A STUDY OF PRIOR SENSITIVITY FOR BAYESIAN PREDICTIVE CLASSIFICATION BASED ROBUST SPEECH RECOGNITION

Qiang Huo[†] *and Chin-Hui Lee*[‡]

[†]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
Currently at: Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
[‡]Dialogue Systems Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA

ABSTRACT

We previously introduced a new *Bayesian predictive classification* (BPC) approach to robust speech recognition and showed that BPC is capable of coping with many types of distortions. We also learned that the efficacy of the BPC algorithm is influenced by the appropriateness of the prior distribution for the mismatch being compensated. If the prior distribution fails to characterize the variability reflected in the model parameters, then the BPC will not help much. In this paper, we show how the *knowledge* and/or *experience* of the interaction between speech signal and the possible mismatch guide us to obtain a better prior distribution which improves the performance of the BPC approach.

1. INTRODUCTION

Most of the current automatic speech recognition (ASR) systems are using the following *plug-in MAP* (maximum *a posteriori*), or *PI-MAP* decision rule in recognition:

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

where $\mathbf{X}$ is the observed feature vector sequence to be recognized, $p_{\Lambda}(\mathbf{X}|W)$ is the acoustic model with parameters Λ, $P_{\Gamma}(W)$ is the language model with parameters Γ, and $\hat{W}$ is the recognized symbol (usually word) sequence of interest embedded in the observation sequence $\mathbf{X}$. This decision rule is known to achieve an *expected* minimum *symbol sequence* recognition error rate *only if* the assumed models and the estimated parameters were correct [6]. In practice this is never true though, and its performance will depend on the following conditions:

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

In reality, we always have to make some assumptions which are often violated for real observed data. Furthermore, in many real applications, there always exists some form of mismatch between training and testing conditions. But an accurate knowledge of the mismatch mechanism is unknown, and the only available information is the test data along with a set of pre-trained speech models and the decision parameters. To achieve the robust ASR in this context, it is thus desirable to develop a general approach that is cable of handling any mismatches which might encounter in real applications.

In the past few years, we have been adopting a Bayesian paradigm to address and formulate the above problem. 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 in which part of the above-mentioned mismatch can be compensated and the decision performance can be improved [1]. The principle behind the BPC approach is rather straightforward: because we assume no knowledge about the possible mismatch, we thus rely on a quite general prior pdf (*probability density function*) $p(\Lambda|\varphi)$ to characterize the variability of the model parameters caused by the possible modeling/estimation errors and/or mismatches between training and testing conditions. We try to average out this variability while making decision and such a BPC rule operates as follows:

$$\hat{W} = \arg\max_{W} \tilde{p}(W|\mathbf{X}) = \arg\max_{W} \tilde{p}(\mathbf{X}|W) \cdot P_{\Gamma}(W) \quad (2)$$

where

$$\tilde{p}(\mathbf{X}|W) = \int_{\Omega} p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \quad (3)$$

is called the *predictive pdf* of the observation $\mathbf{X}$ given the symbol sequence W. The BPC decision rule is known to achieve a minimum *overall symbol sequence error rate* averaged over both the sampling variation of the expected test data and the parameter uncertainty described by the prior distribution [6].

This work was funded by ATR, with additional partial support for Qiang Huo from HK RGC Earmarked Grant under grant number HKU 7016/97E.

One of the factors which greatly influences the efficacy and performance of BPC is the appropriateness of the prior pdf for the mismatch we are compensating for. On the one hand, the BPC procedure does not make rigid assumptions about the possible distortions. Consequently, by using a very simple prior pdf specification method, we have shown in [1] that BPC helps with many types of distortions. Furthermore, if we have access to some testing data, by combining BPC with the data-driven on-line Bayesian adaptation techniques [3], we can make the prior pdf more appropriate and thus the robustness of the ASR system can be further enhanced as shown in [2]. On the other hand, if the prior pdf fails to characterize the variability reflected in the model parameters, then BPC will not help much. In this case, the *knowledge* and/or *experience* of the interaction between speech signal and the possible mismatch will be very helpful to guide us to obtain a better prior pdf which can improve the BPC performance. It is this approach that this paper focuses on.

More specifically, the *knowledge* to be exploited is a very rough one which is based on the observation (*experience*) and/or assumption (modeling intuition) that the effects of many different sources of acoustic variation can be reflected as a displacement of the locations of the poles (and zeros, if used in speech modeling) of the speech signal represented in the z-plane. As a first step, the knowledge applied is the kind of *boundary knowledge*. There is little internal *structural knowledge* involved. We will show in the following sections how this *partial knowledge* can be effectively incorporated into the currently successful but usually labeled by speech scientist as *ignorant* speech modeling framework, via a well-defined mathematical tool (BPC approach), to improve the speech recognition performance.

2. APPROXIMATE BPC APPROACH

In this study, it is assumed that the language model is known and only acoustic models are adjusted. For the simplicity of the discussion, we consider the isolated word recognition case where each word is modeled by an N-state continuous density hidden Markov model (CDHMM) whose parameters are denoted as $\lambda = (\pi, A, \theta)$, where π is the initial state distribution, A is the transition matrix, and θ is the parameter vector composed of mixture parameters $\theta_i = \{\omega_{ik}, m_{ik}, \Sigma_{ik}\}$ for state i. The state observation pdf is assumed to be a mixture of multivariate Gaussian pdf's: $p(\mathbf{x}|\theta_i) = \sum_{k=1}^{K} \omega_{ik}\mathcal{N}(\mathbf{x}|m_{ik}, \Sigma_{ik})$, where $\mathcal{N}(\mathbf{x}|m_{ik}, \Sigma_{ik})$ denotes a Gaussian pdf for random vector $\mathbf{x}$ with m_{ik} being the D-dimensional mean vector and Σ_{ik} being the $D \times D$ diagonal covariance matrix with its d-th diagonal element being σ^2_{ikd}, ω_{ik} is the mixture gain, and K is the number of mixture components. We adopt the so-called *quasi-Bayes predictive classification* (QBPC) approach [1] for recognition where the predictive pdf is computed approximately as follows:

$$\hat{p}(\mathbf{X}|W) \approx p(\mathbf{X}|\Lambda_{MAP}, W) \cdot p(\Lambda_{MAP}|\varphi, W) \cdot (2\pi)^{\mathcal{M}/2} \cdot |V|^{1/2} \tag{4}$$

where $\Lambda_{MAP} = \arg\max_{\Lambda} p(\mathbf{X}|\Lambda, W)p(\Lambda|\varphi, W)$, $\mathcal{M}$ is the number of HMM parameters involved in the integrand in Eq. (3), and V is the $\mathcal{M} \times \mathcal{M}$ approximate modal dispersion matrix evaluated at $\Lambda = \Lambda_{MAP}$. Furthermore, we only consider the uncertainty of the mean vectors. The prior pdf of the means for each word CDHMM is assumed to have a Gaussian pdf $\mathcal{N}(m_{ikd}|\mu, U)$:

$$p(m_{ikd}|W) = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}u_{ikd}} \exp[-\frac{(m_{ikd} - \mu_{ikd})^2}{2u^2_{ikd}}] \tag{5}$$

with a collection of the related mean vectors denoted as $\mu = vec\{\mu_{ikd}\}$ and a diagonal covariance matrix denoted as $U = diag\{u^2_{ikd}\}$. To facilitate the following discussions, we define $\tau_{ikd} = \sigma^2_{ikd}/u^2_{ikd}$. 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 [3] to compute an approximate posterior pdf $p(m_{ikd}|\mathbf{X}, W)$ which is also a Gaussian pdf $\mathcal{N}(m_{ikd}|\tilde{\mu}, \tilde{U})$ with hyperparameters

$$\tilde{\mu}_{ikd} = \frac{\tau_{ikd}\mu_{ikd} + c_{ik}\bar{x}_{ikd}}{\tau_{ikd} + c_{ik}} \tag{6}$$

$$\tilde{u}^2_{ikd} = \frac{\sigma^2_{ikd}}{\tau_{ikd} + c_{ik}} \tag{7}$$

where $c_{ik} = \sum_{t=1}^{T} \zeta_t(i,k)$, $\bar{\mathbf{x}}_{ik} = \sum_{t=1}^{T} \zeta_t(i,k)\mathbf{x}_t/c_{ik}$, and $\zeta_t(i,k) = \Pr(s_t = i, l_t = k|\mathbf{X}, \lambda, W)$. The above QB procedure is implemented by EM algorithm and thus an iterative one. In practice, we observe that several iterations (typically 1 to 3 iterations) are enough to get good recognition results. From the posterior pdf $p(m_{ikd}|\mathbf{X}, W)$, we can easily get the MAP estimate of m_{ikd} as $\tilde{\mu}_{ikd}$. By further replacing V in Eq. (4) with $\tilde{U}$, we can easily evaluate the approximate predictive pdf in Eq. (4) and perform BPC-based recognition.

3. SENSITIVITY OF PRIORS

3.1. Prior Specification

As discussed above, 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 and big enough to cover the interested variability of speech signal which possibly occurs in the testing conditions, then the *method of moment* algorithm presented in [4] can be used to automatically estimate the hyperparameters from the training data $\mathcal{X}$. Otherwise we have to use some *ad hoc* method for hyperparameter estimation. One of such methods is described originally in [3] and also adopted here for BPC-based recognition. In the special case of only considering the mean uncertainty, the related hyperparameters are derived in the last iteration of seed CDHMM's training as follows: $\mu_{ikd} = m_{ikd}$, $\tau_{ikd} = \epsilon_1 \cdot \sum_t \zeta_t(i,k)$, where $\epsilon_1 > 0$ is a weighting coefficient to control the *degree* of the uncertainty of the prior distribution. In this study, the weighting coefficient ϵ_1 was chosen to be $1/W$ with W 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 distincts BPC from the conventional plug-in MAP decoding. Once the prior pdf's are specified for each CDHMM, the QBPC-based speech recognition can be carried out as described in the previous sections. To be more flexible, we can further introduce a refreshing coefficient "rf" for the hyperparameters to control the degree of the uncertainty of the CDHMM parameters, where $rf = 1$ means no change, $rf > 1$ means to decrease the uncertainty of the HMM parameters (i.e., to trust more the current estimate of the HMM parameters), and $rf < 1$ means to increase the uncertainty of the HMM parameters. Accordingly, the updating formulas in Eqs. (6) and (7) are now modified as follows:

$$\tilde{\mu}_{ikd} = \frac{rf \cdot \tau_{ikd} \cdot \mu_{ikd} + c_{ik}\bar{x}_{ikd}}{rf \cdot \tau_{ikd} + c_{ik}} \tag{8}$$

$$\tilde{u}^2_{ikd} = \frac{\sigma^2_{ikd}}{rf \cdot \tau_{ikd} + c_{ik}} \tag{9}$$

The above prior specification method was shown in [1, 2] to work well in compensating for several types of mismatch such as the general cross-condition mismatches in genders, speakers, speaking styles, recording environments, transducers, etc.

3.2. Better Prior From Knowledge and Experience

In the problem we are coping with, we assume we do not have enough knowledge about the possible mismatches and/or distortions. So, we use a "*semi-blind*" compensation type of technique, like BPC, to exploit the information provided by testing data and the existing models themselves to achieve some robustness. A better understanding on how the speech signal is distorted and/or varied in different acoustic conditions will be helpful to design a better prior pdf in BPC 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 [5], 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 the $S(\omega)$,

$$c_d \triangleq \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \cdot e^{j\omega d} \log S(\omega) \ , \tag{10}$$

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:

$$|c_d^{(1)} - c_d^{(2)}| \le C d^{-1} \rho^d \quad \text{for } d \ge 1 \tag{11}$$

where C ($C > 0$) is a proportional term and $0 \le \rho < 1$ 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 LPC (linear predictive coding) derived cepstral coefficients are actually used, the above result still approximately holds for these speech representations. This fact motivates the authors of [5] to adopt a uniform distribution for mean vectors of CDHMM in an uncertainty neighborhood of λ as follows:

$$\eta(\lambda) = \{\lambda \mid \pi_i = \pi_i^*, a_{ij} = a_{ij}^*, \omega_{ik} = \omega_{ik}^*, \Sigma_{ik} = \Sigma_{ik}^*, \\ |m_{ikd} - m_{ikd}^*| \le C\, d^{-1} \rho^d, 1 \le d \le D\} \tag{12}$$

where the hyperparameters C and ρ are used to control respectively the possible mismatch *size* and *shape*, and $\{\pi_i^*$, a_{ij}^*, ω_{ik}^*, m_{ikd}^*, $\Sigma_{ik}^*\}$ denote the pre-trained model parameters. This constrained uniform distribution is shown in [5] to work well in a minimax-based recognition of isolated digits for compensating for the AWGN-caused distortion as well as the cross-condition mismatch between two different databases.

In this study, we try to exploit the above *knowledge* and the *experience* in [5] to get a better hyperparameter estimation for BPC-based recognition. Because we are using a Gaussian pdf $\mathcal{N}(m_{ikd}|\mu, U)$ 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: $\mu_{ikd} = m_{ikd}^*$, $u_{ikd}^2 = \frac{1}{3}C^2\rho^{2d}d^{-2}$. 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.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

A series of speech recognition experiments are designed to examine the viability of the proposed techniques. The task is multi-speaker (8 female and 8 male speakers) isolated word recognition of 20 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. For each speaker and each word, about 10 training utterances and 16 testing utterances are used. The type of mismatch to be examined is caused by additive white Gaussian noise. 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 global (utterance level) signal-to-noise ratio (SNR).

Throughout the following experiments, each word is modeled by a left-to-right 5-state whole word CDHMM with arbitrary state skipping. Each state has 4 Gaussian mixture components with each component having a diagonal covariance matrix. The speech data are down-sampled to 8 KHz. Each feature vector consists of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30ms frame length and a 10ms 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.

Table 1: Performance (word accuracy in %) comparison averaged over 16 speakers of plug-in MAP and QBPC rules as a function of SNR on TI20 AWGN-corrupted word recognition task: QBPC-I corresponds to the case of an inappropriate prior pdf, and QBPC-II refers to that of an improved prior (1 EM iteration for QBPC, $rf = 1.0$)

SNR (dB)	Decoding Methods		
	PI-MAP	QBPC-I	QBPC-II (C, ρ)
∞	97.5	95.6	97.6 (1,0.1)
35	93.4	93.1	94.7 (2,0.9)
30	90.7	89.8	92.3 (5,0.7)
25	85.7	84.2	87.9 (2,0.8)
20	77.5	74.5	81.2 (2,0.8)
15	64.4	58.3	72.2 (4,0.4)
10	43.7	37.7	57.0 (13,0.2)

4.2. Effects of the Knowledge and Experience

Table 1 compares, the average recognition accuracy over 16 speakers of the standard plug-in MAP decision rule to that of the QBPC method at seven different SNR levels. Here SNR $= \infty$ means that no noise is added to the test utterances. One EM iteration is performed for QBPC and the refreshing coefficient "rf" is set to be 1.0. "QBPC-I" corresponds to the case of that the hyperparameters of the prior pdf are estimated with the method described in Subsection 3.1. It is observed that QBPC degrades the performance. This suggests that the current prior pdf is not appropriate for compensating for the AWGN caused mismatch. By using the improved hyperparameter specification method described in Subsection 3.2, the column "QBPC-II" shows the recognition accuracy of the QBPC approach 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 method 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. As an example, we list the recognition performance as a function of the neighborhood parameters C and ρ for QBPC at SNR=15dB in Table 2. It is observed that the recognition performance tends to be relatively insensitive to these control parameters in a reasonably wide range for QBPC. A similar behavior was observed for other SNR values as well. This suggests that 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.

5. DISCUSSION AND CONCLUSION

We have shown in a case study that the *knowledge* and *experience* on how the speech signal is distorted and/or varied under mismatched conditions are helpful to give a better hyperparameter estimation which in turn improves the BPC performance. Although the experiments are for the compensation of AWGN-caused mismatch, we expect that the same formulation will also work for compensating for other type of mismatches whose effects can be characterized by poles' and zeros' perturbation. Furthermore, 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 and the related hyperparameter estimation of the prior pdf's in BPC, and/or a better structural model in structure-based compensation. It will also be crucial for efficient adaptation and compensation to formulate and develop appropriate mathematical tools for discovering a good intrinsic structural model of speech in the acoustic, phonetic and linguistic aspects.

Table 2: Recognition accuracy (in %) as a function of neighborhood parameters C and ρ for QBPC at SNR=15 dB (1 EM iteration; $rf = 1.0$; PI-MAP attains 64.4% correct rate)

$C\backslash\rho$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1	64.9	66.6	68.0	69.4	70.4	71.1	71.6	71.8
2	66.6	68.8	70.3	71.4	71.8	72.0	71.7	70.6
3	67.8	69.8	71.1	71.8	72.2	71.5	70.2	69.0
4	68.6	70.6	71.6	72.2	71.9	70.5	68.7	68.1
5	69.2	70.9	71.8	72.2	71.2	69.7	67.5	66.8
6	69.4	71.1	72.0	72.2	70.6	68.5	66.2	65.7
7	69.9	71.4	72.1	71.6	70.1	67.8	65.3	64.6
8	69.9	71.6	72.2	71.2	69.4	66.9	64.4	63.5
9	70.2	71.7	72.1	70.9	68.8	66.0	63.5	62.3
10	70.3	71.7	72.1	70.4	68.3	65.4	62.9	61.0
11	70.5	71.7	71.8	70.0	67.8	64.8	62.4	59.8
12	70.6	71.6	71.6	69.6	67.4	64.3	61.8	58.6
13	70.6	71.6	71.3	69.4	66.8	63.7	60.9	57.1
14	70.7	71.6	71.3	69.0	66.3	63.4	60.2	55.8
15	70.9	71.7	71.0	68.7	66.0	62.9	59.7	54.2
16	70.9	71.8	70.9	68.4	65.6	62.5	59.5	52.8
17	70.9	71.8	70.7	68.1	65.2	62.2	58.9	51.5
18	70.9	71.8	70.4	68.0	64.9	61.8	58.3	50.3
19	71.0	71.8	70.2	67.6	64.6	61.2	57.8	49.4
20	71.1	71.8	70.1	67.2	64.4	60.6	57.2	48.1

REFERENCES

[1] Q. Huo, H. Jiang and C.-H. Lee, "A Bayesian predictive classification approach to robust speech recognition," *Proc. ICASSP-97*, pp.II-1547-1550.

[2] Q. Huo and C.-H. Lee, "Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition," *Proc. Eurospeech-97*, pp.1847-1850.

[3] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on SAP*, Vol. 5, pp.161-172, 1997.

[4] Q. Huo, C. Chan and C.-H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *IEEE Trans. on SAP*, Vol. 3, pp.334-345, 1995.

[5] N. Merhav and C.-H. Lee, "A minimax classification approach with application to robust speech recognition," *IEEE Trans. on SAP*, Vol. 1, pp.90-100, 1993.

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