 3.2.1 General Formulation . . . . . 9
  - 3.2.2 BPC Formulation for Robust Speech Recognition . . . . . 9
  - 3.2.3 Approximate BPC Approach . . . . . 11
  - 3.2.4 Relation to Other Robust Decision Approaches . . . . . 12
- 3.3 Sensitivity of Priors . . . . . 13

4 **Putting It All Together: A Robust HMM-based ASR System** 14

5 **Discussion and Conclusion** 15

6 **Acknowledgement** 16

**References** 17

# 1 Introduction

In the last two decades, many advances have been achieved in the area of automatic speech recognition (ASR) (see, e.g., [48] for a sample of the state-of-the-art). This is largely attributed to the use of a powerful statistical pattern matching paradigm and the application of dynamic programming search over a structural network representation of acoustic and linguistic knowledge sources. The reader is referred to the seminal contributions in e.g., [8, 9, 35, 5] and a recent overview in [63] for the capabilities and limitations of the pattern recognition approach. For this approach, let's view a *word* $W$ and the associated acoustic observation $\mathbf{X}$ (usually, a feature vector sequence) as a jointly distributed random pair $(W, \mathbf{X})$. Depending on the problem of interest, *word* here could be any linguistic unit, such as a phoneme, a syllable, a word, a phrase, a sentence, etc. Suppose the *true* joint distribution of $(W, \mathbf{X})$ could be modeled by a *true parametric family* of pdf (probability density function) $p(W, \mathbf{X}) = p_\Lambda(\mathbf{X}|W) \cdot P_\Gamma(W)$, where $p_\Lambda(\mathbf{X}|W)$ is known as the acoustic model with parameters $\Lambda$ and $P_\Gamma(W)$ as the language model with parameters $\Gamma$. This decomposition scheme is called *sampling paradigm* in the statistics community to contrast with another so-called *diagnostic* paradigm with $p(W, \mathbf{X}) = P(W|\mathbf{X}) \cdot p(\mathbf{X})$ [66]. Further suppose we have the full knowledge of the parameters $(\Lambda, \Gamma)$ of the above distributions. Then, an *optimal* decoder (speech recognizer) which achieves the *expected* minimum *word* recognition error rate is the following MAP (maximum *a posteriori*) decoder (see, e.g., [66, 15] for a more general discussion on statistical decision theory):

$$\hat{W} = \arg\max_W P(W|\mathbf{X}) = \arg\max_W p_\Lambda(\mathbf{X}|W) \cdot P_\Gamma(W) \tag{1}$$

where $\hat{W}$ is the recognition result. However, in practice, neither do we know the *true* parametric form of $p(W, \mathbf{X})$, nor its *true* parameters. Therefore, the above optimal speech recognizer will never be achievable, but we can only approximate it. A simple heuristic solution is first to assume some parametric form for $p(W, \mathbf{X})$ and then to estimate its parameters $(\Lambda, \Gamma)$ from some training data by using some parameter estimation techniques. Then, we *plug in* the estimate $(\tilde{\Lambda}, \tilde{\Gamma})$ into the optimal but unavailable rule in Equation (1) in place of the correct but unknown $(\Lambda, \Gamma)$ to obtain a *plug in MAP rule* (see a general discussion in e.g., [66]).

# 2 Robust Speech Recognition Problem And Approaches

Theoretically speaking, the performance of the above plug-in decision rule will depend on the following conditions:

- if the assumed parametric models are accurate and flexible enough to appropriately model the highly complex and variable speech signals,
- if the assumed models and the related parameter estimation methods are computationally efficient and robust enough to take care of the possible distortions between models and training samples which might be caused by wrong model assumptions, dependence and/or correlations of training samples, misclassification of training samples, outliers in training samples, etc.,
- if the training data are sufficient and representative enough to guarantee good parameter estimation and generalizability,
- if the distortions between trained models and actual testing data are small enough to avoid the breakdown of the whole approach.

Not all of the above issues have been seriously addressed in the past. Currently, the most widely used and the most successful modeling approach to ASR is to use a set of hidden Markov models (HMM's) as the acoustic models of subword or whole-word units, and to use the statistical $N$-gram model or its variants as lexical language model for words and/or word classes. The reader is referred to two good tutorials in [62] and [36] respectively for an introduction to the above two approaches and their applications. By using the above plug-in MAP approach, it has been repetitively shown by experiments in the past decade that given a large amount of representative training speech and text data, good statistical acoustic and language models can be constructed to achieve a high performance for many ASR tasks. Currently, the most popular training method for HMM parameters is still the *maximum likelihood* (ML) estimate [10, 46, 40]. It is noted in [57] that, if certain assumptions are met, one can argue intuitively that using the ML estimate of HMM's and the plug-in MAP decision rule can lead to a speech recognition system that is asymptotically optimal. Nevertheless, apart from many other issues, inaccuracy alone in modeling the speech signal by HMM may lead to ML models that do not maximize the recognition accuracy [57, 59, 42]. In the past decade, many alternatives to ML training which rely less on the model accuracy assumptions have been investigated. One method is the *minimum discrimination information* (MDI) training [17] which adjusts HMM parameters to minimize a measure (*discrimination information*, or *directed divergence*) between assumed HMM distribution and the best possible distribution derived from the training data under certain constraints embedded in the training data. Unfortunately, no experimental results have been reported to show how MDI works in a speech recognition task. Another type of methods is the so-called *discriminative training*. Some of them such as *maximum mutual information* (MMI) training [6], *conditional maximum likelihood estimate* (CMLE) [59], H-criteria [24] aim indirectly at reducing the error rate of the speech recognizer on training data. Other methods such as *corrective training* [7], *minimum empirical error rate* training [18, 53], *minimum classification error* (MCE) training [44, 42, 43] try to reduce the recognition error rate on training data in a more direct way. It is quite clear that if there are no big mismatches between training and testing conditions and the training data is rich and representative enough, discriminative training can help improving the recognition performance over that of ML training. Otherwise, one should be careful to use the discriminative training [61]. The efficacy of any discriminative training methods is highly dependent on the nature and the size of the training data as well as the task itself and sometimes this limits its generalizability. On the other hand, if appropriately used, the discriminative training can also maximize the separation between models of speech units so that the robustness of a recognizer is improved. The final effect will depend on the result of these two competing factors. For MCE training based on the *generalized probabilistic descent* (GPD) algorithm, it can also be viewed as an adaptive learning algorithm because of its stochastic

approximation nature. However, it converges in probability. Only after a large amount of training data is used, the algorithm starts to converge. This makes MCE/GPD alone not suitable for efficient adaptation purpose.

In many real applications, there always exists some form of mismatch between training and testing conditions. These mismatches may arise from inter- and intra- speaker variabilities, transducer, channel and other environmental variabilities, and many other phonetic and linguistic effects due to the problem of task mismatch. It is the susceptibility of current ASR systems to even moderate acoustic mismatches that prevents the widespread deployment of the ASR systems in a wide range of operating conditions. Robust speech recognition in this context thus refers to the problem of designing an automatic speech recognizer that works well for different tasks and speakers over unexpected and possibly adverse conditions. There are many ways to achieve robust ASR which might include:

- finding invariant or robust features,
- developing better modeling and learning techniques,
- applying signal/feature/model compensation/adaptation techniques, and
- using robust decision strategies.

Along these lines, there have been a great deal of efforts aiming at improving speech recognition and hence enhancing performance robustness in the abovementioned mismatches (see recent reviews in [41, 50, 20] and the references therein).

# 3 Bayesian Approaches to Robust Speech Recognition

In the past few years, we have been adopting a Bayesian paradigm to formulate and address a class of robust speech recognition problem in which

- mismatches between training and testing conditions exist, but
- an accurate knowledge of the mismatch mechanism is unknown,
- the only available information is the test data along with a set of pre-trained speech models and the decision parameters.

We've developed two sets of Bayesian techniques to cope with the acoustic mismatch problem for HMM based speech recognition. The first type of algorithms are targeting those applications involving a recognition session which might consist of a number of testing utterances. Unlike those ASR systems which rely on a *static* design strategy that all the knowledge sources needed in a system are acquired at the design phase and remain fixed during use, we adopt a *dynamic* system design strategy where the new knowledge is acquired sequentially, new information is constantly collected during development and use of the ASR system, and is incorporated into the system using an adaptive learning algorithm, namely *on-line Bayesian adaptive learning* of the HMM parameters [27, 28, 29]. For the second type of techniques, by modifying directly the above plug-in MAP decision rule, we've developed a new robust decision strategy called *Bayesian predictive classification* (BPC) approach in which part of the mismatch can be compensated and the decision performance can be improved [30, 31, 32, 33]. The robustness of the ASR system can be further enhanced by integrating on-line adaptation (OLA) of model parameters with a BPC-based decision rule [31, 32].

In the rest of the report, instead of presenting those comprehensive technical details and the related experimental results which mostly have been and are going to be published elsewhere, we intend to provide here the readers a snapshot of our recent works mentioned above. Special attentions have been paid to provide the background information of the problem, the motivation behind the development of the theory, and the basic principle of the algorithms. This report is thus aiming at triggering the readers' interest to read our other relevant publications and hopefully inspiring other innovations that would potentially lead to better solutions in the context of robust ASR.

## 3.1 Dynamic System Design Strategy: On-line Bayesian Adaptation

### 3.1.1 Background

The on-line adaptation scenario is like this: starting from a previously trained (e.g., speaker and/or task independent [49]) speech recognition system, for a new user (or a group of users) to use the system for a specific task, a small amount of adaptation data is collected from the user. These data are used to construct a speaker adaptive system for the speaker in the particular environment for that specific application. By doing so, the mismatch between the training and testing environments can generally be reduced. The most fascinating adaptation scheme with a practical value is the so called *on-line* (or *incremental, sequential*) adaptation. This scheme makes the recognition system capable of continuously adapting to the new adaptation data (possibly derived from actual test utterances) without the requirement of storing a large set of previously used training data.

Recently, Bayesian adaptive learning of HMM parameters has been proposed and adopted in a number of speech recognition applications. A theoretical framework of Bayesian learning was first proposed by Lee *et al.* [47] for estimating the mean and covariance matrix parameters of a continuous density HMM (CDHMM) with a multivariate Gaussian state observation density. It was then extended to handle all the parameters of a CDHMM with Gaussian mixture state observation densities (e.g., [22]) as well as the parameters of discrete HMMs (DHMMs) and semi-continuous HMMs (SCHMMs, also called tied-mixture HMMs) (e.g., [26]). It was shown that, for HMM-based speech recognition applications, the MAP framework provides an effective way for combining adaptation data and the prior knowledge, and then creating a set of adaptive HMMs to cope with the new acoustic conditions in the test data. The prior knowledge, which is embodied in a set of seed HMMs as well as in the assumed distributions of the model parameters being adapted, is made use of to mitigate the effect of adaptation data shortage to improve the system robustness. This approach works in a *batch* adaptation mode using a history of all the adaptation data. It can also be modified to work in a more attractive *incremental* adaptation mode. A related study was conducted by Matsuoka and Lee [54] in which they used the segmental MAP algorithm to perform *on-line adaptation.* Due to its missing mechanism of updating the hyperparameters of the prior and/or posterior distribution incrementally, all the previously seen adaptation data need to be stored.

The advantage of a sequential algorithm over a batch algorithm is not necessarily in the final result, but in computational efficiency, reduced storage requirements, and the fact that an outcome may be provided without having to wait for all the data to be processed. Moreover, the parameters of interest are sometimes subject to changes, e.g., they are time varying just like previously mentioned acoustic mismatch problem frequently encountered in real speech recognition applications. In such cases, different data segments often correspond to different parameter values. Processing of all the available data jointly is no longer desirable, even if we can afford the computational load of the batch algorithm. To alleviate such problems, a better *on-line* Bayesian adaptation approach should be able to update both the hyperparameters of the prior and/or posterior distributions and the HMM parameters themselves simultaneously upon the presentation of the latest adaptation data. A sequential algorithm can also be designed to adaptively track the varying parameters by further introducing some forgetting mechanism to adjust the contribution of previously observed sample utterances. Recursive Bayesian inference theory provides a good vehicle to formulate this problem.

### 3.1.2 Bayesian Approaches to On-line Adaptation

Let $\mathcal{X}_1^n = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$ be $n$ independent sets of observation samples which are used to estimate/adapt the HMM parameters $\Lambda$. Our initial knowledge about $\Lambda$ is assumed to be contained in a known joint *a priori* density $p(\Lambda)$, with $\Lambda \in \Omega$, where $\Omega$ denotes an admissible region of the HMM parameter space. In denoting the prior pdf $p(\Lambda)$, we do not explicitly show the parameters of the prior pdf (often referred to as the *hyperparameters*) which are assigned values by the investigator.

Such prior information may, for example, come from subject matter considerations and/or from previous experiences. Let's assume the samples $\mathcal{X}_i$'s are given successively one by one, we can obtain a recursive expression for the *a posteriori* pdf of $\Lambda$, given $\mathcal{X}_1^n$, as

$$p(\Lambda|\mathcal{X}_1^n) = \frac{p(\mathcal{X}_n|\Lambda) \cdot p(\Lambda|\mathcal{X}_1^{n-1})}{\int_\Omega p(\mathcal{X}_n|\Lambda) \cdot p(\Lambda|\mathcal{X}_1^{n-1}) d\Lambda} . \tag{2}$$

Starting the calculation of posterior pdf from $p(\Lambda)$, repeated use of the equation (2) produces the sequence of densities $p(\Lambda|\mathcal{X}_1^1)$, $p(\Lambda|\mathcal{X}_1^2)$, and so forth. This provides a basis of making formal recursive Bayesian inference of parameters $\Lambda$ and thus a good solution for on-line HMM adaptation. However, there are some serious computational difficulties to directly implement this learning procedure [28]. Consequently, some approximations are needed in practice. One such approach called quasi-Bayes (QB) learning was firstly developed in [26, 27] for adapting the mixture coefficients of SCHMM parameters and then extended to incremental adaptive learning of all of the CDHMM parameters in [28]. Based on the theory of recursive Bayesian inference, the QB algorithm is designed to incrementally update the hyperparameters on the approximate posterior distribution and the CDHMM parameters simultaneously. By further introducing some *forgetting mechanisms*, namely *exponential forgetting* and *hyperparameter refreshing*, to adjust the contribution of previously observed sample utterances, the algorithm is truly adaptive in nature and capable of performing an on-line adaptive learning using only the current sample utterance. On the other hand, the QB framework is also flexible enough to include the batch and/or block mode MAP/ML learning as special cases. This algorithm has also been implemented into ATR's speech recognition package [69] and was shown to work well in ATR's large vocabulary conversational speech recognition system [78]. More recently, based on the above general QB framework, a sequential learning method of mean vectors of CDHMM based on a finite mixture approximation of their prior/posterior densities has also been investigated [38, 39].

In a conventional HMM-based Bayesian adaptation framework, HMM parameters of different speech units are usually assumed independent. Therefore, each model can only be adapted if the corresponding speech unit has been observed in the current adaptation data. Consequently, only after all units have been observed enough times, all of the HMM parameters can thus be effectively adapted. To enhance the efficiency and the effectiveness of the Bayes adaptive training, it is desirable to introduce some constraints on HMM parameters based on all possible sources of knowledge. Therefore all the model parameters can be adjusted at the same time in a consistent and systematic way even though some units are not seen in adaptation data. A simple way to achieve the above objective is to introduce the parameter tying. Consequently, the formulation in [28] can be straightforwardly modified to accommodate the on-line adjustment of the tied parameters. Another way to achieve the above objective is to explicitly consider the correlation of HMM parameters corresponding to different speech units. However, it is too difficult to define a joint prior distribution for all sets of HMM parameters, if not impossible. A tractable case could be to assume all mean vectors are correlated and have a joint prior distribution [45]. By restricting ourselves to this special case, in [29], we have extended our QB learning framework to cope with the correlated CDHMM's with Gaussian mixture state observation densities in which all mean vectors are assumed to be correlated and have a joint Gaussian distribution. Considering the difficulties of parameter updating and initial hyperparameters' estimation arisen from the introduction of correlation between different models, we proposed a successive approximation algorithm based on pairwise correlations to update the mean vectors of CDHMM's as well as the corresponding hyperparameters. As an example, by applying the method to on-line speaker adaptation application, the algorithm is experimentally shown to be asymptotically convergent as well as being able to enhance the efficiency and the effectiveness of the Bayes learning by taking into account the correlation information between different model parameters. The technique can be used to cope with the time-varying nature of some acoustic and environmental variabilities, including mismatches caused by changing speakers, channels, transducers, environments and so on.

### 3.1.3 Relation to Other Adaptation Approaches

Now, we are ready to compare our approach to other related methods in the literature. In speech and pattern recognition area, to our knowledge, it was Lasry and Stern that first proposed a formulation of the MAP estimate (called extended MAP, or EMAP) in [45] for the mean vectors of a set of Gaussian pdf's in which those mean vectors are assumed to have a joint Gaussian prior distribution. They applied the EMAP method to the dynamic speaker adaptation in a feature-based isolated word recognition application [73]. To avoid the difficulty of the initial hyperparameters estimation, a classifier with a decision-tree structure is adopted. At each node of the decision tree, the utterance is classified into a small number of decision categories, based on a relatively small number of features that are relevant to the classification in question. Consequently, every time, they only make use of the correlation information among a small number of classes for adaptation and thus can afford the memory requirement and the computational complexity of the related algebraic operations. To avoid the repeated inversion of a big matrix in the standard EMAP implementation for dynamic speaker adaptation, later, in the context of SCHMM, Rozzi and Stern developed a least mean square (LMS) algorithm to implement the correlated means adaptation which is supposed to be more computationally efficient, but at the expense of a finite misadjustment [68]. On the other hand, the initial hyperparameters estimation problem still exists. More recently, Zavaliagkos *et al* applied EMAP into a large scale CDHMM-based speech recognition systems [79, 80]. With a similar motivation as in [45, 73], they adopted a hierarchical class tying technique to ease the abovementioned difficulties of the EMAP implementation. In [29], we integrate EMAP into our quasi-Bayes learning framework [26, 27, 28] and propose a successive approximation algorithm to ease the implementation. The algorithm does not involve any big matrix operation, thus becomes very computationally efficient. On the other hand, even if we can have an initial estimate of a non-singular correlation matrix, the successive approximation algorithm can not guarantee its nonsingularity after each iteration. However, because the implementation of the algorithm does not rely on the assumption of the nonsingularity of the correlation matrix, this in turn eases the problem of the initial hyperparameters' estimation.

We also wish to draw the reader's attention to the work of Shahshahani [71], who has a very similar motivation to our work in the sense of exploiting model correlations for efficient Bayesian adaptation where a Gibbs distribution is adopted to serve as the joint prior pdf of the mean vectors of the all CDHMM's. However, in that work, only conventional batch mode adaptation is formulated and it's very difficult to extend this method for a true on-line adaptive learning.

We can also set up the links between our approach and two other techniques, namely MAP/VFS (e.g., [74, 75, 76, 60, 25]) and regression based model prediction (RMP) methods (e.g., [19, 12, 2]). Our parameter updating equation in [29] is very similar to the so-called *interpolation* step in MAP/VFS method [74, 75] except that 1.) we use a different weighting coefficient, and 2.) every time, we only use the information from one mixture component to predict the mean vector of the mixture component without observations. But in our approach, by successively changing the role of the mixture components, we can achieve the similar effects as those of both *interpolation* and *smoothing* steps in MAP/VFS formulation. Furthermore, by updating the correlation coefficient, the algorithm can autonomously control the importance of the correlation information and thus make the estimations of the mean vectors of CDHMM asymptotically converge to their MAP or ML estimates without considering correlation. On the other hand, in MAP/VFS case, to avoid early saturation of the adaptation, some heuristic methods have to be employed [76]. We can also view our parameter updating equation as a simple linear regression function with one *explanatory variable* and the adaptive regression coefficients. Once again, by successive approximation, we can achieve the similar effect as that of RMP in [12, 2].

In the context of efficient adaptation, our method also shares the similarity with another type of transformation-based adaptation methods (e.g., [13, 52, 21]) in a more general sense of global mapping. The basic idea of both types of methods is to bind HMM parameters together (via correlation structure in our case and some shared transformations among different model parameters

in the latter case), and then to adjust them globally in a consistent and systematic way. For the transformation-based approaches, in order to achieve a better asymptotic convergence, one has to either dynamically increase the number of shared transformations according to the amount of available adaptation data (e.g., [52, 21]) or just combined with the Bayesian approach (e.g., [14]), both in a heuristic way.

From above discussions, we can see that our Bayesian learning procedure has a more consistent formulation as well as an intuitively pleasing behavior (an improved adaptation efficiency for short adaptation data and a good asymptotic property for increasing number of adaptation data). By activating the *forgetting mechanism*, the algorithm can also be used to cope with the continuously changing conditions. We expect that discovering an appropriate acoustic space configuration and a good definition of an appropriate correlation structure among states and/or phones could be helpful for enhancing the efficiency of the OLA of the correlated CDHMM's. It will be interesting to see how it works by combining our approach with other techniques such as tree-structured Gaussians to explore acoustic space structure (e.g., [72]), phone-dependence tree to explore phonetic dependency structure (e.g., [67]), and their combination (e.g., [34]). We believe this is an area that deserves a further research from both a theoretical and a practical point of view.

## 3.2 Robust Decision Strategy: Bayesian Predictive Classification

### 3.2.1 General Formulation

The conventional *plug-in* MAP decision rule for speech recognition is known to achieve an optimal Bayes decision if the assumed models and parameters of the rule were correct. However, in real world situations, we rarely have the full knowledge about the nature of the classification data to warrant optimal decisions. Some recent approaches have focused on modifying the *decision rule* to improve the decision performance.

In a Bayesian framework, we intend to consider the uncertainty of the HMM parameters $\Lambda$ by treating them as if they were random. Our prior knowledge about $\Lambda$ is assumed to be summarized in a known joint *a priori* density $p(\Lambda)$. Suppose a training set of the form $\mathcal{X} = \{\mathbf{X}^{(q,r)}\}$ is available, with $\mathbf{X}^{(q,r)}$ denoting the $r$th training observation sequence of length $T^{(q,r)}$ associated with the $q$-th speech unit, and each unit has $W_q$ such observation sequences. A posterior distribution can now be constructed as

$$p(\Lambda|\mathcal{X}) = \frac{p(\mathcal{X}|\Lambda) \cdot p(\Lambda)}{\int_\Omega p(\mathcal{X}|\Lambda) \cdot p(\Lambda) d\Lambda} \tag{3}$$

to update our knowledge about $\Lambda$. This posterior pdf $p(\Lambda|\mathcal{X})$ includes all of the information inherited from the prior knowledge and learned from the training data. Conventionally, we derive a *point estimate* $\tilde{\Lambda}$ from $p(\Lambda|\mathcal{X})$ (e.g., MAP estimate [47, 22, 26, 28]) and then use the plug-in MAP decision rule in Equation (1) for recognition. If we want to account for HMM model parameters' uncertainty in *recognition*, an *optimal Bayes solution*, namely *Bayesian predictive classification* (BPC) approach exists which chooses a speech recognizer to minimize the *overall recognition error* when the average is taken both with respect to the sampling variation in the expected testing data and the uncertainty described by the prior/posterior distribution (The reader is referred to [58, 66] for a brief proof of the optimality of the BPC rule). If we assume that the language model is known and only acoustic models are adjusted, such a BPC rule operates as follows:

$$\hat{W} = \underset{W}{\text{argmax}}\, \tilde{p}(W|\mathbf{X}) = \underset{W}{\text{argmax}}\, \tilde{p}(W, \mathbf{X}) = \underset{W}{\text{argmax}}\, \tilde{p}(\mathbf{X}|W) \cdot P_\Gamma(W) \tag{4}$$

where

$$\tilde{p}(\mathbf{X}|W) = \int_\Omega p(\mathbf{X}|\Lambda, W) p(\Lambda|\mathcal{X}, W) \mathrm{d}\Lambda \tag{5}$$

is called the predictive pdf [3, 23, 66] of the observation $\mathbf{X}$ given the word $W$. The computation of this predictive pdf is usually the most difficult part of the BPC procedure. The crucial difference between the plug-in and predictive classifiers is that the former acts as if the estimated model parameters were the true ones whereas the predictive methods average over the uncertainty in parameters.

Historically, the predictive approach receives little attention in many classical statistics textbooks despite the existence of many good works. This may be because it usually makes little difference from plug-in approaches within the problems and the tightly constrained parametric families many statisticians used or considered. Nonetheless, it will become important when we consider much larger families and formulate the problem appropriately. The books of Aitchison & Dunsmore (1975) [3] and Geisser (1993) [23] are devoted to the predictive approach. Both contain brief accounts of classification, in Aitchison & Dunsmore's Chapter 11 under the heading of "diagnosis" and in Geisser's Chapter 9 under the heading of "classification" respectively. Ripley (1996) [66] also contains a concise treatment of the topic.

### 3.2.2 BPC Formulation for Robust Speech Recognition

In speech recognition area, to our knowledge, it was Nadas who first adopted a BPC formulation and pointed out its potential in speech recognition applications [58]. He was using the posterior pdf $p(\Lambda|\mathcal{X})$ derived from the *training set* $\mathcal{X}$ directly to serve as the prior pdf in predictive decision making and gave a simple example in which reproducing density existed. No experimental results

were reported and the paper closed by briefly discussing the difficulty of applying the theory to HMM-based speech recognition.

Started with Nadas's formulation, Merhav and Ephraim [55] suggested a so-called *approximate Bayesian decision rule* (AB) for speech recognition which was based on the generalized likelihood ratios computed from the available training and testing data. Such an AB rule operates as follows:

$$\hat{W} = \arg\max_{W} \frac{\max_{\Lambda}[p(\mathbf{X}|\Lambda, W) \cdot p(\mathcal{X}|\Lambda, W)]}{\max_{\Lambda} p(\mathcal{X}|\Lambda, W)]} P_{\Gamma}(W) \ . \tag{6}$$

It is clear that if the training sequences $\mathcal{X}$ are considerably longer than the test sequence $\mathbf{X}$ which is the case in most speech recognition applications, the parameter set $\Lambda$ that maximizes the denominator of Equation (6) is very close to the parameter set that maximizes the numerator, hence the factor $p(\mathcal{X}|\Lambda, W)$ in both numerator and denominator is essentially canceled. This makes the AB decision rule of little difference from the plug-in MAP decision rule using ML estimate of $\Lambda$. The AB decision rule is also computationally expensive because the maximization of $[p(\mathbf{X}|\Lambda, W) \cdot p(\mathcal{X}|\Lambda, W)]$ over $\Lambda$ must be performed for every test sequence $\mathbf{X}$. Furthermore, all of the training data must be stored. All of these facts make the AB decision rule impractical for most of the speech recognition applications.

Similarly, if we directly apply the decision rule in Equation (4) as suggested by Nadas to speech recognition, it will also make little difference from the conventional plug-in MAP rule. This is because whatever the initial prior pdf, $p(\Lambda)$, is used, when a large amount of training data $\mathcal{X}$ are available, we will get a posterior pdf $p(\Lambda|\mathcal{X})$ with a sharp mode. This makes the predictive pdf in Equation (5) of little difference from $p(\mathbf{X}|\tilde{\Lambda}, W)$ with the ML estimate $\tilde{\Lambda}$. In an extreme case, if $p(\Lambda|\mathcal{X}) = \delta(\Lambda - \tilde{\Lambda})$ with $\delta(\cdot)$ denoting the Kronecker delta function, namely, the posterior probability mass of $\Lambda$ is concentrated at the ML estimate $\tilde{\Lambda}$ obtained from $\mathcal{X}$, then it is easy to see from Equations (4) and (5) that the BPC decision rule coincides with the plug-in MAP decision rule.

In the robust speech recognition problem we are considering, it is assumed that there are mismatches between training and testing conditions which often result in a performance degradation in comparison with the matched conditions. Because of the nature of many speech recognition applications, the mismatch involved could be of any types as discussed before. It is thus desirable to develop a general robust speech recognition approach that is cable of handling any mismatches which might encounter in real applications. By considering the reality of the statistical pattern recognition paradigm and the modeling techniques we are using, we always have to make some assumptions which are often violated for real observed data. One way to achieve the performance robustness is thus to design and construct a robust decision rule which considers the *model uncertainty* and thus is not so sensitive to definite types of distortions. Because we are using a parametric model namely HMM, a simple thus also limited way to consider the model uncertainty is via considering the *model parameter uncertainty*, i.e., any perturbation of the model parameter values will cause a perturbation of the range of the observed data which the assumed model can correctly represent. Thus the principle behind the BPC approach is rather straightforward: Because we assume no knowledge about the possible mismatch, we thus rely on a quite general prior pdf to characterize the variability of the HMM parameters caused by the possible modeling/estimation errors and/or mismatches between training and testing conditions. We try to average out this variability while making decision with BPC. More specifically, we start with where Nadas [58] left off, with an *empirical Bayes* method in which a specific parametric pdf $p(\Lambda|\varphi)$ is adopted to represent the prior/posterior pdf of the CDHMM parameters. Its hyperparameters $\varphi$ could be estimated from some training data, or specified based on some empirical reasoning, or their combination [26, 28]. Consequently, the predictive pdf required for BPC decoding will be computed as:

$$\tilde{p}(\mathbf{X}|W) = \int_{\Omega} p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \tag{7}$$

The crucial difference of $p(\Lambda|\varphi)$ and $p(\Lambda|\mathcal{X})$ is that the former is *inflated* appropriately and thus

less sharp than $p(\Lambda|\mathcal{X})$. This provides the BPC chance to make a difference from the conventional plug-in MAP decoder.

### 3.2.3 Approximate BPC Approach

In the CDHMM case, due to the nature of the *missing data* problem in HMM formulation, it is not easy to compute the following true predictive pdf:

$$\tilde{p}(\mathbf{X}|W) = \sum_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X},\mathbf{s},\mathbf{l}|\Lambda,W)p(\Lambda|\varphi,W)\mathrm{d}\Lambda \tag{8}$$

where $\mathbf{s}$ is the unobserved state sequence and $\mathbf{l}$ is the associated sequence of the unobserved mixture component labels corresponding to the observation sequence $\mathbf{X}$. Consequently, some approximations are needed.

One way to compute an approximate predictive pdf is to use the Monte Carlo method. We can use the Monte Carlo simulation of the hidden processes (state sequence and mixture label sequence) of the CDHMM and then perform integration and averaging. We can also perform a double-fold Monte Carlo simulation of both the hidden processes and the HMM parameters, and then perform only averaging. Because it's computationally expensive, the Monte Carlo method has only of academic interest in the stage of performing speech recognition.

Another way to compute the approximate predictive pdf is to use the following Viterbi approximation:

$$\tilde{p}(\mathbf{X}|W) \approx \max_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X},\mathbf{s},\mathbf{l}|\Lambda,W)p(\Lambda|\varphi,W)\mathrm{d}\Lambda \tag{9}$$

A detailed algorithm to implement the above approximation and the related experimental results are reported in [37]. The resultant BPC rule is called Viterbi BPC (VBPC) rule.

The third way is to adopt a numerical approximation technique, namely, *Laplace approximation* for integral, to compute the approximate predictive pdf as follows:

$$\tilde{p}(\mathbf{X}|W) \approx p(\mathbf{X}|\Lambda_{MAP},W) \cdot p(\Lambda_{MAP}|\varphi,W) \cdot (2\pi)^{\mathcal{M}/2} \cdot |V|^{1/2} \tag{10}$$

where $\Lambda_{MAP}$ is the MAP estimate as shown in Equation (14), $\mathcal{M}$ is the number of HMM parameters involved in the integrand in Equation (7), and $V$ is the $\mathcal{M} \times \mathcal{M}$ modal dispersion matrix, i.e., $-V^{-1}$ is the Hessian matrix of second derivatives of

$$\hbar(\Lambda) = \log\{p(\mathbf{X}|\Lambda,W)p(\Lambda|\varphi,W)\} \tag{11}$$

evaluated at $\Lambda = \Lambda_{MAP}$. We note that the Laplace approximation is essentially derived by retaining the quadratic term in the Taylor expansion of $\hbar(\Lambda)$ and are thus equivalent to *normal-like approximation* to the integrand, namely, using a normal pdf $\mathcal{N}(\Lambda|\Lambda_{MAP},V)$ to approximate the posterior pdf $p(\Lambda|\mathbf{X},W)$. So, this approximation technique is also known as the *normal approximation* method in the Bayesian community. In the case of CDHMM, to compute $V$ directly is still too computationally involved. So, we have to make further approximation. If we only consider the uncertainty of the mean vectors in CDHMM for BPC decoding, we can use the QB algorithm in [28] or [29] to compute an approximate posterior pdf $\mathcal{N}(\Lambda;\Lambda_{MAP},\tilde{U})$ and then replace $V$ in equation (10) with $\tilde{U}$. The resulted BPC rule is thus named as the QBPC (*quasi-Bayesian predictive classification*) rule [30, 31, 32, 33].

Both QBPC and VBPC methods have been shown via a series of comparative experiments in [30, 31, 32, 33, 37, 39] to be able to greatly enhance the robustness when mismatches exist between training and testing conditions. We can actually go one step further. By combing BPC decision strategy with the on-line model adaptation strategy to continuously update our prior knowledge about the uncertainty of the model parameters, we can approach a performance achieved by the plug-in MAP rule under a matched condition [31, 32, 39].

### 3.2.4 Relation to Other Robust Decision Approaches

In addition to the above BPC approach, another way to achieve performance robustness in the unknown mismatch case via considering the *model uncertainty* is to adopt the so-called *minimax principle* in which the essence is to try and protect against the worst possible mismatch within *some classes*. Therefore, the minimax approach is considered to be the most conservative decision strategy. Merhav and Lee presented a case study of minimax classification for robust speech recognition in [56]. In that approach, instead of only using the estimated values of $\Lambda$ as in plug-in MAP rule, like in BPC, it is also assumed that the *true* parameters $\Lambda$ are uncertain (random variables) and randomly distributed in a *neighborhood* region $\Omega$ around the estimated ones. If we have no further knowledge about $\Lambda$, a reasonable decision is to warrant the optimal outcome (e.g. minimum classification error) in the possibly worst-case condition (e.g. maximum mismatch in the assumed uncertainty neighborhood). Under some assumptions, Merhav and Lee proposed such a *minimax decision rule* which minimizes the *worst-case probability of classification error*. It turns out to be too difficult to implement this decision rule. They then suggested a weaker decision rule which seeks to minimize an *upper bound* of the *worst-case probability of classification error*. Although it was still called a minimax decision rule, it has a much loose meaning than what it usually means in statistics literature. Such a minimax decision rule operates as follows:

$$\hat{W} = \arg\max_{W}[P_{\Gamma}(W) \cdot \max_{\Lambda \epsilon \Omega} p(\mathbf{X}|\Lambda, W)] \tag{12}$$

As discussed in [56], it can also be viewed as a two-step procedure. First, each testing utterance is treated to possibly belong to any word sequence and a constrained ML estimate of the related HMM parameters is obtained. Then, a plug-in MAP rule is used for speech recognition by using the updated HMM parameters. This intuitive interpretation opens up the possibilities to use other estimation approaches, e.g., MAP approach, in the first step. This is exactly what we did in a set of comparative experiments of BPC with minimax approach in [30, 31, 32, 33]. Such a modified minimax decision rule works as follows:

$$\hat{W} = \arg\max_{W} p(\mathbf{X}|\Lambda_{MAP}, W) \cdot P_{\Gamma}(W) \tag{13}$$

where

$$\Lambda_{MAP} = \arg\max_{\Lambda \in \Omega_W} p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W) \tag{14}$$

The minimax strategy tries to secure the decision in the worst case of the assumed mismatch, thus usually does not perform nearly as well as in a less malign situation and/or those techniques which use some prior information of the possible mismatches. In some applications, if a rough knowledge of the distortion is available, then it can be used to design a *structural* model which takes advantage of some structural constraints and thus only includes a small number of *nuisance parameters* to characterize the systematic distortion structure. The compensation can then be performed via on-line estimation of these nuisance parameters from the given pre-trained models and the available testing data. A so-called *stochastic matching* approach described in [70] is such a natural extension of structure-based compensation from minimax approach. The similar extension can also be applied to the BPC approach.

## 3.3 Sensitivity of Priors

In the Bayesian framework we are adopting, one of the factors which greatly influences the efficacy of the OLA and BPC approaches is the appropriateness of the prior pdf. Generally speaking, prior density estimation and the choice of density parameters depend on the particular application of interest. Because we have already assumed a specific parametric form for the prior pdf in our study, this turns out to be a hyperparameter specification/estimation problem. If the training data set $\mathcal{X}$ is rich and big enough to cover the interested variability of speech signal which possibly occurs in the testing conditions, then the *method of moment* algorithm presented in [26] can be used to automatically estimate the hyperparameters from the training data $\mathcal{X}$. Otherwise we have to use some *ad hoc* method for hyperparameter estimation. One of such methods is described in [28]. If application scenario allows us to have access to some testing data, by using sequential Bayesian learning method in [28, 29], we can also make the prior pdf more appropriate. Furthermore, the *knowledge* and/or *experience* of the interaction between speech signal and the possible mismatch will also be very helpful to guide us to obtain a better prior pdf as shown in [32, 33].

# 4 Putting It All Together: A Robust HMM-based ASR System

Because both OLA and BPC are formulated under a unified Bayesian paradigm to address respectively the model parameter inference problem and the decision problem, they can be naturally combined to produce an enhanced algorithm to cope with the robust ASR problem as described before. Such a robust ASR system is schematically shown in Figure 1.

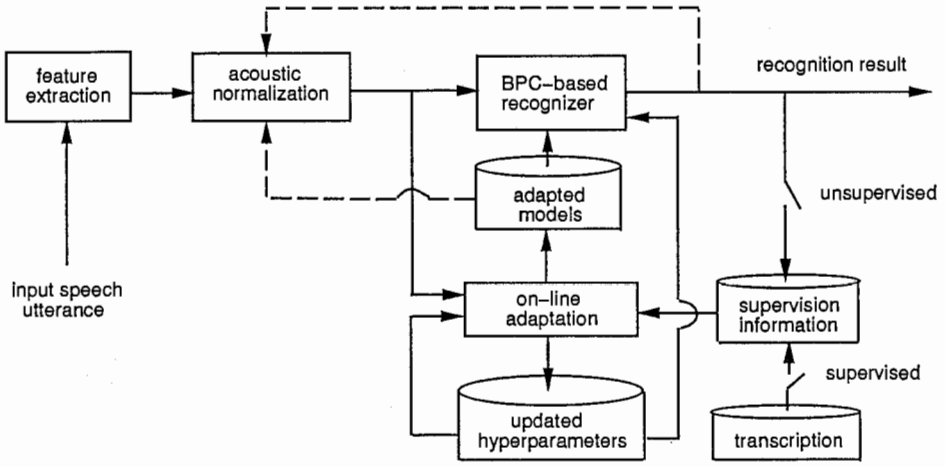


Figure 1: A block diagram of a robust HMM-based speech recognition system

Given a new block of input speech, feature extraction (usually spectral analysis) is first performed to derive the feature vector sequences used to characterize the speech input. It is followed by some kind of acoustic normalization to reduce the possible mismatch in the feature vector space. The processed feature vector sequences are then recognized based on the current set of HMM's by using BPC approach. After the recognition of the current block of utterances, the HMM's and the posterior distributions of the related speech units are adapted and the updated models are used to recognize future input utterance(s). In this way, we can get a better and better posterior/prior pdf (i.e., more and more accurate knowledge about the uncertainty of the model parameters), and this in turn makes the BPC-based recognition system approach a performance achieved by the plug-in MAP rule under a matched condition.

For the acoustic normalization/equalization module shown in Figure 1, many existing techniques can be applied. They include, for example, the popular cepstral mean subtraction algorithm [4], different cepstral normalization methods (e.g. CDCN and others in [1]), ML-based feature space stochastic matching methods (e.g., [11, 81, 70]), signal conditioning techniques (e.g., [64, 65]), etc. Acoustic normalization could even be integrated into the feature extraction stage, e.g., speaker normalization via vocal tract length normalization using frequency warping (e.g., [77, 16, 51]). Encouraging results have also been demonstrated in combined acoustic normalization and model adaptation based on a small amount of calibration data (e.g., [81, 82]).

On-line model adaptation is a data-driven method and its strength comes from the availability of a certain amount of test data. If the application involves a recognition session which might consist of a number of testing utterances, then a combined BPC decoding and on-line adaptation of the prior of the HMM parameters will provide a good solution to enhance the robustness towards varying environments, microphones, channels, speakers, and other general mismatches or distortions. For real-world applications, unsupervised on-line adaptation is usually more realistic and desirable. One of the remaining research issues is how to guide the unsupervised OLA when the recognition rate is initially low. Different degree of parameter tying and/or smoothing might be helpful. Incorporating some data validation mechanism will also be useful and more theoretical works are needed to develop a better verification paradigm.

# 5 Discussion and Conclusion

We have presented an overview of the Bayesian approaches to robust speech recognition we've been developing in the past few years. From the lessons we learned thus far, we expect that a better understanding and more experience on how the speech signal is distorted and/or varied under different acoustic conditions will be helpful to design a better parametric form and the related hyperparameter estimation of the prior pdf's for both OLA and BPC. It will also be helpful to design a better structural model in structure-based compensation. It will be crucial for efficient adaptation and compensation to formulate and develop appropriate mathematical tools for discovering a good intrinsic structural model of speech in the acoustic, phonetic and linguistic aspects. We are continuously exploring other possibilities for

- robust pattern verification,
- structural model design,
- efficient learning,
- robust and selective learning,
- intelligent and flexible dialogue control,

which by combining with other techniques, will lead to a robust and intelligent spoken dialogue system for many useful applications. The greatest challenge might come from those applications which only involve a couple of utterances, but every utterance involves a distinct "distortion channel" from the intended message to the received signal. How to reliably and efficiently recover and/or extract the interested message from this signal pose a big challenge for the so-called robust ASR in this context.

# 6 Acknowledgement

I would like to thank my collaborators, Dr. Chin-Hui Lee and Mr. Hui Jiang, for their contributions on the works described in this report. I also want to thank Dr. Y. Yamazaki, past President, Dr. S. Yamamoto, current President, ATR Interpreting Telecommunications Research Laboratories, and Dr. Y. Sagisaka, Head, Speech Recognition Department of the ATR-ITL, for their support of this work during my stay at ATR from April 1995 to December 1997. I appreciate it very much that Mr. Dimitry Rtischev implemented, with the help of Mr. Harald Singer [69], the on-line adaptation algorithm described in [28] into the ATRSPREC system which makes it possible for other interested people (e.g., Mr. Hirofumi Yamamoto [78]) to perform more experiments, show the usefulness of the algorithm, and hopefully refine and extend it to a better solution.

## References


[1] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, Kluwer Academic Publishers, 1993.

[2] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 11, pp.187-206, 1997.

[3] J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis*, Cambridge, UK: Cambridge University Press, 1975.

[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, Vol. 55, No. 6, pp.1304-1312, 1974.

[5] L. R. Bahl, F. Jelinek and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, No. 2, pp.179-190, March 1983.

[6] L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. ICASSP-96*, 1996, pp.49-52.

[7] L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy," *IEEE Trans. Speech and Audio Processing*, Vol. 1, No. 1, pp.77-83, 1993.

[8] J. K. Baker, "Stochastic modeling for automatic speech understanding," in *Speech Recognition* (D. R. Reddy, ed.), New York: Academic, 1975, pp.521-542.

[9] J. K. Baker, "The DRAGON system – an overview," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-23, pp.24-29, Feb. 1975.

[10] L. E. Baum, "An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, Vol. 3, pp.1-8, 1972.

[11] S. Cox and J. Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting," *Proc. ICASSP-89*, pp.294-297.

[12] S. Cox, "Predictive speaker adaptation in speech recognition," *Computer Speech and Language*, Vol. 9, pp.1-17, 1995.

[13] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 5, pp.357-366, 1995.

[14] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 4, pp.294-300, 1996.

[15] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.

[16] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Proc. ICASSP-96* (Atlanta), May 1996, pp.346-349.

[17] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. on Information Theory*, Vol. 35, No. 5, pp.1001-1013, 1989.

[18] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Trans. on Information Theory*, Vol. 36, No. 2, pp.372-380, 1990.

[19] S. Furui, "A training procedure for isolated word recognition systems," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 2, pp.129-136, 1980.

[20] S. Furui, "Recent advances in robust speech recognition," *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition For Unknown Communication Channels*, Pont-a-Mousson, France, April 1997, pp.11-20.

[21] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, Vol. 10, pp.249-264, 1996.

[22] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp.291-298, April 1994.

[23] S. Geisser, *Predictive Inference: An Introduction*, New York: Chapman & Hall, 1993.

[24] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo and M. A. Picheny, "Decoder selection based on cross-entropies," in *Proc. ICASSP-88*, 1988, pp.20-23.

[25] H. Hattori and S. Sagayama, "Vector field smoothing principle for speaker adaptation," *Proc. ICSLP-92*, Oct. 1992, pp.381-384.

[26] Q. Huo, C. Chan and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 5, pp.334-345, 1995.

[27] Q. Huo, C. Chan and C.-H. Lee, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, pp.141-144, 1996.

[28] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, pp.161-172, 1997.

[29] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," to appear in *IEEE Trans. on Speech and Audio Processing*. See also a condensed version with the same title in *Proc. ICSLP-96*, pp.985-988, 1996.

[30] Q. Huo, H. Jiang and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *Proc. ICASSP-97* (Munich, Germany), April 1997, pp.II-1547-1550.

[31] Q. Huo and C.-H. Lee, "Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition," *Proc. Eurospeech-97*, Rhodes, Greece, September 1997.

[32] Q. Huo and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," submitted to *IEEE Trans. on Speech and Audio Processing*, August 1997.

[33] Q. Huo and C.-H. Lee, "A study of prior sensitivity for Bayesian predictive classification based robust speech recognition," submitted to *Proc. ICASSP-98*, Seattle, May 12-15, 1998.

[34] J. Ishii, M. Tonomura, and S. Matsunaga, "Speaker adaptation using tree structured shared-state HMMs," *Proc. ICSLP-96* (Philadelphia), October 1996.

[35] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, Vol. 64, No. 4, pp.532-556, April 1976.

[36] F. Jelinek, R. L. Mercer and S. Roukos, "Principles of lexical language modeling for speech recognition," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi (eds.), New York: Marcel Dekker, 1991, pp.651-699.

[37] H. Jiang, K. Hirose and Q. Huo, "Robust speech recognition based on Bayesian prediction approach", submitted to *IEEE Trans. on Speech and Audio Processing*, August 1997. See also a condensed version titled "Robust speech recognition based on Viterbi Bayesian predictive classification," in *Proc. ICASSP-97*, pp.II-1551-1554, 1997.

[38] H. Jiang, K. Hirose and Q. Huo, "Sequential Bayesian learning of CDHMM based on finite mixture approximation of its prior/posterior density," *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU'97), Santa Barbara, December 1997.

[39] H. Jiang, K. Hirose and Q. Huo, "Improving Viterbi Bayesian predictive classification via sequential Bayesian learning in robust speech recognition," submitted to *Proc. ICASSP-98*, Seattle, May 12-15, 1998.

[40] B.-H. Juang, S. E. Levinson and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. on Information Theory*, Vol. IT-32, No. 2, pp.307-309, 1986.

[41] B.-H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, Vol. 5, pp.275-294, 1991.

[42] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, Vol. 40, No. 12, pp.3043-3054, 1992.

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

[44] S. Katagiri, C.-H. Lee and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 1991, pp.299-308.

[45] M. J. Lasry and R. M. Stern, "*A posteriori* estimation of correlated jointly Gaussian mean vectors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 4, pp.530-535, 1984.

[46] L. R. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. on Information Theory*, Vol. IT-28, pp.729-734, 1982.

[47] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. on Signal Processing*, Vol. 39, No. 4, pp.806-814, 1991.

[48] C.-H. Lee, F.-K. Soong and K.-K. Paliwal (eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, Boston: Kluwer Academic Publishers, 1996.

[49] C.-H. Lee, B.-H. Juang, W. Chou and J. J. Molina-Perez, "A study on task-independent subword selection and modeling for speech recognition," *Proc. ICSLP-96* (Philadelphia), pp. 1820-1823, October 1996.

[50] C.-H. Lee, "On feature and model compensation approach to robust speech recognition," *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition For Unknown Communication Channels*, Pont-a-Mousson, France, April 1997, pp.45-54.

[51] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," *Proc. ICASSP-96* (Atlanta), May 1996, pp.353-356.

[52] 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.

[53] A. Ljolje, Y. Ephraim, and L. R. Rabiner, "Estimation of hidden Markov model parameters by minimizing empirical error rate," in *Proc. ICASSP-90*, 1990, pp.709-712.

[54] T. Matsuoka and C.-H. Lee, "A study of on-line Bayesian adaptation for HMM-based speech recognition," *Proc. EUROSPEECH-93* (Berlin, Germany), 1993, pp.815-818.

[55] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. on Signal Processing*, Vol. 39, No. 10, pp.2157-2166, 1991.

[56] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 1, pp.90-100, 1993.

[57] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-31, No. 4, pp.814-817, 1983.

[58] A. Nadas, "Optimal solution of a training problem in speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 1, pp.326-329, 1985.

[59] A. Nadas, D. Nahamoo, and M. A. Picheny, "On a model-robust training method for speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-36, No. 9, pp.1432-1436, 1988.

[60] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," *Proc. ICSLP-92*, Oct. 1992, pp.369-372.

[61] D. B. Paul, J. K. Baker and J. M. Baker, "On the interaction between true source, training, and testing language models," in *Proc. ICASSP-91*, 1991, pp.569-572.

[62] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp.257-286, 1989.

[63] L. R. Rabiner, B.-H. Juang and C.-H. Lee, "An overview of automatic speech recognition," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F.-K. Soong & K.-K. Paliwal, eds., Boston: Kluwer Academic Publishers, 1996, pp.1-30.

[64] M. G. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 1, pp.19-30, 1996.

[65] M. G. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp.107-109, 1996.

[66] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge, UK: Cambridge University Press, 1996.

[67] O. Ronen and M. Ostendorf, "A dependence tree model of phone correlation," *Proc. ICASSP-96* (Atlanta), May 1996, pp.873-876.

[68] W. A. Rozzi and R. M. Stern, "Speaker adaptation in continuous speech recognition via estimation of correlated mean vectors," *Proc. ICASSP-91* (Toronto), May 1991, pp.865-868.

[69] D. Rtischev, Q. Huo, and H. Singer, "Implementation and testing of quasi-Bayes speaker adaptation algorithm," *Technical Report* TR-IT-0185, ATR Interpreting Telecommunications Research Laboratories (2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan), September 1996.

[70] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp.190-202, 1996.

[71] B. M. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, pp.183-191, 1997.

[72] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDL principle," *Proc. ICASSP-96* (Atlanta), May 1996, pp.717-720.

[73] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 6, pp.751-763, 1987.

[74] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation," *Computer Speech and Language*, Vol. 11, pp.127-146, 1997.

[75] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum *a posteriori* probability estimation," *Computer Speech and Language*, Vol. 10, pp.117-132, 1996.

[76] M. Tonomura, T. Kosaka, S. Matsunaga, and A. Monden, "Speaker adaptation fitting training data size and contents," *Proc. EUROSPEECH-95* (Madrid), September 1995, pp.1147-1150.

[77] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," *Proc. ICASSP-96* (Atlanta), May 1996, pp.339-341.

[78] H. Yamamoto, H. Singer, A. Nakamura, and Q. Huo, "Unsupervised quasi-Bayes on-line speaker adaptation using language information," *Proceedings of the 1997 Fall Meeting of the Acoustical Society of Japan* (ASJ'97 Fall Meeting), Sapporo, Japan, September 16-19, 1997, pp.21-22, (written in Japanese).

[79] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," *Proc. ICASSP-95* (Detroit), May 1995, pp.I–676–679.

[80] G. Zavaliagkos, R. Schwartz, J. McDonough, and J. Makhoul, "Adaptation algorithms for large scale HMM recognizers," *Proc. EUROSPEECH-95* (Madrid), September 1995, pp.1131-1134.

[81] Y.-X. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 3, pp.380-394, 1994.

[82] Y.-X. Zhao, "Self-learning speaker and channel adaptation based on spectral variation source decomposition," *Speech Communication*, Vol. 18, pp.65-77, 1996