ELSEVIER

Speech Communication 34 (2001) 175–194

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

# Robust speech recognition based on adaptive classification and decision strategies

Qiang Huo [a,*], Chin-Hui Lee [b,1]

[a] *Department of Computer Science & Information Systems, The University of Hong Kong, Pokfulam Road, Hong Kong, People's Republic of China*

[b] *Dialogue Systems Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA*

**Abstract**

We examine key research issues in adaptively modifying the conventional *plug-in MAP decision rules* in order to improve the robustness of the classification and decision strategies used in automatic speech recognition (ASR) systems. It is well known that the commonly adopted plug-in MAP decoder does not achieve the minimum error rate desired in ASR because the joint probability distribution of speech and language is usually not known exactly. The optimality issue becomes even more serious when there exists acoustic mismatch between training and testing conditions. We review in detail two recently proposed classification rules, namely *minimax classification* and *Bayesian predictive classification*. Both of them model classifier parameter uncertainty and modify the classification rules to satisfy some desired robustness properties. We also present an overview on a number of related techniques and discuss how these algorithms can be used to improve the robustness of speech recognizers. 

*Keywords:* Decision rules; Plug-in MAP decision rule; Minimax decision rule; Bayesian predictive classification; Adaptive classification; Adaptation; Compensation; Robust automatic speech recognition

## 1. Introduction

In the last two decades, many advances have been achieved in the area of automatic speech recognition (ASR) (see (Lee et al., 1996) for a sample of the state-of-the-art). This is largely attributed to the use of a powerful statistical pattern recognition paradigm. In this approach, let us view a "word", $W$, and its corresponding acoustic observation (in practice, usually a feature vector sequence extracted from the speech signal), $X$, as a jointly distributed random pair $(W,X)$. 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, a semantic attribute, etc. If the joint probability density function (pdf) $p(W,X)$ is known exactly, an *optimal decoder* (speech recognizer) which achieves the expected minimum word recognition error rate is the following maximum a posteriori (MAP) decoder (e.g., Ripley, 1996):

$$\hat{W} = \arg\max_W P(W|X) = \arg\max_W p(W,X) = \arg\max_W p(X|W)\cdot P(W), \quad (1)$$

* Corresponding author. Tel.: +852-2857-8459; fax: +852-2559-8447.

*E-mail addresses:* qhuo@csis.hku.hk (Q. Huo), chl@research.bell-labs.com (C.-H. Lee).

[1] Tel.: +908-582-5226; fax: +908-582-7308

0167-6393/01/$ - see front matter 
PII: S0167-6393(00)00053-4

where $\hat{W}$ is the recognition result. However, in practice, neither do we know the true parametric form of $p(W,X)$, nor do we have the knowledge about its true parameter values. We shall say that we have *prior uncertainty* in this case. Therefore, the above optimal speech recognizer is never realizable. A simple approximation is first to *assume* some parametric form for $p(W,X)$, e.g., $p(W,X) = p_\Lambda(X|W) \cdot P_\Gamma(W)$, where $p_\Lambda(X|W)$ is known as the acoustic model with parameters $\Lambda$, and $P_\Gamma(W)$ as the language model with parameters $\Gamma$; and then to *estimate* its parameters $(\Lambda, \Gamma)$ from some training data $\mathscr{X}$. The pdf estimate $\{p_{\hat{\Lambda}}(X|W), P_{\hat{\Gamma}}(W)\}$ (with the model parameter estimate denoted as $\{\hat{\Lambda}, \hat{\Gamma}\}$) is then plugged into the rule in Eq. (1) in place of the correct but unknown $\{p(X|W), P(W)\}$ to obtain a *plug-in MAP decision rule,*

$$\begin{aligned}\hat{W} &= \arg\max_W P(W|X) \\ &= \arg\max_W p_{\hat{\Lambda}}(X|W) \cdot P_{\hat{\Gamma}}(W). \end{aligned} \tag{2}$$

Readers are referred to (Kharin, 1996; Huo, 1999) for discussions of the optimality of the above plug-in MAP decision rule.

Currently, the most widely used and the most successful modeling approach to ASR is to use a set of hidden Markov models (HMMs) as the acoustic models of subword or whole-word units (e.g., Rabiner, 1989), and to use the statistical $N$-gram model or its variants as lexical language models for words and/or word classes (e.g., Jelinek et al., 1991). By using the abovementioned plug-in MAP decision rule, it has been repeatedly shown by experiments in the past three decades that given a large amount of *representative* training speech and text data, good statistical acoustic and language models can be constructed which achieve a high performance for many ASR tasks. This has given the speech research community a certain level of confidence in believing that the Discrete HMM (DHMM) and the Gaussian-mixture Continuous Density HMM (CDHMM), together with $N$-gram model, provide a good approximate parametric form for $p_\Lambda(X|W)$ and $P_\Gamma(W)$ (these models are apparently imperfect but very flexible and mathematically well-defined). Based on the belief that these models are good approximations, the maximum likelihood (ML) estimate for the HMM parameters and $N$-gram model parameters has been the most popular parameter estimation method.

However, the principles for the construction of the above-mentioned optimal MAP decision rule and plug-in MAP rule are based on some assumptions which may be violated in practice (Kharin, 1996; Huo, 1999). From a computational modeling point of view, a classification of main distortion types which produce violations of assumptions is as follows (Kharin, 1996):

- distortions caused by small-sample effects;
- distortions of models for training samples;
- distortions of models for testing observations to be classified.

The distortions caused by small-sample effects are typical for all statistical plug-in procedures. They arise from the noncoincidence of the statistical estimates $\{P_{\hat{\Gamma}}(W), p_{\hat{\Lambda}}(X|W)\}$ of probability characteristics and their true values $\{P(W), p(X|W)\}$. So, the design and/or collection of the training samples become very critical. The key is to make the samples in $\mathscr{X}$ follow the intended distribution $p(W,X)$ as closely as possible. Otherwise, some more intelligent ways of using the available training data must be developed. As for the distortions of the models for the training samples, they can be caused by the wrong assumptions and/or inflexible parametric forms of the model; the misclassification of training samples; outliers in training samples, etc. To cope with this problem, better models need to be found and techniques need to be designed for robust learning from data. The biggest problem for ASR might be caused by the third type of distortions, the distortions of the models for the observations to be classified. In most real applications, there always exists some form of mismatch which causes a distortion between the trained models and the test data. A conceptual illustration is shown in Fig. 1 (taken from (Sankar and Lee, 1996)). $D_1(\cdot)$, $D_2(\cdot)$ and $D_3(\cdot)$ characterize the possible distortion in the signal, feature and model spaces, respectively. These mismatches may arise from inter- and intra-speaker variabilities; transducer, channel and other environmental variabilities; and many other phonetic and

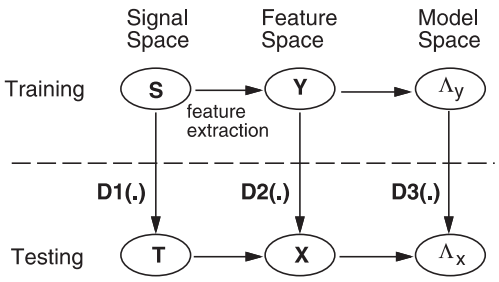


Fig. 1. Mismatch between training and testing.

linguistic effects caused by mismatch in training and testing task definitions. Robust speech recognition in this context thus refers to the topic of designing speech recognizers that work well for different tasks and speakers over unexpected and possibly adverse operating conditions.

Traditional approaches to robust speech recognition include finding invariant or robust features and developing better modeling and learning techniques. In addition, three major classes of statistical techniques to improve ASR robustness can be defined:

- modifying recognizer parameters to new speakers or speaking conditions based on adaptation data;
- applying feature and model compensation techniques based only on testing data;
- using robust decision strategies.

*Adaptation* and *compensation* are two closely related topics that have been intensively studied. Many techniques that were originally developed for adaptation can be extended to compensation which can be considered as an *unsupervised adaptation* with testing data. *Robust decision* is a rather new research area and serves as the focus of this paper.

The main purposes of this paper are to present an overview of some recent techniques on the topics of adaptation and compensation, and to review in detail the topic of robust decision rules for automatic speech recognition. The remainder of the paper is organized as follows. In Section 2, we first establish the connection between adaptation/compensation and robust decision, and then define the decision rule robustness. In Sections 3 and 4, we present two examples of the robust decision rules, namely the *minimax decision rule* and the *Bayesian predictive classification rule*. In Section 5, we review some other related robust decision approaches. In Section 6, we report on some illustrative experimental results to show how these robust decision rules work. Section 7 summarizes the ideas discussed in the paper.

## 2. From adaptation to compensation to robust decision

One of the effective ways to handle mismatch seems to be finding invariant (or robust) features so as to minimize the acoustic mismatch between training and testing environments. Even though some features have been shown less affected by a certain type of distortion, such as linear microphone or channel effect, it has not been discovered yet any feature that is invariant across all adverse acoustic conditions. Further research in front-end signal processing and feature extraction is definitely needed to improve on the currently "standard" acoustic analysis for ASR (Hermansky, 1998; Morgan, 1999; Hunt, 1999). In addition to finding the more robust features for ASR, another more straightforward solution is to collect additional training data in a specific testing condition and then to adapt the recognizer parameters accordingly to work in the prescribed scenario.

### 2.1. Classifier parameter adaptation

For HMM-based speech recognition systems, adaptation is usually accomplished in two ways:

- *direct adaptation* of the HMM parameters;
- *indirect adaptation* of a set of transformation parameters which induces the adapted HMM parameters through transformation.

Bayesian parameter learning is the dominant approach to direct adaptation (e.g. Lee et al., 1991; Gauvain and Lee, 1994; Huo et al., 1995). It provides a mathematical framework for combining prior information embedded in a general set of stochastic models (or a set of training data) and the information embedded in a specific set of adaptation data. It also has the nice asymptotic property that the more adaptation data we use, the

closer the adapted model parameters to their desired ML estimate which typically leads to a better recognition performance in practice. The simplest direct adaptation scenario aims at *local* adaptation of units seen in adaptation data. In many situations when not all the units are observed with enough instances, Bayesian adaptation does not work as effectively as anticipated. To enhance the efficiency and the effectiveness of Bayesian adaptive learning, it is desirable to introduce some constraints on the HMM parameters. By this means 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 the adaptation data. A simple way to achieve the above objective is to introduce parameter tying. Another way to achieve the above objective is to explicitly consider the correlation of HMM parameters corresponding to different speech units (e.g., Lasry and Stern, 1984; Stern and Lasry, 1987; Zavaliagkos et al., 1995a,b; Shahshahani, 1997; Huo and Lee, 1998a).

As for indirect adaptation, the most popular technique assumes that the adapted model parameters are obtained through a linear transformation of the original model parameters. The transformation parameters can be estimated from adaptation data under the ML or MAP criterion by using an EM algorithm (Dempster et al., 1977). If ML criterion is used, the technique is known as the maximum likelihood linear regression (MLLR) (e.g., Leggetter and Woodland, 1995a; Gales and Woodland, 1996; Gales, 1998), the *maximum likelihood constrained estimation of HMM* (Digalakis et al., 1995; Diakoloukas and Digalakis, 1999), and the *stochastic matching* (Sankar and Lee, 1996). If MAP criterion is used, the technique is known as the MAP-LR (e.g., Chien and Wang, 1997; Paul, 1997; Shinoda and Lee, 1997, 1998; Siohan et al., 1999; Mokbel and Delphin-Poulat, 1999; Chesta et al., 1999; Chou, 1999). In these methods, sharing transformations among HMM parameters provides another way of imposing a *global constraint* over all the model units so that *all* the parameters can be adapted at the same time even though not all the units have been observed in the adaptation data. This approach works quite well especially in situations when only a small amount of adaptation data is used. However, the performance often saturates fast and settles at a performance level not as good as that achieved by Bayesian adaptation when a large amount of adaptation data can be used (e.g. Digalakis and Neumeyer, 1996; Chien et al., 1997; Huo and Ma, 1999). Another potential limitation of MLLR is the need to determine the number of transformation matrices according to the size of adaptation data. So far only heuristic selection rules have been adopted. Recently, there have been increasing efforts to combine the merits of the above two methodologies. A series of hybrid methods have been developed (e.g., Digalakis and Neumeyer, 1996; Zhao, 1994, 1996; Chien et al., 1997; Huo and Ma, 1999; Siohan et al., 2000).

Once a recognition system has been designed, it can further be improved based on a *dynamic* strategy such that new knowledge and information are acquired and incorporated incrementally during the development and use of the ASR system. One type of the adaptive learning algorithm is often referred to as *on-line Bayesian adaptive learning* (Huo and Lee, 1997a, 1998a; Huo and Ma, 1999; Jiang et al., 1999b; Chien, 1999). Another type of algorithm is known as the on-line MLLR (Leggetter and Woodland, 1995b; Digalakis, 1999). There are also many other on-line adaptation algorithms in literature (e.g., Zavaliagkos et al., 1995a,b; Zhao, 1996; Takahashi and Sagayama, 1997; Shinoda and Lee, 1998; Wang and Zhao, 1999). In addition, there is a great potential for a unified Bayesian framework based on the concepts of variability source decomposition for prior elicitation, on-line prior fusion for improving compensation and robust decision, and multiple-stream prior evolution and posterior pooling for continuously improving adaptation and robust decision (Huo and Ma, 1999). More powerful adaptation algorithms will soon appear based on this framework. In Section 6.5, we will show a simple example of how to use prior evolution to improve the robust decision.

Whether supervision is available makes a significant difference when adopting a dynamic strategy for adaptation. *Supervised adaptation* in ASR often refers to the situation in which the *transcription* corresponding to the adaptation

utterances is available. *Unsupervised adaptation*, on the other hand, means an assumed transcription has to be obtained with the speech recognizer and then used to aid the adaptation process. Since the ASR systems usually make transcription errors, unsupervised adaptation does not work as well as supervised adaptation. It has also been observed that transformation-based adaptation performs better than Bayesian adaptation in large vocabulary recognition using unsupervised adaptation with short adaptation data. Part of the reason is that a rough and global adaptation strategy does not rely as much on correct supervision as the local Bayesian adaptation scenario.

### 2.2. Feature and model compensation

Compensation can be considered as a form of unsupervised adaptation in which only the testing data are used. Many other names have also been adopted, e.g., *self adaptation*, *auto adaptation*, *instantaneous adaptation*, or *stochastic matching* (e.g., Zavaliagkos et al., 1995a; Zhao, 1996; Sankar and Lee, 1996). For robust speech recognition, compensation can be accomplished in the signal, feature and model spaces in order to reduce the distortions shown in Fig. 1. The readers are referred to a recent review on the topic of feature and model compensation (Lee, 1998).

One of the earliest studies on feature compensation is cepstral mean normalization (CMN) (Atal, 1974) which removes the cepstral mean of each utterance before training and testing. CMN was shown to be robust to microphone and channel distortions in many systems. By making CMN more effective for different sounds in different speaking conditions, codeword-dependent cepstral normalization (CDCN) (Acero, 1993) was then developed. A simplified version known as signal bias removal (SBR) or signal conditioning was shown to be effective for several applications (e.g., Zhao, 1994; Rahim and Juang, 1996; Rahim et al., 1996; Lawrence and Rahim, 1999). Typically, a codebook is used to represent the reference acoustic space and then a set of biases can be derived to compensate cepstral difference between testing feature vectors and reference codebook. When no training data is available to create the codebook, a natural extension is to use the information embedded in the acoustic HMMs to aid the feature compensation process (e.g., Zhao, 1994; Sankar and Lee, 1996; Lawrence and Rahim, 1999). In *stochastic matching* (Sankar and Lee, 1996), the entire set of HMMs is used to perform feature compensation and solved for the recognized sentence.

Although model-based feature compensation is effective in some situations, there are many types of distortions that cannot easily be realized by a simple feature transformation. Sometimes the exact distribution of the transformed feature vectors cannot be derived in a useful form for decoding, i.e. a numerical procedure might be required. Model compensation provides an attractive alternative. For example, if the feature bias is time varying, i.e. $x_t = y_t + b_t$ with $b_t$ being a *stochastic bias* (Sankar and Lee, 1996), then the feature compensation vector cannot be computed exactly. If $b_t$ is a random vector with mean vector, $\mu_b$, and covariance matrix, $\Sigma_b$, and is independent of the speech features $y_t$, then it is equivalent to estimating the bias density parameters by the following model transformations: $\mu_x = \mu_y + \mu_b$ and $\Sigma_x = \Sigma_y + \Sigma_b$. The *nuisance* parameters, $\mu_b$ and $\Sigma_b$, are estimated with the EM algorithm under either an ML criterion (Sankar and Lee, 1996) or under a MAP criterion (Chien et al., 1997). Other structures can also be employed to reduce the number of parameters while improving compensation efficiency and effectiveness (Lee, 1998).

### 2.3. Robust decision strategies

As we discussed in Section 1, the plug-in MAP decoder minimizes the recognition error only if the form of the distributions of the data to be recognized and the corresponding parameters are known exactly. The above adaptation and compensation strategies improve the robustness of speech recognition systems by making the distribution $p_{\hat{\Lambda}}(X|W)$ reflect more faithfully the true distribution of $p(X|W)$ for utterance $X$ to be recognized, while keeping the plug-in MAP classification and decision rules intact. Another possibility to improve the robustness of an ASR

system is to modify the plug-in MAP decoder. This area has not attracted much research attention partly because the dynamic programming (DP)-based search strategies are by far the most efficient implementation for solving speech recognition solutions. Any modification of the prevailing DP search algorithm requires a considerable amount of work. However, there exist robust decision rules that can be implemented without changing too much of the existing DP algorithms. The robust decision rule provides a new tool for studying a class of robust speech recognition problems 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.

Before we go further, let us formally define what we mean by a decision rule.

Let us assume our ASR problem is to classify a speech observation $X$ into one of $M$ classes, $W \in \Omega_W$, where $\Omega_W = \{W_1, W_2, \ldots, W_M\}$ denotes the set of $M$ classes. Let us assume that $X$ belongs to a suitable signal space, $\Omega_x$. The ASR problem is, in principle, equivalent to finding a *decision rule*, $d(\cdot)$, in a set of possible decision rules $\mathscr{D}$, such that $d : \Omega_x \to \Omega_W$, or simply

$$W = d(X) \quad \text{for } X \in \Omega_x, \ W \in \Omega_W \quad \text{and} \quad d(\cdot) \in \mathscr{D}, \tag{3}$$

with $W$ being one of the $M$ possible class labels in $\Omega_W$. In this case, the *decision space*, $\{d(X) : X \in \Omega_x\}$, of the decision rule $d(\cdot)$ is the same as the $\Omega_W$. A decision rule $d(\cdot) \in \mathscr{D}$ implies a mapping from the sample space to the class label space. This mapping is known as a *nonrandomized decision rule* (Kharin, 1996). Define $\Omega_x(W_i) = \{X : X \in \Omega_x, d(X) = W_i\}$ to be a subset of $\Omega_x$ corresponding to the region of $X$ being mapped as class $W_i$ with the decision rule $d(\cdot)$, then the construction of a decision rule amounts to finding a partition, $\Omega_x(d(\cdot)) = \{\Omega_x(W_1), \Omega_x(W_2), \ldots, \Omega_x(W_M)\}$, of the observation space $\Omega_x$ under the following constraints:

$$\bigcup_{i=1}^{M} \Omega_x(W_i) = \Omega_x, \quad \Omega_x(W_i) \bigcap \Omega_x(W_j) = \emptyset \quad \text{for } i \neq j; \ i, j = 1, 2, \ldots, M. \tag{4}$$

There may exist an infinite set of decision rules for the same given classification problem. Not all of them are of equal value in practice though. To determine whether a decision rule is "good" one has to agree on a reasonable set of criteria for assessing the "goodness". Intuitively speaking, a decision strategy (rule) is called robust if it is not very sensitive to the previously discussed prior uncertainty (or distortions). In the following, let us formally define what we mean by a robust decision rule.

Let $d(\cdot) = d(X; \mathscr{X})$ be an arbitrary decision rule constructed under some hypothetical model $\mathscr{M}_0$, where $d(X) \in \Omega_W$ is the class to which the observation $X \in \Omega_x$ will be assigned; $\mathscr{X}$ is a training sample set used for the construction of the decision rule. Let $\mathscr{M}_\epsilon$ denote an arbitrary admissible distorted data model for the distortion types discussed in Section 1, where $\epsilon \geqslant 0$ is used to characterize the distortion level. Let $\mathscr{M}^*_\epsilon$ denote the set of admissible distorted data models. The classification performance of the decision rule $d(\cdot)$ in a situation where data are fitted to the distorted model $\mathscr{M}_\epsilon \in \mathscr{M}^*_\epsilon$ will be characterized by the risk functional:

$$r_\epsilon(d(\cdot)) = \boldsymbol{E}[\ell(W, d(X))],$$

where $\boldsymbol{E}[\cdot]$ denotes the expectation with respect to the probability distribution of $(W, X)$ corresponding to the distorted model $\mathscr{M}_\epsilon \in \mathscr{M}^*_\epsilon$; $\ell(W, d(X))$ is a *loss function* associated with making a decision, $d(X)$, if the true class is $W$. One would like the loss function to have the following property:

$$0 \leqslant \ell(W, W) \leqslant \ell(W, d(X) \neq W). \tag{5}$$

Let us call the functional

$$r_+ = r_+(d(\cdot)) = \sup_{\mathscr{M}_\epsilon \in \mathscr{M}^*_\epsilon} r_\epsilon(d(\cdot))$$

the *guaranteed (upper) risk* (Kharin, 1996) for the decision rule $d(\cdot)$ in the presence of distortions

$\mathscr{M}_\epsilon \in \mathscr{M}_\epsilon^*$. If we know the distribution of $\mathscr{M}_\epsilon$ on $\mathscr{M}_\epsilon^*$, we can further define the following functional:

$$\tilde{r} = \tilde{r}(d(\cdot)) = \boldsymbol{E}[r_\epsilon(d(\cdot))],$$

where $\boldsymbol{E}[\cdot]$ denotes the expectation with respect to the distribution of $\mathscr{M}_\epsilon$ on $\mathscr{M}_\epsilon^*$. We call $\tilde{r}(d(\cdot))$ the *overall risk*. Apparently, both $r_+(d(\cdot))$ and $\tilde{r}(d(\cdot))$ can be used as optimality criteria in searching for *robust* (*with respect to distortions* $\mathscr{M}_\epsilon^*$) *decision rules*. A decision rule $d^*(X;\mathscr{X})$ with the minimal value of the guaranteed risk for all admissible distortions,

$$d^*(\cdot) = \arg\min_{d(\cdot)} r_+(d(\cdot)), \tag{6}$$

is referred to as a *minimax decision rule*. A decision rule $\tilde{d}(X;\mathscr{X})$ with the minimal value of the overall risk for all admissible distortions,

$$\tilde{d}(\cdot) = \arg\min_{d(\cdot)} \tilde{r}(d(\cdot)), \tag{7}$$

is referred to as a *predictive decision rule*.

The construction of these robust decision rules will depend on how the admissible distortions $\mathscr{M}_\epsilon \in \mathscr{M}_\epsilon^*$ are defined, and also, for the case of the predictive decision rule, the distribution of the distortion on $\mathscr{M}_\epsilon^*$. In the following two sections, we show two examples of such robust decision rules, namely *minimax decision rule* and *Bayesian predictive decision rule*, respectively. Both of them assume that

- the distributions $p(X|W)$ and $P(W)$ are known up to some specifiable parameters in the forms of $p_\Lambda(X|W)$ and $P_\Gamma(W)$;
- the true parameters of these distributions, $\Lambda$ and $\Gamma$, lie in a neighborhood of the estimated (or hypothetical) ones; therefore,
- the prior uncertainty can be modeled by defining an *uncertainty neighborhood* of the model parameters $\Lambda$, $\Gamma$, and/or possibly a distribution of model parameters $p(\Lambda,\Gamma)$ on this uncertainty neighborhood.

With these assumptions, the specific minimax decision rule and predictive decision rule can be constructed accordingly to satisfy some desired robustness properties. To simplify our discussion, we further assume that we do not consider the uncertainty of $P(W)$ and use $P_{\Gamma_0}(W)$ as the language model, with $\Gamma_0$ being the set of language model parameters estimated from the training text data.

## 3. Minimax classification

### 3.1. Basic formulation

Let $\eta_\epsilon(\Lambda_0)$ denote the uncertainty neighborhood of the true model parameters $\Lambda$, i.e., $\Lambda \in \eta_\epsilon(\Lambda_0)$, where $\Lambda_0$ is the set of model parameters estimated from the training data $\mathscr{X}$, and $\epsilon$ can be viewed as a generic parameter to characterize the degree of the distortion. Then, we have

$$\mathscr{M}_\epsilon^* = \{p_\Lambda(X|W) \mid \Lambda \in \eta_\epsilon(\Lambda_0)\},$$

where $\mathscr{M}_\epsilon^*$ is the set of distorted models, and

$$r_+ = r_+(d(\cdot)) = \sup_{\Lambda\in\eta_\epsilon(\Lambda_0)} \sum_{W\in\Omega_W} P_{\Gamma_0}(W) \int_{X\in\Omega_x} \ell(W, d(X)) p_\Lambda(X|W)\, \mathrm{d}X.$$

To construct a minimax decision rule which minimizes the above-guaranteed risk $r_+(d(\cdot))$ is not an easy task. In practice, some more relaxed criteria have to be adopted. One possibility is to use the upper bound of $r_+(d(\cdot))$, which we denote $r_{++}(d(\cdot))$,

$$r_{++} = r_{++}(d(\cdot)) = \sum_{W\in\Omega_W} \int_{X\in\Omega_x} \sup_{\Lambda\in\eta_\epsilon(\Lambda_0)} \ell(W, d(X)) p_\Lambda(X|W) P_{\Gamma_0}(W)\, \mathrm{d}X.$$

Furthermore, assume that we use a (0,1)-loss function

$$\ell(W, d(X)) = \begin{cases} 0 & \text{if } W = d(X) \text{ (correct decision)}, \\ 1 & \text{if } W \neq d(X) \text{ (wrong decision)} \end{cases} \tag{8}$$

for $W \in \Omega_W$, $d(X) \in \Omega_W$. Then, we have

$$r_{++} = r_{++}(d(\cdot)) = \sum_{W\in\Omega_W} P_{\Gamma_0}(W) \int_{X\notin\Omega_x(W)} \sup_{\Lambda\in\eta_\epsilon(\Lambda_0)} p_\Lambda(X|W)\, \mathrm{d}X. \tag{9}$$

A decision rule which minimizes the above $r_{++}(d(\cdot))$ is as follows:

$$\hat{W} = d_{++}(X) = \arg\max_{W}\left[P_{\Gamma_0}(W)\max_{\Lambda\in\eta_\epsilon(\Lambda_0)} p_\Lambda(X|W)\right]. \tag{10}$$

This is the so-called *minimax decision rule* which was first studied by Merhav and Lee (1993). It can be solved in two steps as illustrated in Fig. 2. First, we estimate the underlying parameters using the ML approach within each neighborhood $\eta_\epsilon(\Lambda_0^{(W)})$, i.e.

$$\hat{\Lambda}_W = \arg\max_{\Lambda\in\eta_\epsilon(\Lambda_0^{(W)})} p_\Lambda(X|W), \tag{11}$$

where $\Lambda_0^{(W)}$ denotes pre-trained model parameters for word $W$. Then, we apply the plug-in MAP decision rule with $\hat{\Lambda}_W$ replacing the original $\Lambda_0^{(W)}$. Therefore, conceptually, the minimax decision rule described in Eq. (10) can be viewed as a procedure which modifies the plug-in MAP decoder shown in Eq. (2) with an extra step as in Eq. (11) to find a modified point estimate in the neighborhood $\eta_\epsilon(\Lambda_0) = \{\eta_\epsilon(\Lambda_0^{(W)})\}$ of the original classifier parameters $\Lambda_0 = \{\Lambda_0^{(W)}\}$.

The above robust minimax classification rule makes no assumption about the form of the distortion. However, its efficacy does depend on an appropriate specification of the parameter uncertainty neighborhood $\eta_\epsilon(\Lambda_0) = \{\eta_\epsilon(\Lambda_0^{(W)})\}$. In the past several years, some other specific techniques have also been developed to implement the above *minimax decision rule* in HMM-based ASR systems (e.g., Moon and Hwang, 1997; Huo et al., 1997; Jiang et al., 1998). They are shown to be effective in dealing with noisy speech recognition and the mismatch caused by different recording conditions.

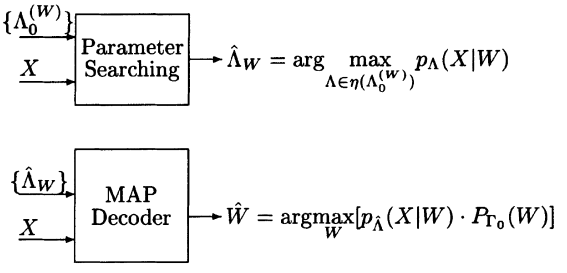


Fig. 2. Two-step minimax classification rule.

### 3.2. Extension to continuous speech recognition

So far we have only dealt with situations that each class $W$ is modeled by a whole word HMM. However, if a sequence of smaller units is used to characterize each word, such as in continuous speech recognition (CSR), the two-step minimax rule shown in Fig. 2 cannot be applied directly due to the increase in search complexity. We discuss two possible ways here to address this problem.

The first approach is to use the *N-best search* paradigm to approximate the minimax rule in three steps. First, the original set of models, $\Lambda_0 = \{\Lambda_0^{(W)}\}$, is used to construct $N$-best candidates for the testing utterance, $X$. Then, the parameter search algorithm shown in the top half of Fig. 2 is applied to each segmented unit of each candidate string on the $N$-best list in the neighborhood $\eta_\epsilon(\Lambda_0^{(W)})$ of each unit model to find a modified model, $\hat{\Lambda}_W$. Finally the new set of models, $\{\hat{\Lambda}_W\}$, is plugged in to rescore all the segmented units in all candidates. These candidates are then re-ordered and the string with the best modified likelihood is the recognized result. These three steps can also be repeated to find a new set of $N$-best candidates and a re-ordered list iteratively.

The second approach is to use a *lattice search* paradigm. First, the original set of models, $\Lambda_0$, is used to construct a word lattice. Then, a recursive minimax search algorithm described in (Jiang et al., 1998) can be used to find the recognition result in the word lattice. For some small vocabulary CSR problems, a direct recursive minimax search on the full recognition network is feasible. Such an example applied to robust connected digit recognition was shown to improve recognition performance in noise. Extension of minimax classification to subword-based, large vocabulary CSR has yet to be studied.

### 3.3. Minimax classification and stochastic matching

There are also other possibilities to model the admissible distortions $\mathscr{M}_\epsilon^*$. For example, if we use

$$\mathscr{M}_\epsilon^* = \{p_\Lambda(X|W) \mid \Lambda = \mathscr{T}_\vartheta(\Lambda_0)\},$$

where $\mathscr{T}_\vartheta(\Lambda_0)$ denotes a specific transformation of $\Lambda_0$ with parameters $\vartheta$. In this way, the uncertainty

of $\Lambda$ can be characterized by the uncertainty of $\vartheta$. Then the minimax decision rule with respect to the above $\mathscr{M}_\epsilon^*$ will be

$$\hat{W} = \arg\max_W \left[ P_{\Gamma_0}(W) \max_\vartheta p(X|W, \Lambda = \mathscr{T}_\vartheta(\Lambda_0)) \right].$$

The so-called *model-space stochastic matching* method described in (Sankar and Lee, 1996; Surendran et al., 1999) can be theoretically justified in this way.

## 4. Bayesian predictive classification

As we discussed before, minimax classification tries to handle the worst case mismatch by assuming a uniform distribution in the uncertainty neighborhood for all possible deviations from the nominal parameters, $\Lambda_0$. Instead of assigning another point estimate, $\hat{\Lambda}$, as done in the minimax classification rule shown in Fig. 2, one can also *average out* the effect of the possible modeling and estimation errors by assuming a general prior pdf for $\Lambda$ to characterize the parameter variability while making classification decisions. In this way, a new robust decision strategy can be derived and is often referred to as a Bayesian predictive classification (BPC) rule (e.g., Nadas, 1985; Huo et al., 1997). We will discuss BPC in the following.

### 4.1. BPC formulation for robust ASR

The principle behind the BPC approach is quite straightforward. As we assume no knowledge about the possible distortions, we rely on a quite general prior pdf to characterize the variability of the HMM parameters caused by the possible mismatches and errors in modeling and estimation. Let us consider the uncertainty of the model parameters $\Lambda$ by treating them as if they were random. Our *prior uncertainty* about $\Lambda$ is then assumed to be summarized in a known joint a priori density $p(\Lambda|\varphi)$, with $\Lambda \in \Omega_\Lambda$, where $\Omega_\Lambda$ denotes an admissible region of $\Lambda$, and $\varphi$ is the set of parameters of the prior pdf (often referred to as *hyperparameters*). In this way, we are essentially considering the following admissible distorted set of data model $\mathscr{M}_\epsilon^*$:

$$\mathscr{M}_\epsilon^* = \{ p_\Lambda(X|W) \mid \Lambda \sim p(\Lambda|\varphi); \Lambda \in \Omega_\Lambda \},$$

where we can view the $\epsilon$ as a parameter to characterize the broadness of the distribution $p(\Lambda|\varphi)$, or equivalently, the degree of the distortion. Based on the above $\mathscr{M}_\epsilon^*$, the *overall risk* is $\tilde{r}(d(\cdot))$,

$$\begin{aligned}
\tilde{r}(d(\cdot)) &= \boldsymbol{E}_{W,X}\boldsymbol{E}_\Lambda[\ell(W, d(X))] \\
&= \sum_{W\in\Omega_W} P_{\Gamma_0}(W) \int_{X\in\Omega_x} \int_{\Omega_\Lambda} \ell(W, d(X)) \\
&\qquad \times p(X|W, \Lambda) p(\Lambda|\varphi)\, \mathrm{d}\Lambda\, \mathrm{d}X \\
&= \sum_{W\in\Omega_W} P_{\Gamma_0}(W) \int_{X\in\Omega_x} \ell(W, d(X)) \\
&\qquad \times \left[ \int_{\Omega_\Lambda} p(X|W, \Lambda) p(\Lambda|\varphi)\, \mathrm{d}\Lambda \right] \mathrm{d}X \\
&= \sum_{W\in\Omega_W} P_{\Gamma_0}(W) \int_{X\in\Omega_x} \ell(W, d(X)) \tilde{p}(X|W)\, \mathrm{d}X,
\end{aligned}$$

where

$$\tilde{p}(X|W) = \int p(X|\Lambda, W) p(\Lambda|\varphi)\, \mathrm{d}\Lambda \tag{12}$$

is called the *predictive pdf* (e.g., Aitchison and Dunsmore, 1975; Geisser, 1993; Ripley, 1996) of the observation $X$ given the word $W$. Then, under the (0,1)-loss function, the predictive decision rule which minimizes the above $\tilde{r}(d(\cdot))$ is as follows:

$$\begin{aligned}
\hat{W} &= \tilde{d}(X) = \arg\max_W \tilde{p}(W|X) \\
&= \arg\max_W \tilde{p}(X|W) \cdot P_{\Gamma_0}(W).
\end{aligned} \tag{13}$$

This decision rule $\tilde{d}(\cdot)$ will be referred to as the BPC rule. 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. Three key issues thus arise in BPC, namely,

- the definition of the prior density $p(\Lambda|\varphi)$ for modeling the uncertainty of the HMM parameters;
- the specification of the hyperparameters, $\varphi$;
- the evaluation of the predictive density.

They are briefly discussed in the following.

### 4.2. Prior specification

The first two issues are related to prior specification. Generally speaking, prior density estimation and the choice of density parameters depend on the particular application of interest. The efficacy of the BPC approach largely depends on the appropriateness of the prior pdf for the distortions we are compensating. If the prior pdf fails to cover the variability reflected in the model parameters, then BPC will not help very much. Therefore, the priors should be carefully specified to make BPC work for robust speech recognition. As in most applications we will assume a specific parametric form for the prior pdf, this problem turns out to be a hyperparameter specification/estimation problem. Many techniques developed in Bayesian adaptation can be used here too. If the training data set $\mathscr{X}$ is rich enough to cover the variability of interest of speech signal which might possibly occur in the testing conditions, then some automatic estimation techniques such as the *method of moments* algorithm presented in (Huo et al., 1995) may be used to automatically estimate the hyperparameters from the training data $\mathscr{X}$. Otherwise, we have to use some ad hoc method for hyperparameter estimation. Readers are referred to (Huo and Lee, 1997a; Huo et al., 1997) for some examples. If the application scenario allows us to have access to some testing data, then by using the *prior evolution* method in (Huo and Lee, 1997a, 1998a; Huo and Ma, 1999; Jiang et al., 1999b), we can obtain an increasingly improved prior pdf (i.e., more and more accurate knowledge about the uncertainty of the model parameters). By using this improved prior pdf, the BPC-based recognition system can approach the performance achieved by the plug-in MAP rule under matched conditions (e.g., Huo and Lee, 1997b; Jiang et al., 1999b). Furthermore, if some knowledge on how the speech signal is distorted and/or varied in different acoustic conditions is available, this will guide us to design a better prior pdf and/or develop a better hyperparameter estimation method (Huo and Lee, 1998b, 2000). A promising future research direction is to develop new methods for deriving efficiently and effectively an appropriate prior pdf for BPC on-the-fly for individual testing utterance by using both the pre-prepared information and the information embedded in the testing utterance itself.

### 4.3. Approximate BPC approach

As for the third issue, due to the nature of the *missing data* problem in the HMM formulation, it is not easy to compute the following true predictive pdf:

$$\begin{aligned}\tilde{p}(X|W) &= \int p(X|\Lambda, W)p(\Lambda|\varphi)\,\mathrm{d}\Lambda \\ &= \sum_{\boldsymbol{s},\boldsymbol{l}} \int p(X, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W)p(\Lambda|\varphi)\,\mathrm{d}\Lambda, \end{aligned} \quad (14)$$

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_T)$ is the unobserved state sequence and $\boldsymbol{l} = (l_1, l_2, \ldots, l_T)$ is the associated sequence of the unobserved mixture component labels corresponding to the observation sequence $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ (we are using Gaussian-mixture CDHMM as an example). Consequently, some approximations are needed.

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

$$\tilde{p}(X|W) \approx \max_{\boldsymbol{s},\boldsymbol{l}} \int p(X, \boldsymbol{s}, \boldsymbol{l}|\Lambda, W)p(\Lambda|\varphi)\,\mathrm{d}\Lambda. \quad (15)$$

A detailed algorithm to implement the above approximation and the related experimental results are reported in (Jiang et al., 1999a).

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

$$\begin{aligned}\tilde{p}(X|W) \approx\ & p(X|\Lambda_{\mathrm{MAP}}, W) \cdot p(\Lambda_{\mathrm{MAP}}|\varphi, W) \\ & \cdot (2\pi)^{\mathscr{L}/2} \cdot |V|^{1/2}, \end{aligned} \quad (16)$$

where $\Lambda_{\mathrm{MAP}}$ is the following MAP estimate

$$\Lambda_{\mathrm{MAP}} = \arg\max_{\Lambda \in \Omega_\Lambda} p(X|\Lambda, W)p(\Lambda|\varphi), \quad (17)$$

$\mathscr{L}$ is the number of HMM parameters involved in the integrand in Eq. (12), and $V$ is the $\mathscr{L} \times \mathscr{L}$ modal dispersion matrix, i.e., $-V^{-1}$ is the Hessian matrix of second derivatives of

$$\hbar(\Lambda) = \log\{p(X|\Lambda, W)p(\Lambda|\varphi)\} \tag{18}$$

evaluated at $\Lambda = \Lambda_{\mathrm{MAP}}$. This is essentially equivalent to using a normal pdf $\mathcal{N}(\Lambda|\Lambda_{\mathrm{MAP}}, V)$ to approximate the posterior pdf $p(\Lambda|X, W)$. In the case of CDHMM, to compute $V$ directly is still too computationally involved. Therefore, we have to make further approximations. If we only consider the uncertainty of the mean vectors in CDHMM for BPC decoding, we can use the Quasi-Bayes (QB) algorithm in (Huo and Lee, 1997a, 1998a) to compute an approximate posterior pdf $\mathcal{N}(\Lambda; \Lambda_{\mathrm{MAP}}, \tilde{U})$ and then replace $V$ in Eq. (16) with $\tilde{U}$. The resulting BPC rule is thus named as the Quasi-Bayes Predictive Classification (QBPC) rule (Huo et al., 1997; Huo and Lee, 2000).

Both QBPC and VBPC methods have been shown to enhance robustness when mismatches exist between training and testing conditions (e.g., Huo et al., 1997; Huo and Lee, 1998b, 2000; Jiang et al., 1999a,b).

### 4.4. BPC based on structural parameters

As we discussed in compensation techniques, for 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. As shown in Section 3.3, the stochastic matching approach described in (Sankar and Lee, 1996) is such a natural extension of structure-based compensation from the minimax approach. The same kind of extension can also be applied to the BPC approach. For example, we can use

$$\mathscr{M}^*_{\epsilon} = \{p_{\Lambda}(X|W) \mid \Lambda = \mathscr{T}_{\vartheta}(\Lambda_0), \vartheta \sim p(\vartheta)\},$$

where $\mathscr{T}_{\vartheta}(\Lambda_0)$ denotes a specific transformation of $\Lambda_0$ with parameters $\vartheta$. In this way, the uncertainty of $\Lambda$ can be characterized by the uncertainty of $\vartheta$. The BPC decision rule in (13) can then be modified using the following predictive pdf:

$$\tilde{p}(X|W) = \int_{\Omega_{\vartheta}} p(X|W, \Lambda = \mathscr{T}_{\vartheta}(\Lambda_0))p(\vartheta)\,\mathrm{d}\vartheta.$$

The above issue of prior specification will then be translated into the specification of $p(\vartheta)$.

## 5. Other related robust decision approaches

Historically, the predictive approach receives little attention in many classical statistics literature despite the existence of many good books (e.g. Aitchison and Dunsmore, 1975; Geisser, 1993; Ripley, 1996). To our knowledge, it was Nadas who first adopted a BPC formulation and pointed out its potential in speech recognition applications (Nadas, 1985). Nadas was suggesting to use the posterior pdf $p(\Lambda|\mathscr{X})$ derived from the *training set* $\mathscr{X}$ directly to serve as the prior pdf in predictive decision making. However, in this case, BPC will make little difference from the conventional plug-in MAP rule in many applications. This is because whatever initial prior pdf is used, when a large amount of training data $\mathscr{X}$ are available, we will get a posterior pdf $p(\Lambda|\mathscr{X})$ with a sharp peak. This makes the predictive pdf in Eq. (12) of little difference from $p_{\hat{\Lambda}}(X|W)$ with the ML estimate $\hat{\Lambda}$. In the limit, if all the posterior probability mass is concentrated at the ML estimates $\hat{\Lambda}$ obtained from $\mathscr{X}$, it is easy to see from Eqs. (13) and (12) that the BPC decision rule coincides with the plug-in MAP decision rule. In (Nadas, 1985), a simple example is given in which a *reproducing density* exists such that a closed-form solution can be derived for calculating the predictive pdf. No experimental results were reported and the paper ended by briefly discussing the difficulty of applying the theory to HMM-based speech recognition.

### 5.1. Use of training set in decision

Starting from Nadas's formulation, Merhav and Ephraim (1991) suggested a so-called approximate Bayesian (AB) decision rule 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:

$$\hat{W} = \arg\max_{W} \frac{\max_{\Lambda}[p(X|\Lambda, W) \cdot p(\mathscr{X}|\Lambda, W)]}{\max_{\Lambda} p(\mathscr{X}|\Lambda, W)} P_{\Gamma_0}(W). \tag{19}$$

It is clear that if the training sequences $\mathscr{X}$ are considerably longer than the test sequence $X$ which is the case in most speech recognition applications, the parameter set $\Lambda$ that maximizes the denominator of Eq. (19) is very close to the parameter set that maximizes the numerator; hence, the factor $p(\mathscr{X}|\Lambda, W)$ in both the numerator and denominator is essentially canceled. This makes the AB decision rule of little difference from the plug-in MAP decision rule using an ML estimate of $\Lambda$. The AB decision rule is also computationally expensive because the maximization of $[p(X|\Lambda, W) \cdot p(\mathscr{X}|\Lambda, W)]$ over $\Lambda$ must be performed for every test sequence $X$. Furthermore, all the training data must be stored. All these factors make the AB decision rule impractical for most speech recognition applications.

### 5.2. Bayesian minimax rule

As discussed previously, the minimax classification rule can be viewed as a two-step procedure and implemented in Eq. (10). First, each testing utterance is treated as possibly belonging 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. the MAP approach, in the first step. Such a modified minimax decision rule works as follows:

$$\hat{W} = \arg\max_{W}[p(X|\Lambda_{\mathrm{MAP}}, W) \cdot P_{\Gamma_0}(W)] \tag{20}$$

where $\Lambda_{\mathrm{MAP}}$ is the MAP estimate as shown in Eq. (17). For the convenience of reference, we call this modified minimax decision rule as a *Bayesian minimax rule* to emphasize its difference from the original minimax approach in (Merhav and Lee, 1993). The readers are referred to Merhav and Lee (1993), Huo et al. (1997) and Jiang et al. (1998, 1999a) for a performance comparison of different implementations of the minimax decision rule on several speech recognition tasks.

### 5.3. Model compensation based on component predictive density

We have previously discussed the BPC approach as a new decision rule which averages out the sampling error in HMM parameter estimation. A related but simpler approach can also be used. Instead of directly modifying the basic decision rule, one can also assume that the CDHMM parameters are uncertain. Then, one uses the *Bayesian predictive density* of each Gaussian mixture component to serve as the compensated distribution of that component and plug these compensated distributions into the MAP decision rule in Eq. (1). We thus call the approach *Bayesian predictive density-based model compensation* method, or shortly BP-MC method, to differentiate it from the BPC rule defined in Eq. (13). In (Shahshahani, 1997), such an idea is explored in the context of Bayesian speaker adaptation where a Gaussian prior pdf for mean vector is adopted. In (Jiang et al., 1999a), a similar idea is applied to noisy speech recognition where a uniform prior pdf on a pre-specified uncertainty neighborhood for mean vector is adopted. More recently, Surendran and Lee (1998, 1999) apply a similar idea to the transformation-based model compensation by using the predictive pdf of the transformation parameters. The prior pdf of the transformation parameters is specified under the guidance of certain information embedded in the testing utterance. Such an approach was shown to be robust to speaker and channel differences (Surendran and Lee, 1998, 1999).

## 6. Illustrative experimental results

So far, we have presented an overview of some recent techniques on the topics of adaptation, compensation, and robust decision. In this section, we present some experimental results on robust decision rules we discussed before to give readers a rough idea of how these techniques work in robust speech recognition.

### 6.1. Experimental setup

Two isolated word recognition tasks are chosen to address the issue of the vocabulary confusability. The first one is the recognition of 26 English letters which are highly confusable and their discrimination is weak even without mismatch. Two severely mismatched databases namely the OGI ISOLET and TIalpha (alphabet subset of TI46 corpus) corpora are used (e.g., Huo et al., 1995; Huo and Lee, 1997a, 1998a). The second task is the recognition of 20 less confusable English words which include 10 digits and 10 commands namely enter, erase, go, help, no, rubout, repeat, stop, start, yes. The 20-word subset (TI20) of the TI46 corpus is used.

Throughout the following experiments, each word, $W$, is modeled by a left-to-right $N$-state whole word CDHMM, $\lambda_W$, with arbitrary state skipping. For each state $i$, the state observation pdf is assumed to be a mixture of multivariate Gaussian pdf's,

$$p(\boldsymbol{x}|s_t = i) = \sum_{k=1}^{K} \omega_{ik} \mathcal{N}(\boldsymbol{x}|m_{ik}, \Sigma_{ik}),$$

where the set of mixture coefficients $\{\omega_{ik}\}$ satisfy the constraint $\sum_{k=1}^{K} \omega_{ik} = 1$, and $\mathcal{N}(\boldsymbol{x}|m_{ik}, \Sigma_{ik})$ is the $k$th normal mixture component with $m_{ik}$ being the $D$-dimensional mean vector and $\Sigma_{ik}$ being the $D \times D$ covariance matrix with its $d$th diagonal element being $\sigma^2_{ikd}$. Here, we use five states per CDHMM (i.e., $N = 5$) and four Gaussian mixture components per state (i.e. $K = 4$) with each component having a diagonal covariance matrix.

The speech data in all the corpora are downsampled to 8 kHz. Each feature vector used in this study consists of 12 (i.e. $D = 12$) bandpass-liftered LPC-derived cepstral coefficients with a 30 ms frame length and a 10 ms frame shift (e.g., Lee et al., 1990). Utterance-based cepstral mean normalization is applied for acoustic normalization both in training and testing. In the plug-in MAP recognition, the decision rule determines the recognized word as the one which attains the highest forward–backward probability.

In the following subsections, a series of experiments are designed to examine (i) the algorithmic characteristics of the QBPC and Bayesian minimax approaches in Section 6.3; (ii) the effect of the class confusability in Section 6.4; (iii) the effect of the prior evolution in Section 6.5. Throughout the experiments, two types of mismatches, namely the general cross-condition mismatch and cross-gender mismatch are studied, respectively. Comparisons between the QBPC and Bayesian minimax approaches are also made. Before that, let us provide some details on how to specify the prior pdf.

### 6.2. Prior specification

The prior pdf of the means for each word CDHMM is assumed to have a Gaussian pdf $\mathcal{N}(\{m_{ikd}\}|\boldsymbol{\mu}, U)$,

$$p(\{m_{ikd}\}|W) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}u_{ikd}} \exp\left[-\frac{(m_{ikd}-\mu_{ikd})^2}{2u^2_{ikd}}\right], \quad (21)$$

with a collection of the related mean vectors denoted as $\boldsymbol{\mu} = \text{vec}\{\mu_{ikd}\}$ and a diagonal covariance matrix denoted as $U = \text{diag}\{u^2_{ikd}\}$. To facilitate the following discussions, we define $\tau_{ikd} = \sigma^2_{ikd}/u^2_{ikd}$. The related hyperparameters are derived in the last iteration of seed CDHMM's training as follows:

$$\mu_{ikd} = m_{ikd}, \quad (22)$$

$$\tau_{ikd} = \epsilon_1 \cdot \sum_t \xi_t(i,k), \quad (23)$$

where $\xi_t(i,k) = \Pr(s_t = i, l_t = k|\mathcal{X}, \Lambda_0)$ is the probability of observing $\boldsymbol{x}_t$ in mixture component $k$ of state $i$ (given training data $\mathcal{X}$ and trained HMM parameters $\Lambda_0$); and $\epsilon_1 > 0$ is a weighting coefficient to control the *degree* of the uncertainty of the prior distribution. In this study, the weighting coefficient $\epsilon_1$ was chosen to be $1/J$ with $J$ being the number of training tokens corresponding to each HMM. Thus, roughly speaking, the prior distribution contains the same information as would, on average, a single observation contain. This seems to be a reasonable representation of the common situation where there is a little, but not much, prior information. It also makes the contributions from the prior and a

single testing token comparable and thus distinguishes BPC and Bayesian minimax from conventional plug-in MAP decoding. Once the prior pdf's are specified for each CDHMM, the QBPC-based and Bayesian minimax-based speech recognition can be carried out for an unknown utterance $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$. Both of them need to use QB algorithm in e.g. (Huo and Lee, 1997a) to compute an approximate posterior pdf $p(\{m_{ikd}\}|X, W)$, which is also a Gaussian pdf $\mathcal{N}(\{m_{ikd}\}|\tilde{\boldsymbol{\mu}}, \tilde{U})$ with hyperparameters

$$\tilde{\mu}_{ikd} = \frac{\mathrm{rf} \cdot \tau_{ikd} \cdot \mu_{ikd} + c_{ik}\bar{x}_{ikd}}{\mathrm{rf} \cdot \tau_{ikd} + c_{ik}}, \tag{24}$$

$$\tilde{u}^2_{ikd} = \frac{\sigma^2_{ikd}}{\mathrm{rf} \cdot \tau_{ikd} + c_{ik}}, \tag{25}$$

where

$$\xi_t(i,k) = \Pr(s_t = i, l_t = k|X, \lambda, W), \tag{26}$$

$$c_{ik} = \sum_{t=1}^{T} \xi_t(i,k), \tag{27}$$

$$\bar{\boldsymbol{x}}_{ik} = \sum_{t=1}^{T} \xi_t(i,k)\boldsymbol{x}_t / c_{ik}. \tag{28}$$

The above QB procedure is implemented by an iterative EM algorithm. In practice, we observe that several iterations (typically 1–3 iterations) are enough to get a good recognition result. Furthermore, the refreshing coefficient "rf" for the hyperparameters is used to control the degree of the uncertainty of the CDHMM parameters, where $\mathrm{rf} = 1$ means no change, $\mathrm{rf} > 1$ means to decrease the uncertainty of the HMM parameters (i.e., to trust more the current estimate of the HMM parameters), and $\mathrm{rf} < 1$ means to increase the uncertainty of the HMM parameters.

### 6.3. Characteristics of QBPC and Bayesian minimax approaches

The first set of experiments are designed to examine the algorithmic characteristics and the behavior of the QBPC and Bayesian minimax approaches which include

- the efficacy of the QBPC and Bayesian minimax approaches under a simple prior specification for compensating a general cross-condition mismatch;
- the effect of the number of EM iterations;
- the effect of the refreshing coefficient.

The recognition vocabulary is the 26 English letters. We train a set of speaker-independent (SI) CDHMMs as well as their initial prior pdf's by using the OGI ISOLET database which consists of 150 speakers, 75 females and 75 males, each speaking each of the letters twice. The hyperparameters of the prior pdf's are estimated with the method described in previous subsection. Then, we do SI testing on TIalpha which consists of 16 speakers, 8 females and 8 males. To keep consistent with our previous experimental setup in (Huo and Lee, 1997a, 1998a), about 8 tokens per letter for each speaker are used for testing. Due to the strong mismatch between the training and testing databases, we are effectively considering the general mismatch conditions of those in speaker, speaking style, transducer, recording environments and conditions, sampling rate and quantization resolution, etc. Table 1 compares the average recognition accuracy over 8 female speakers using the standard plug-in MAP decision rule to that of the QBPC and the Bayesian minimax methods with different EM iterations and different values of the refreshing coefficient. For the Bayesian minimax method here, in comparison with (Merhav and

Table 1
Performance (word accuracy in %) comparison averaged over 8 female speakers of plug-in MAP, QBPC and Bayesian minimax rules on English letter recognition task: training on ISOLET and testing on TIalpha

| EM iterations | Plug-in MAP | QBPC with different rf | | | | | Bayesian minimax with different rf | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2.0 | 1.5 | 1.0 | 0.75 | 0.5 | 2.0 | 1.5 | 1.0 | 0.75 | 0.5 |
| 1 | 49.1 | 52.9 | 53.2 | 53.7 | 54.1 | 53.3 | 54.2 | 53.9 | 52.0 | 51.8 | 50.9 |
| 2 | N/A | 54.3 | 54.5 | 55.1 | 55.0 | 53.5 | 51.8 | 51.3 | 49.6 | 49.1 | 45.6 |
| 3 | N/A | 54.8 | 55.3 | 56.2 | 55.5 | 53.6 | 51.8 | 50.2 | 48.2 | 46.0 | 44.3 |

Lee, 1993), we are using a more informative parametric form (Gaussian) for the prior instead of a uniform distribution in an uncertainty neighborhood surrounding the ML-trained HMM parameters. The experimental results show that the QBPC is achieving the best performance with around 14% relative recognition error rate reduction over that of the standard plug-in method. With more EM iterations, QBPC achieves a better performance while that of Bayesian minimax degrades. It is also observed that in a reasonably wide range of values of the control parameters (both the number of EM iterations and the refreshing coefficient), both the QBPC and Bayesian minimax methods achieve improvement over the conventional plug-in MAP method.

### 6.4. Effects of class confusability

The second set of experiments are designed to examine the effect of the vocabulary confusability on the performance of the QBPC and Bayesian minimax approaches. The recognition tasks are TIalpha and TI20, respectively. Another type of mismatch, namely the gender difference is examined. We train two sets of gender-dependent models (both CDHMMs and their initial prior pdf's) from 8 female and 8 male speakers in TI46 using about 10 training tokens per word for each speaker. We then perform cross-gender testing (testing on 8 female speakers by using male seed models and vice versa) using about 8 tokens per letter for each speaker in TIalpha case (same as previous experiments) and about 16 tokens per word for each speaker in TI20 case. Tables 2 and 3 compare, on TIalpha and TI20 word recognition tasks, respectively, the average recognition accuracies over 8 female and 8 male speakers of the standard plug-in MAP decision rule to that of the QBPC approach and the Bayesian minimax method under matched (training on one gender and testing on the same gender) and mismatched (cross-gender testing) conditions. One EM iteration is performed for QBPC in matched condition testing and for all of the Bayesian minimax testing cases. Two and three EM iterations are performed for QBPC in the TIalpha and TI20 cases, respectively. The refreshing coefficient rf is set to be 1.0. The first observation from the experimental results is that both the QBPC and Bayesian minimax methods achieve better performance than the conventional plug-in MAP decoding (denoted as "PI-MAP" in the tables) in mismatched condition testing. More improvement is achieved in a less confusable vocabulary, i.e., TI20. The second observation is that the QBPC and Bayesian minimax methods degrade the performance in the matched condition testing cases. Although, in principle, we

Table 2
Performance (word accuracy in %) comparison averaged over 8 female and 8 male speakers of plug-in MAP, QBPC and Bayesian minimax rules on TIalpha English letter recognition task

| Testing speakers | Matched condition | | | Mismatched condition | | |
|---|---|---|---|---|---|---|
| | PI-MAP | QBPC | Bayesian minimax | PI-MAP | QBPC | Bayesian minimax |
| Female | 82.0 | 79.1 | 75.1 | 29.9 | 34.0 | 34.3 |
| Male | 84.1 | 81.7 | 75.2 | 27.4 | 37.9 | 40.2 |

Table 3
Performance (word accuracy in %) comparison averaged over 8 female and 8 male speakers of plug-in MAP, QBPC and Bayesian minimax rules on TI20 word recognition task

| Testing speakers | Matched condition | | | Mismatched condition | | |
|---|---|---|---|---|---|---|
| | PI-MAP | QBPC | Bayesian minimax | PI-MAP | QBPC | Bayesian minimax |
| Female | 98.4 | 97.5 | 95.0 | 45.4 | 54.1 | 50.5 |
| Male | 98.4 | 98.2 | 95.3 | 40.5 | 53.5 | 55.4 |

can make both QBPC and Bayesian minimax methods achieve a similar performance to that of the plug-in MAP decoding in matched condition testing by using a bigger value of the refreshing coefficient rf (i.e., a sharper prior pdf), how to automatically achieve this goal is an interesting open research topic. The third observation is that in spite of the performance improvement by using the QBPC or Bayesian minimax methods in the mismatched condition case, the absolute recognition rate is still far inferior to that of the matched condition testing result. How to bridge this performance gap is still a challenging topic for further research. One possible way to achieve further improvement is to explore better prior pdf's, from either incrementally available testing data or knowledge and experience. We present a case study in the following subsection to show how improved prior pdf with *prior evolution* helps.

### 6.5. Effects of prior evolution

As suggested in Section 4.2, if we have an access to some testing data, the prior pdf can be made more appropriate by using the prior evolution (PE) techniques (e.g., Huo and Lee, 1997a, 1998a; Huo and Ma, 1999). The third set of experiments are thus designed to examine the viability of this technique. We use the QB method described in (Huo and Lee, 1997a) for prior evolution and the QBPC and Bayesian minimax methods discussed before for recognition. The baseline system uses plug-in MAP decision rule and on-line adaptation (OLA) of CDHMM parameters (Huo and Lee, 1997a). The adapted CDHMM parameters are derived as a point estimate (here, MAP) from the evolving prior pdf. Note that in this study, although we evolve the prior pdf's of all CDHMM parameters except for variances in prior evolution, we only consider, for QBPC and Bayesian minimax, the uncertainty of the mean vectors of CDHMMs which is characterized by a set of Gaussian pdf's as discussed before. Also note that all the prior evolution experiments are performed in a supervised mode. The framework used in the experiments for these combined prior evolution and robust decision rule procedures is schematically shown in Fig. 3.

As in Section 6.3, starting from SI seed models and the corresponding initial prior pdf's trained from the ISOLET database, we perform prior evolution for condition and speaker adaptation on the speakers in the TIalpha database. One token per letter is used incrementally for PE. The same 8 tokens per letter for each speaker as in the experiments of Section 6.3 are used for testing after each PE step. Fig. 4 shows the performance comparison averaged over 8 female speakers on the English letter recognition task as a function of the total number of adaptation tokens per speaker among methods by combining on-line adaptation with the plug-in MAP, Bayesian minimax, and QBPC

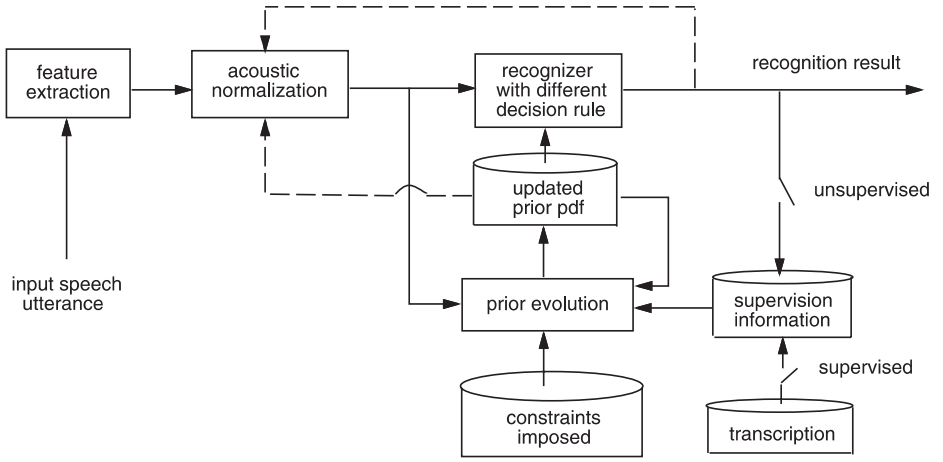


Fig. 3. The framework used in the experiments for the combined prior evolution and robust decision rules.

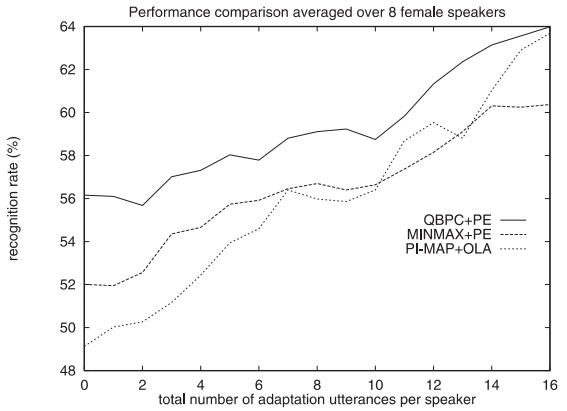


Fig. 4. Performance (word accuracy in %) comparison averaged over 8 female speakers on English letter recognition task as a function of total amount of adaptation data among methods by combining prior evolution with plug-in MAP, Bayesian minimax and QBPC decodings (SI seed models from ISOLET; 3 EM iterations for on-line adaptation, 3 for QBPC and 1 for Bayesian minimax; rf = 1.0).

decision rules. The refreshing coefficient rf is set to be 1.0. One EM iteration is performed for Bayesian minimax and three EM iterations for QBPC. Without any compensation, as expected, the cross-condition SI recognition rate is very low. With OLA and conventional plug-in MAP decoding (denoted as "PI-MAP + OLA"), the performance is continuously improved with an increasing amount of adaptation data. By combining PE with Bayesian minimax (denoted as "MINMAX + PE" for simplicity) and QBPC decoding (denoted as "QBPC + PE"), the performance is further improved before a certain point where a good enough model parameter estimation warrants the plug-in MAP decision to surpass either Bayesian minimax or QBPC. It is also observed that QBPC performs better than Bayesian minimax in general.

Experiments for a similar performance comparison are also performed with the cross-gender on-line speaker adaptation and testing on TIalpha and TI20, respectively. Facts similar to the above observations are also observed. As a future work, it will be interesting to see how the QBPC performs if combined with the PE algorithm for correlated CDHMMs in (Huo and Lee, 1998a) and other more advanced PE algorithms in (Huo and Ma, 1999).

### 6.6. Discussion of experimental results

We observed that the efficacy of the BPC and Bayesian minimax approaches depends on the appropriateness of the prior distribution for the mismatch we are compensating and the confusability of the classes we are comparing. Both approaches improve the performance robustness in mismatched conditions at the expense of decreasing the discriminative ability of the models. More likely, they will perform better in a less confusable classes case simply because we have more chances to use a broader prior pdf to accommodate a higher degree of distortions which is evidenced by our experimental results shown earlier. If the application scenario allows, by combining BPC and Bayesian minimax approaches with prior evolution, we can make the prior distribution more appropriate, thus achieve a better performance. For real-world applications, unsupervised prior evolution is usually more realistic and desirable. One of the remaining research issues is how to guide the unsupervised prior evolution when the recognition rate is initially low.

We have shown elsewhere (Huo and Lee, 1998b, 2000) that knowledge and experience on how the speech signal is varied under mismatched conditions are helpful to give a better hyperparameter estimation which in turn improves the BPC and Bayesian minimax performance. We expect that a better understanding and more experience of the type under different acoustic conditions will also be helpful to design a better parametric form of the prior pdf's and the estimation of corresponding hyperparameters.

As far as the issue of computational complexity is concerned, the QBPC algorithm is relatively simple to implement and there is no big increase in computational complexity when compared with the conventional plug-in MAP decoding. The overhead of the QBPC approach is mainly determined by the number of EM iterations in the quasi-Bayesian approximation of computing the approximate posterior density. In the case of one EM iteration, in comparison with the standard plug-in MAP approach, the increased computation of the QBPC involved in Eqs. (16), (24) and (25) is negligible. In the case of multiple, say $N$ EM

iterations, the decoding speed of the QBPC is approximately $N$ times that of the plug-in MAP decoder. In our experiments, we observed that for the QBPC approach, one EM iteration is usually enough in matched-condition testing, and more iterations (one to three seems enough) help in mismatched-condition testing. When applying the QBPC approach to the continuous ASR problem, it can be operated under an $N$-best hypothesises rescoring mode. As a remark, the Bayesian minimax method has a similar computational complexity to the QBPC approach.

As for the BPC in general, two issues remain to be addressed. One is the question of whether a more accurate approximation method in the BPC procedure to compute the approximate predictive pdf for classification will lead to a better performance. Another concerns the sufficiency of considering only the uncertainty of the mean vectors of CDHMM. More theoretical work is needed if we want to consider the uncertainty of the other parameters in BPC.

## 7. Summary

We have presented an overview of some recent techniques on the topics of adaptation and compensation. We have also collectively discussed in detail two recently proposed classification rules, namely minimax classification and Bayesian predictive classification. Both of them model classifier parameter uncertainty and modify the classification rules to satisfy some desired robustness properties. By combining robust decision strategies with Bayesian adaptation and compensation, many new algorithms can be designed to improve the performance of speech recognition in mismatched and/or adverse conditions. We want to emphasize again that the previously reviewed techniques provide some useful tools. Intelligent use of these flexible tools for different purposes in different applications will be an important part of the future research. We are continuously developing the new techniques in the following research areas:

- *Intelligent and efficient learning techniques* to exploit different knowledge sources from large amount of task- and condition-independent speech data in an efficient way.
- *Information fusion techniques* to combine different knowledge sources (guided by task specifications and the information embedded in task- and condition-dependent speech data), and then to derive a set of models which work well for the target task.
- *Adaptive learning techniques* to continuously improve the performance with the increasing amount of task- and condition-dependent speech data.
- *Robust decision rules* to make the best decision during recognition based on all the available information.

A new theory is under development to unify the above topics together. It is our hope that the previous in-depth discussions may inspire further innovations that will lead to better solutions for robust ASR.

## Acknowledgements

The authors would like to thank the anonymous referees for their suggestions that improved the presentation of the paper. Some of the works by Q. Huo described in this paper was partially supported by a grant from the RGC of the Hong Kong SAR (Project No. HKU7016/97E) and a HKU CRCG research initiation grant.

## References

Acero, A., 1993. Acoustical and environmental robustness in automatic speech recognition. Kluwer Academic Publishers, Dordrecht.

Aitchison, J., Dunsmore, I.R., 1975. Statistical Prediction Analysis. Cambridge University Press, UK.

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55 (6), 1304–1312.

Chesta, C., Siohan, O., Lee, C.-H., 1999. Maximum a posterior linear regression for hidden Markov model adaptation. In: Proc. Eurospeech-99. Budapest, Hungary.

Chien, J.-T., 1999. Online hierarchical transformation of hidden markov models for speech recognition. IEEE Trans. Speech Audio Process. 7 (6), 656–667.

Chien, J.-T., Wang, H.-C., 1997. Telephone speech recognition based on Bayesian adaptation of hidden Markov models. Speech Communication 22, 369–384.

Chien, J.-T., Lee, C.-H., Wang, H.-C., 1997. A hybrid algorithm for speaker adaptation using MAP transformation and adaptation. IEEE Signal Process. Lett. 4 (6), 167–168.

Chou, W., 1999. Maximum a posterior linear regression with elliptically symmetric matrix variate priors. In: Proc. Eurospeech-99. Budapest, Hungary.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39 (1), 1–38.

Diakoloukas, V.D., Digalakis, V.V., 1999. Maximum-likelihood stochastic-transformation adaptation of hidden Markov models. IEEE Trans. Speech Audio Process. 7 (2), 177–187.

Digalakis, V.V., 1999. Online adaptation of hidden Markov models using incremental estimation algorithms. IEEE Trans. Speech Audio Process. 7 (3), 253–261.

Digalakis, V.V., Neumeyer, L.G., 1996. Speaker adaptation using combined transformation and Bayesian methods. IEEE Trans. Speech Audio Process. 4 (4), 294–300.

Digalakis, V.V., Rtischev, D., Neumeyer, L.G., 1995. Speaker adaptation using constrained estimation of Gaussian mixtures. IEEE Trans. Speech Audio Process. 3 (5), 357–366.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Language 12, 75–98.

Gales, M.J.F., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Comput. Speech Language 10, 249–264.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Geisser, S., 1993. Predictive Inference: An Introduction. Chapman & Hall, New York.

Hermansky, H., 1998. Should recognizers have ears?. Speech Communication 25, 3–27.

Hunt, M.J., 1999. Spectral signal processing for ASR. In: Proc. 1999 IEEE Workshop on Automatic Speech Recognition and Understanding. Keystone, CO.

Huo, Q., 1999. An introduction to decision rules for automatic speech recognition. Technical Report TR-99-07, Department of Computer Science and Information Systems, The University of Hong Kong (available at URL: http://www.csis.hku.hk/publications/).

Huo, Q., Lee, C.-H., 1997a. On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate. IEEE Trans. Speech Audio Process. 5 (2), 161–172.

Huo, Q., Lee, C.-H., 1997b. Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition. In: Proc. Eurospeech-97. Rhodes, Greece.

Huo, Q., Lee, C.-H., 1998a. On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition. IEEE Trans. Speech Audio Process. 6 (4), 386–397.

Huo, Q., Lee, C.-H., 1998b. A study of prior sensitivity for Bayesian predictive classification based robust speech recognition. In: Proc. ICASSP-98. Seattle, pp. 741–744.

Huo, Q., Lee, C.-H., 2000. A Bayesian predictive classification approach to robust speech recognition. IEEE Trans. Speech Audio Process. 8 (2), 200–204.

Huo, Q., Ma, B., 1999. On-line adaptive learning of continuous density hidden Markov models based on multiple-stream prior evolution and posterior pooling. Submitted to IEEE Trans. Speech Audio Process. See also a condensed version in: Proc. Eurospeech-99. Budapest, Hungary, pp. 2721–2724.

Huo, Q., Chan, C., Lee, C.-H., 1995. Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition. IEEE Trans. Speech Audio Process. 3 (5), 334–345.

Huo, Q., Jiang, H., Lee, C.-H., 1997. A Bayesian predictive classification approach to robust speech recognition. In: Proc. ICASSP-97. Munich, pp. 1547–1550.

Jelinek, F., Mercer, R.L., Roukos, S., 1991. Principles of lexical language modeling for speech recognition. In: Furui, S., Sondhi, M.M., (Eds.), Advances in Speech Signal Processing. Marcel Dekker, New York, pp. 651–699.

Jiang, H., Hirose, K., Huo, Q., 1998. A minimax search algorithm for CDHMM based robust continuous speech recognition. In: Proc. ICSLP-98. Sydney, pp. II-389–392.

Jiang, H., Hirose, K., Huo, Q., 1999a. Robust speech recognition based on a Bayesian prediction approach. IEEE Trans. Speech Audio Process. 7 (4), 426–440.

Jiang, H., Hirose, K., Huo, Q., 1999b. Improving Viterbi Bayesian predictive classification via sequential Bayesian learning in robust speech recognition. Speech Communication 28 (4), 313–326.

Kharin, Y., 1996. Robustness in Statistical Pattern Recognition. Kluwer Academic Publishers, Boston.

Lasry, M.J., Stern, R.M., 1984. A posteriori estimation of correlated jointly Gaussian mean vectors. IEEE Trans. Pattern Anal. Machine Intell. 6 (4), 530–535.

Lawrence, C., Rahim, M., 1999. Integrated bias removal techniques for robust speech recognition. Comput. Speech Language 13, 283–298.

Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication 25 (1–3), 29–47.

Lee, C.-H., Rabiner, L.R., Pieraccini, R., Wilpon, J.G., 1990. Acoustic modeling for large vocabulary speech recognition. Comput. Speech Language 4, 127–165.

Lee, C.-H., Lin, C.-H., Juang, B.-H., 1991. A study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Trans. Signal Process. 39 (4), 806–814.

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

Leggetter, C.J., Woodland, P.C., 1995a. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Language 9, 171–185.

Leggetter, C.J., Woodland, P.C., 1995b. Flexible speaker adaptation for large vocabulary speech recognition. In: Proc. Eurospeech-95. Madrid, Spain, pp. 1155–1158.

Merhav, N., Ephraim, Y., 1991. A Bayesian classification approach with application to speech recognition. IEEE Trans. Signal Process. 39 (10), 2157–2166.

Merhav, N., Lee, C.-H., 1993. A minimax classification approach with application to robust speech recognition. IEEE Trans. Speech Audio Process. 1 (1), 90–100.

Mokbel, C., Delphin-Poulat, L., 1999. A unified framework for auto-adaptive speech recognition. In: Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. Tampere, pp. 227–230.

Moon, S., Hwang, J.-N., 1997. Robust speech recognition based on joint model and feature space optimization of hidden Markov models. IEEE Trans. Neural Networks 8 (2), 194–204.

Morgan, N., 1999. Temporal signal processing for ASR. In: Proc. 1999 IEEE Workshop on Automatic Speech Recognition and Understanding. Keystone, CO.

Nadas, A., 1985. Optimal solution of a training problem in speech recognition. IEEE Trans. Acoust. Speech Signal Process. 33 (1), 326–329.

Paul, D.B., 1997. Extensions to phone-state decision-tree clustering: single tree and tagged clustering. In: Proc. ICASSP-97. Munich, Germany, pp. II-1487–1490.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Rahim, M.G., Juang, B.-H., 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. IEEE Trans. Speech Audio Process. 4 (1), 19–30.

Rahim, M.G., Juang, B.-H., Chou, W., Buhrke, E., 1996. Signal conditioning techniques for robust speech recognition. IEEE Signal Process. Lett. 3 (4), 107–109.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, UK.

Sankar, A., Lee, C.-H., 1996. A maximum likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech Audio Process. 4 (3), 190–202.

Shahshahani, B.M., 1997. A markov random field approach to Bayesian speaker adaptation. IEEE Trans. Speech Audio Process. 5 (2), 183–191.

Shinoda, K., Lee, C.-H., 1997. Structural MAP speaker adaptation using hierarchical priors. In: Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding. Santa Barbara, pp. 381–388.

Shinoda, K., Lee, C.-H., 1998. Unsupervised adaptation using structural Bayes approach. In: Proc. ICASSP-98. Seattle, pp. 793–796.

Siohan, O., Chesta, C., Lee, C.-H., 1999. Hidden Markov model adaptation using maximum a posteriori linear regression. In: Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. Tampere, pp. 147–150.

Siohan, O., Chesta, C., Lee, C.-H., 2000. Joint maximum a posteriori adaptation of transformation and HMM parameters, In: Proc. ICASSP-2000. Istanbul, Turkey.

Stern, R.M., Lasry, M.J., 1987. Dynamic speaker adaptation for feature-based isolated word recognition. IEEE Trans. Acoust. Speech Signal Process. 35 (6), 751–763.

Surendran, A., Lee, C.-H., 1998. Predictive adaptation and compensation for robust speech recognition. In: Proc. ICSLP-98. Sydney.

Surendran, A., Lee, C.-H., 1999. Bayesian predictive approach to adaptation of HMMs. In: Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions. Tampere, pp. 155–158.

Surendran, A.C., Lee, C.-H., Rahim, M., 1999. Nonlinear compensation for stochastic matching. IEEE Trans. Audio Speech Process. 7 (6), 643–655.

Takahashi, J., Sagayama, S., 1997. Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation. Comput. Speech Language 11, 127–146.

Wang, S.-J., Zhao, Y.-X., 1999. On-line Bayesian tree-structured transformation of hidden Markov models for speaker adaptation. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. Keystone, CO.

Zavaliagkos, G., Schwartz, R., Makhoul, J., 1995a. Batch, incremental and instantaneous adaptation techniques for speech recognition. In: Proc. ICASSP-95. Detroit, pp. 676–679.

Zavaliagkos, G., Schwartz, R., McDonough, J., Makhoul, J., 1995b. Adaptation algorithms for large scale HMM recognizers, In: Proc. Eurospeech-95. Madrid, Spain, pp. 1131–1134.

Zhao, Y.-X., 1994. An acoustic-phonetic based speaker adaptation technique for improving speaker independent continuous speech recognition. IEEE Trans. Speech Audio Process. 2 (3), 380–394.

Zhao, Y.-X., 1996. Self-learning speaker and channel adaptation based on spectral variation source decomposition. Speech Communication 18, 65–77.