Title	A Bayesian predictive classification approach to robust speech recognition
Author(s)	Huo, Q; Lee, CH
Citation	IEEE Transactions on Speech and Audio Processing, 2000, v. 8 n. 2, p. 200-204
Issued Date	2000
URL	http://hdl.handle.net/10722/43650
Rights	Creative Commons: Attribution 3.0 Hong Kong License

A Bayesian Predictive Classification Approach to Robust Speech Recognition

Qiang Huo and Chin-Hui Lee

***Abstract*—We introduce a new decision strategy called *Bayesian predictive classification* (BPC) for robust speech recognition where unknown mismatch between training and testing conditions exists. We then propose and focus on one of the approximate BPC approaches called *quasi-Bayes predictive classification* (QBPC). In a series of comparative experiments where the mismatch is caused by additive white Gaussian noise, we show that the proposed QBPC approach achieves a considerable improvement over the conventional *plug-in MAP* decision rule.**

***Index Terms*—Bayesian predictive classification (BPC), plug-in maximum a posteriori (MAP) decision, quasi-Bayes approximation, robust automatic speech recognition.**

I. Introduction

The modern automatic speech recognition (ASR) technology is based on a communication theoretical view of the generation, acquisition and transmission, and perception of speech (e.g., [2]). It builds upon a statistical pattern recognition paradigm. 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, a semantic attribute, etc. We make the following assumptions.

- The *true* joint distribution of $(W, \mathbf{X})$ can 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 Λ and $P_\Gamma(W)$ as the language model with parameters Γ;
- The full knowledge of the parameters (Λ, Γ) of the above distributions is known.

With these assumptions, an *optimal* decoder (speech recognizer) which achieves the *expected* minimum *word* recognition error rate is the following MAP (maximum *a posteriori*) decoder (see [16] for a more general discussion on statistical decision theory):

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

where $\mathbf{X}$ is the observation and $\hat{W}$ is the recognition result. However, in practice, neither do we know the *true* parametric form of $p(W, \mathbf{X})$, nor its *true* parameters. We shall say that we have *prior uncertainty* in this case. Therefore, the above optimal speech recognizer will never be realizable. Approximation to the optimal decoder is often needed. A simple heuristic solution is first to *assume* some parametric form for $p(W, \mathbf{X})$ and then to *estimate* its parameters (Λ, Γ) from some training data by using particular parameter estimation techniques. Then, the estimate $(\tilde{\Lambda}, \tilde{\Gamma})$ is plugged into the optimal, but unrealizable, rule in (1) in place of the correct but unknown (Λ, Γ) to obtain a *plug-in MAP (PI-MAP) rule*. The performance of any such nonconservative rule depends on the accuracy of the model assumptions, the choice of parameter estimation methods, the nature and size of the training data, the nature and degree of the mismatch between training and testing conditions.

In the past few years, we have been adopting a Bayesian paradigm to address and formulate 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.

One way to achieve the performance robustness is to design and construct a robust decision rule, by taking into account the *prior uncertainty*, which makes it less sensitive to the distortions of models for observations to be recognized. By directly modifying the above PI-MAP decision rule, we've been studying and developing a new robust decision strategy called *Bayesian predictive classification* (BPC) approach to improve the robustness of an HMM-based ASR system [5]–[7], [11].

In this paper, the BPC formulation for robust speech recognition is first introduced in Section II. The formulation of the approximate quasi-Bayes predictive classification approach is proposed in Section III. The important issue of prior specification is discussed in Section IV. In Section V, a series of experimental results along with discussions are reported. Finally, we summarize our findings in Section VI.

II. Bayesian Predictive Classification Approach

In our study, it is assumed that the language model is known and only acoustic models are adjusted. Suppose there are M speech units in the recognizer, each being modeled by a Gaussian mixture continuous density HMM (CDHMM). Consider a collection of M such CDHMM's $\Lambda = \{\lambda_q\}_{q=1,\cdots,M}$, where $\lambda_q = (\pi^{(q)}, A^{(q)}, \theta^{(q)})$ denotes the set of parameters of the qth N-state CDHMM used to characterize the qth speech unit, of which, $\pi^{(q)}$ represents the initial state distribution, $A^{(q)}$ is the transition probability matrix, and $\theta^{(q)}$ is the parameter vector composed of mixture parameters $\theta_i^{(q)} = \{\omega_{ik}^{(q)}, m_{ik}^{(q)}, \Sigma_{ik}^{(q)}\}$ for state i. The state observation pdf is assumed to be a mixture of multivariate Gaussian pdf's:

$$p(\mathbf{x}|\theta_i^{(q)}) = \sum_{k=1}^{K} \omega_{ik}^{(q)} \mathcal{N}(\mathbf{x}|m_{ik}^{(q)}, \Sigma_{ik}^{(q)}) \quad (2)$$

where the set of mixture coefficients $\{\omega_{ik}^{(q)}\}$ satisfy the constraint $\sum_{k=1}^{K} \omega_{ik}^{(q)} = 1$, and $\mathcal{N}(\mathbf{x}|m_{ik}^{(q)}, \Sigma_{ik}^{(q)})$ is the kth normal mixture component with $m_{ik}^{(q)}$ being the D-dimensional mean vector and $\Sigma_{ik}^{(q)}$ being the $D \times D$ covariance matrix with its dth diagonal element being $\sigma_{ikd}^{(q)2}$. For notational convenience, it is assumed that all the state observation pdf's have the same number of mixture components.

A. Modeling Uncertainty

In order to model the abovementioned distortions or *prior uncertainty*, the observation $\mathbf{X}^{(q)}$ for the qth class (unit) to be recognized is assumed to have a pdf $p_q(\cdot) \in \mathcal{P}_q(\epsilon_q)$ (where $\mathcal{P}_q(\epsilon_q)$ is a set of admissible distorted densities for the qth class, and $\epsilon_q \geq 0$ is the distortion level). In the special case of $\epsilon_q = 0$ (i.e., no distortion),

Manuscript received August 27, 1997; revised June 21, 1999. This work was supported by ATR and Hong Kong RGC Earmarked Grant under Grant HKU 7016/97E. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wu Chou.

Q. Huo was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. He is now with Department of Computer Science and Information Systems, University of Hong Kong, Hong Kong (e-mail: qhuo@cs.hku.hk).

C.-H. Lee is with Multimedia Communications Research Laboratory, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA.

Publisher Item Identifier S 1063-6676(00)01721-1.

$\mathcal{P}_q(0) = \{p(\mathbf{X}^{(q)}|\lambda_q^{(0)})\}$ is a singleton set consisting of the hypothetical (ideal, non-distorted) pdf of the qth class observation $\mathbf{X}^{(q)}$ with the model parameters $\lambda_q^{(0)}$ estimated from a training set. If $\epsilon_q > 0$ (i.e., some distortions exist), there are many ways to model the possible distortions between pre-trained models and testing observations. These will depend on whether parametric or nonparametric descriptions of the set $\mathcal{P}_q(\epsilon_q)$ are used [12]. A simple way to construct $\mathcal{P}_q(\epsilon_q)$ is to consider the *model parameter uncertainty* as follows:

$$\mathcal{P}_q(\epsilon_q) = \{p(\mathbf{X}|\lambda_q); \lambda_q \in \Omega(\epsilon_q)\}$$

where $\Omega(\epsilon_q)$ denotes an admissible region of the HMM parameter space. The Bayesian inference approach provides a good way to formalize this parameter uncertainty modeling problem.

In a Bayesian framework, we intend to consider the uncertainty of the HMM parameters Λ by treating them as if they were random. Our prior knowledge about Λ is assumed to be summarized in a known joint *a priori* density $p(\Lambda|\varphi^{(0)})$, with $\Lambda \in \Omega$, where Ω denotes an admissible region of the HMM parameter space, and $\varphi^{(0)}$ is the parameter set 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. We will drop the notation $\varphi^{(0)}$ from time to time in cases where there is no confusion. Suppose a training set of the form $\mathcal{X} = \{\mathbf{X}^{(q,r)}\}$ is available, with $\mathbf{X}^{(q,r)}$ denoting the rth training observation sequence associated with the qth speech unit. A posterior distribution can now be constructed as

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

to update our knowledge about Λ. 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) and then use the plug-in MAP decision rule in (1) for recognition. The conventional plug-in MAP decision rule based on the ML estimate of the HMM parameters can be treated as a special case of the above MAP estimate with a non-informative prior.

B. BPC Formulation for Robust Speech Recognition

The principle behind the BPC approach is quite straightforward. Because we assume no knowledge about the possible distortions, we thus 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. If we want to account for model parameters' uncertainty in *recognition*, an *optimal Bayes solution*, namely the *Bayesian predictive classification* (BPC) approach exists which selects a speech recognizer to minimize the *overall recognition error* (this is 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). Readers are referred to [15], [16] for a brief proof of the optimality of the BPC rule. Such a BPC rule operates as follows:

$$\hat{W} = \arg \max_W \tilde{p}(W|\mathbf{X}) = \arg \max_W \tilde{p}(\mathbf{X}|W) \cdot P_\Gamma(W) \tag{4}$$

where

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

is called the *predictive pdf* [1], [3], [16] 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. However, if we directly apply the decision rule in (4) and (5) as suggested in [15] to speech recognition, it will make little difference from the conventional plug-in MAP rule. This is because whatever 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 peak. This makes the predictive pdf in (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 Λ is concentrated at the ML estimate $\tilde{\Lambda}$ obtained from $\mathcal{X}$, then it is easy to see from (4) and (5) that the BPC decision rule coincides with the plug-in MAP decision rule.

In our approach here, we adopt an *empirical Bayes* method in which a specific parametric pdf $p(\Lambda|\varphi)$ is used to represent the prior/posterior pdf of the CDHMM parameters. Consequently, the predictive pdf required for BPC decoding will be computed as

$$\tilde{p}(\mathbf{X}|W) = \int_\Omega p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W)\, d\Lambda. \tag{6}$$

Using $p(\Lambda|\varphi)$, instead of $p(\Lambda|\mathcal{X})$, to represent *prior uncertainty* provides a flexible way to incorporate and make use of possibly available knowledge sources. For example, the set of hyperparameters, φ, could be estimated from some training data, or specified based on some empirical reasoning, or their combination [7]–[9]. This provides the BPC approach a way to be different from the conventional plug-in MAP decoder. As for the relation between our BPC approach and other robust decision approaches such as the *approximate Bayesian decision rule* in [13], the *minimax decision rule* in [14], readers are referred to [4] for a detailed discussion.

Three key issues thus arise in the BPC formulation, namely,

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

In the following two sections, we will discuss how to address the above three issues for the robust speech recognition applications.

III. Approximate BPC Approach

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

$$\begin{aligned}\tilde{p}(\mathbf{X}|W) &= \int p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W)\, d\Lambda \\ &= \sum_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) p(\Lambda|\varphi, W)\, d\Lambda\end{aligned} \tag{7}$$

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. The simplest way is to first generate random samples $\Lambda_1, \Lambda_2, \cdots, \Lambda_n$ from $p(\Lambda|\varphi, W)$. According to the law of large

numbers, there is convergence of the average $(1/n)\ \Sigma_{i=1}^{n}\ p(\mathbf{X}|\Lambda_i, W)$ to the right-hand-side (RHS) of (6) when n goes to ∞. Similarly, we can also perform a double-fold Monte Carlo simulation of both the HMM parameters and the hidden processes (state sequences and mixture label sequences) of the CDHMM. Following this, we then perform averaging of the $p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W)$ over the generated random samples of $\{\mathbf{s}, \mathbf{l}, \Lambda\}$ and hence approximate the RHS of (7). Because of their computational expense, the above Monte Carlo methods are only of academic interest in speech recognition.

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

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

A detailed algorithm to implement the above approximation and the related experimental results are reported in [11].

In this study, we adopt a numerical approximation technique, namely, the *Laplace approximation*, for the integral (e.g., [17]), to compute the approximate predictive pdf. Let us define

$$\hbar(\Lambda) = \log\{p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W)\}. \tag{9}$$

The value of Λ that maximizes $\hbar(\Lambda)$ is the following MAP estimate, Λ_{MAP}

$$\Lambda_{MAP} = \arg \max_{\Lambda \in \Omega_W} p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W). \tag{10}$$

Let's consider a Taylor series expansion of $\hbar(\Lambda)$ about Λ_{MAP},

$$\begin{aligned}\hbar(\Lambda) = {} & \hbar(\Lambda_{MAP}) + (\Lambda - \Lambda_{MAP})^t \hbar'(\Lambda_{MAP}) \\ & + \tfrac{1}{2}(\Lambda - \Lambda_{MAP})^t \hbar''(\Lambda_{MAP})(\Lambda - \Lambda_{MAP}) \\ & + o(\|\Lambda - \Lambda_{MAP}\|^2)\end{aligned} \tag{11}$$

where the superscript "t" denotes the matrix transpose, $\hbar'(\Lambda)$ is the vector of first partial derivatives of $\hbar(\Lambda)$, and $\hbar''(\Lambda)$ is the Hessian matrix of the second partial derivatives of $\hbar(\Lambda)$. Now, $\hbar'(\Lambda_{MAP}) = 0$ because $\hbar(\Lambda)$ reaches a maximum at Λ_{MAP} and so its first derivative is equal to zero at that point. Denote $V^{-1} = -\hbar''(\Lambda_{MAP})$, i.e., V is the $\mathcal{M} \times \mathcal{M}$ modal dispersion matrix with $\mathcal{M}$ being the number of HMM parameters involved in the integrand in (6). Thus

$$\hbar(\Lambda) \approx \hbar(\Lambda_{MAP}) - \tfrac{1}{2}(\Lambda - \Lambda_{MAP})^t V^{-1} (\Lambda - \Lambda_{MAP}). \tag{12}$$

The approximation in (12) does not always hold unless Λ is close to Λ_{MAP}, or $\hbar(\Lambda)$ is highly peaked about its maximum Λ_{MAP}. It follows that

$$\begin{aligned}\tilde{p}(\mathbf{X}|W) = {} & \int_{\Omega} p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W)\ d\Lambda \\ = {} & \int \exp[\hbar(\Lambda)]\ d\Lambda \\ \approx {} & \exp[\hbar(\Lambda_{MAP})] \int \\ & \exp[-\tfrac{1}{2}(\Lambda - \Lambda_{MAP})^t V^{-1} (\Lambda - \Lambda_{MAP})]\ d\Lambda\end{aligned} \tag{13}$$

by (12). Recognizing the integrand in (13) as proportional to a multivariate normal density gives the result

$$\begin{aligned}\tilde{p}(\mathbf{X}|W) \approx {} & p(\mathbf{X}|\Lambda_{MAP}, W) \\ & \cdot p(\Lambda_{MAP}|\varphi, W) \cdot (2\pi)^{\mathcal{M}/2} \cdot |V|^{1/2}.\end{aligned} \tag{14}$$

This is equivalent to use a normal distribution $\mathcal{N}(\Lambda|\Lambda_{MAP}, V)$ 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.

For the simplicity of the discussion, let's consider the isolated word recognition case where each word is modeled by a CDHMM. Let's also only consider the uncertainty of the mean vectors in CDHMM for BPC decoding. The prior pdf of the means for each word CDHMM is assumed to have a Gaussian pdf $\mathcal{N}(\{m_{ikd}\}|\boldsymbol{\mu}, U)$:

$$\begin{aligned}p(\{m_{ikd}\}|W) = {} & \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{d=1}^{D} \\ & \cdot \frac{1}{\sqrt{2\pi} u_{ikd}} \exp\left[-\frac{(m_{ikd} - \mu_{ikd})^2}{2 u_{ikd}^2}\right]\end{aligned} \tag{15}$$

with a collection of the related mean vectors denoted as $\boldsymbol{\mu} = vec\{\mu_{ikd}\}$ and a diagonal covariance matrix denoted as $U = \text{diag}\{u_{ikd}^2\}$. To facilitate the following discussions, we define $\tau_{ikd} = \sigma_{ikd}^2 / u_{ikd}^2$. Given an unknown utterance to be recognized $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$, let $\mathbf{s} = (s_1, s_2, \cdots, s_T)$ be the unobserved state sequence, and $\mathbf{l} = (l_1, l_2, \cdots, l_T)$ be the associated sequence of the unobserved mixture component labels. We can use the *quasi-Bayes* (QB) algorithm in [9], [10] to compute an approximate posterior pdf $p(\{m_{ikd}\}|\mathbf{X}, W)$ which is also a Gaussian pdf $\mathcal{N}(\{m_{ikd}\}|\tilde{\boldsymbol{\mu}}, \tilde{U})$ with hyperparameters

$$\tilde{\mu}_{ikd} = \frac{\tau_{ikd}\mu_{ikd} + c_{ik}\overline{x}_{ikd}}{\tau_{ikd} + c_{ik}} \tag{16}$$

$$\tilde{u}_{ikd}^2 = \frac{\sigma_{ikd}^2}{\tau_{ikd} + c_{ik}} \tag{17}$$

where

$$\xi_t(i,k) = \Pr(s_t = i, l_t = k|\mathbf{X}, \lambda, W) \tag{18}$$

$$c_{ik} = \sum_{t=1}^{T} \xi_t(i,k) \tag{19}$$

$$\overline{\mathbf{x}}_{ik} = \sum_{t=1}^{T} \xi_t(i,k)\mathbf{x}_t / c_{ik}. \tag{20}$$

The above QB procedure is implemented by an iterative EM algorithm. In practice, we observe that several iterations (typically 1 to 3 iterations) are enough to obtain a good recognition result. Now, we can use the Gaussian pdf $\mathcal{N}(\{m_{ikd}\}|\tilde{\boldsymbol{\mu}}, \tilde{U})$ to *approximate* the $\mathcal{N}(\Lambda|\Lambda_{MAP}, V)$. So, we obtain the MAP estimate of m_{ikd} as $\tilde{\mu}_{ikd}$. By further replacing V in (14) with $\tilde{U}$, we can evaluate the approximate predictive pdf in (14) and perform BPC-based recognition. The resulting BPC rule is thus named as the *quasi-Bayes predictive classification*, or QBPC, rule.

IV. Prior Specification

In principle, the efficacy of the BPC approach depends on the appropriateness of the prior pdf for the mismatch we are compensating. If the prior pdf fails to cover the variability reflected in the CDHMM parameters, then BPC will not help much. Therefore, the prior should be carefully specified to make it work for robust speech recognition. Because we have already assumed a specific parametric form for the prior pdf, this turns out to be a hyperparameter specification/estimation problem. If the training data set $\mathcal{X}$ is rich enough to cover the interested variability of speech signal which might possibly occur in the testing conditions, then the *method of moment* algorithm presented

in [8] 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. Readers are referred to [5], [7] for some examples. If the application scenario allows us to have access to some testing data, then by using the sequential Bayesian learning method in [9], [10], 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 [6]. Furthermore, if some knowledge on how the speech signal is distorted and/or varied in different acoustic conditions is available, it will guide us to design a better prior pdf and/or develop a better hyperparameter estimation method. We give an example here for additive white Gaussian noise (AWGN) compensation to show how *knowledge* and *experience* help.

In [14], the power spectral density (PSD) of a block of speech signal (one speech frame of short-time spectral analysis), $S(\omega)$, is assumed to be represented by a rational function of $e^{j\omega}$. If the cepstral coefficients are defined as the inverse Fourier transform of $\log S(\omega)$

$$c_d \triangleq \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \cdot e^{j\omega d} \log S(\omega) \tag{21}$$

then the perturbation reflected in the cepstral coefficients caused by a spectral mismatch between two PSD's $S_1(\omega)$ and $S_2(\omega)$ is bound above as follows:

$$|c_d^{(1)} - c_d^{(2)}| \leq C d^{-1} \rho^d \quad \text{for } d \geq 1 \tag{22}$$

where C $(C > 0)$ is a proportional term and $0 \leq \rho < 1$ denotes the maximum modulus among those zeroes and poles of $S(\omega)$'s. Although in many practical speech recognition systems, some empirical cepstral representations such as MFCC (mel-frequency cepstral coefficients) and LPCC (linear predictive coding cepstrum) are actually used, the above result still approximately holds for these speech representations. This fact motivates the authors of [14] to adopt a uniform distribution for mean vectors of CDHMM in an uncertainty neighborhood of λ as follows:

$$\begin{aligned}\eta(\lambda) = \{\lambda | &\pi_i = \pi_i^*, a_{ij} = a_{ij}^*, \omega_{ik} = \omega_{ik}^*, \Sigma_{ik} = \Sigma_{ik}^*, \\ &\cdot |m_{ikd} - m_{ikd}^*| \leq C\ d^{-1} \rho^d, \\ &1 \leq i \leq N, 1 \leq k \leq K, 1 \leq d \leq D\}\end{aligned} \tag{23}$$

where the hyperparameters C and ρ are used to control respectively the possible mismatch *size* and *shape*, and $\{\pi_i^*, a_{ij}^*, \omega_{ik}^*, m_{ikd}^*, \Sigma_{ik}^*\}$ denote the pre-trained model parameters. This constrained uniform distribution is shown in [14] to work well in a minimax-based recognition of isolated digits for compensating the AWGN-caused distortion as well as the cross-condition mismatch between two different databases. How to choose the optimal values of C and ρ for different mismatches is still an interesting open question though.

In this study, we try to exploit the above *knowledge* and the *experience* in [14] to get a better hyperparameter estimation for BPC-based recognition. Because we are using a Gaussian pdf $\mathcal{N}(\{m_{ikd}\}|\boldsymbol{\mu}, U)$ to serve as the prior, we *set* the mean and variance of this Gaussian distribution to be the mean and variance of the above uniform distribution respectively as follows:

$$\mu_{ikd} = m_{ikd}^* \tag{24}$$

$$u_{ikd}^2 = \tfrac{1}{3} C^2 \rho^{2d} d^{-2}. \tag{25}$$

TABLE I
PERFORMANCE (WORD ACCURACY IN PERCENT) COMPARISON AVERAGED OVER 16 SPEAKERS OF PLUG-IN MAP AND QBPC RULES AS A FUNCTION OF SNR ON TI20 AWGN-CORRUPTED WORD RECOGNITION TASK

Decoding	SNR (dB)						
Methods	∞	35	30	25	20	15	10
PI-MAP	97.5	93.4	90.7	85.7	77.5	64.4	43.7
QBPC	97.6	94.7	92.3	87.9	81.2	72.2	57.0
(C, ρ)	(1,0.1)	(2,0.9)	(5,0.7)	(2,0.8)	(2,0.8)	(4,0.4)	(13,0.2)

This is known to be the best *normal approximation* to the above uniform distribution to minimize the Kullback-Leibler directed divergence of any normal pdf from the above uniform distribution. Its effectiveness will be examined in the following experimental section.

V. EXPERIMENTS AND RESULTS

A series of speech recognition experiments are designed to examine the viability of the proposed BPC algorithm. The task is multispeaker (eight female and eight male speakers) recognition of 20 isolated English words which include ten digits and ten commands namely enter, erase, go, help, no, rubout, repeat, stop, start, yes. The 20-word subset (TI20) of the TI46 corpus is used [8]–[10]. Throughout the following experiments, each word is modeled by a left-to-right five-state whole word CDHMM with arbitrary state skipping. Each state has four Gaussian mixture components with each component having a diagonal covariance matrix. The speech data are down-sampled to 8 KHz. Each feature vector used in this study consists of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30 ms frame length and a 10 ms frame shift. Utterance-based cepstral mean subtraction (CMS) 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.

The type of mismatch to be examined is caused by additive white Gaussian noise (AWGN). For each speaker and each word, about ten training utterances and 16 testing utterances are used. While training is performed on the original clean data, in the testing phase, machine-generated, zero-mean, white Gaussian noise, with various levels of intensity, is added to the original waveform prior to the preprocessing to get the desired signal-to-noise ratio (SNR). The SNR is defined in a global manner (utterance level), that is, if the clean signal $s(t)$ of one utterance contains T samples and the noise samples is $n(t)$, then

$$\text{SNR} \triangleq 10 \log_{10} \frac{\sum_{t=1}^{T} s^2(t)}{\sum_{t=1}^{T} n^2(t)}. \tag{26}$$

By using the hyperparameter specification method described in Section IV, Table I compares, for several SNR values, the recognition accuracy of the standard plug-in MAP decision rule to that of the QBPC approach (1 EM iteration is used) for the best mismatch neighborhood parameter values: C in the range [1,20], and ρ in the range [0,1]. As can be seen, the QBPC introduces considerable improvement, especially at low SNR values. Strictly speaking, the performance of QBPC depends on the appropriate choice of ρ and C, which in turn depends on the unknown nature and the amount of mismatch. However, in our experiments, it is observed that the

recognition performance tends to be relatively insensitive to these control parameters in a reasonably wide range for QBPC [7]. This suggests that the exact knowledge of ρ and C is not crucial to achieve improvement. However, in order to achieve the maximal performance improvement, it will be important to develop a simple on-line adjusting procedure to tune the neighborhood parameters based on only very few training/adaptation data which remains a topic for future research. Readers are also referred to [5]–[7], [11] for more comparative experimental results among approaches of QBPC, Viterbi BPC (VBPC), and minimax on other types of mismatch such as general cross-condition mismatch, cross-gender mismatch, and mismatches caused by many other types of additive noise.

As far as the issue of computational complexity is concerned, the QBPC algorithm is relatively simple to implement and 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-Bayes 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 (14), (16) and (17) 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. When applying the QBPC approach to the continuous ASR problem, it can be operated under an N-best hypotheses re-scoring mode.

As for the QBPC itself, 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.

VI. Summary

In this paper, we introduce a new decision strategy called *Bayesian predictive classification* for robust speech recognition where unknown mismatch between training and testing conditions exists. We propose and focus on one of the approximate BPC approaches called QBPC. In a series of comparative experiments, we have shown how the QBPC approach leads to a considerable reduction of the recognition error rate over the standard *plug-in MAP* scheme. The BPC procedure relies on a quite general prior distribution to characterize the variability of the HMM parameters and does not make rigid assumptions about the possible distortions. Consequently, it might help for many distortion types. This suggests the potential of the BPC approach to serve as a general tool for robust ASR in real applications where any types of mismatch might happen. It is believed that a better understanding and more experience of the *knowledge* and *experience* on how the speech signal is varied under different acoustic conditions will guide us to design a better prior pdf. Although some success has been observed for certain problems, the general issues related to mismatch and robustness are still largely unresolved. 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 poses a big challenge for the so-called robust ASR in this context.

References

[1] J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1975.

[2] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 179–190, Mar. 1983.

[3] S. Geisser, *Predictive Inference: An Introduction*. London, U.K.: Chapman & Hall, 1993.

[4] Q. Huo, "Adaptive learning and compensation of hidden Markov model for robust speech recognition," in *Proc. 1998 Int. Symp. Chinese Spoken Language Processing*, Singapore, Dec. 1998, pp. 31–43.

[5] Q. Huo, H. Jiang, and C.-H. Lee, "A bayesian predictive classification approach to robust speech recognition," in *Proc. ICASSP-97*, Munich, Germany, Apr. 1997, pp. II-1547–1550.

[6] Q. Huo and C.-H. Lee, "Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition," in *Proc. Eurospeech-97*, Rhodes, Greece, Sept. 1997, pp. 1847–1850.

[7] ——, "A study of prior sensitivity for Bayesian predictive classification based robust speech recognition," in *Proc. ICASSP-98*, Seattle, WA, May 1998, pp. II-741–744.

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

[9] 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. Speech Audio Processing*, vol. 5, pp. 161–172, Mar. 1997.

[10] ——, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 386–397, July 1998.

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

[12] Y. Kharin, *Robustness in Statistical Pattern Recognition*. Boston, MA: Kluwer, 1996.

[13] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 2157–2166, Oct. 1991.

[14] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 90–100, Jan. 1993.

[15] A. Nadas, "Optimal solution of a training problem in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 1, pp. 326–329, 1985.

[16] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[17] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *J. Amer. Statist. Assoc.*, vol. 81, no. 393, pp. 82–86, Mar. 1986.