A BAYESIAN PREDICTIVE CLASSIFICATION APPROACH TO ROBUST SPEECH RECOGNITION

Qiang Huo[a], Hui Jiang [b] and Chin-Hui Lee[c]

[a]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
[b]Department of Information and Communication Engineering, University of Tokyo, Tokyo, Japan
[c]Multimedia Communications Research Lab, Bell Laboratories, Murray Hill, NJ 07974, USA

ABSTRACT

We introduce a new *Bayesian predictive classification* (BPC) approach to robust speech recognition and apply the BPC framework to Gaussian mixture continuous density hidden Markov model based speech recognition. We propose and focus on one of the approximate BPC approach called *quasi-Bayesian predictive classification* (QBPC). In comparison with the standard plug-in maximum *a posteriori* decoding, when the QBPC method is applied to speaker independent recognition of a *confusable vocabulary*, namely 26 English letters, where a broad range of mismatches between training and testing conditions exist, the QBPC achieves around 14% relative recognition error rate reduction. While the QBPC method is applied to cross-gender testing on a *less confusable vocabulary*, namely 20 English digits and commands, the QBPC method achieves around 24% relative recognition error rate reduction.

1. INTRODUCTION

In this paper, we introduce a new *Bayesian predictive classification* (BPC) approach to robust speech recognition. The conventional *plug-in maximum a posteriori* (MAP) decision rule is known to achieve an optimal Bayes decision if the assumed models and parameters of the rule were correct. However, in real world situations, we rarely have the full knowledge about the nature of the classification data to warrant optimal decisions. Furthermore, we often encounter situations 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 the given MAP decision rule and the decision parameters. Some recent approaches have focused on modifying the decision rule and the model parameters so that part of the mismatch can be compensated and the decision performance can be improved. One such approach is the *minimax classification* algorithm [5] which assumes the best decision parameters for the given test data lie in the neighborhoods of the given parameters and adjusts the decision rule and the corresponding parameters accordingly. The minimax classification is thus geared to protect against the possibility of the worst mismatch.

The first author would like to thank Drs. Y. Yamazaki and Y. Sagisaka of ATR-ITL for their support of this work.

The proposed BPC framework improves upon the conservative minimax classification approach by taking into account some prior knowledge about the given plug-in decision rule and parameters. We apply the proposed BPC framework to hidden Markov model (HMM) based speech recognition. Specifically, we model each speech unit with a continuous density HMM (CDHMM) in which each HMM state is characterized by a mixture of multivariate Gaussian densities. Because of the nature of the *missing-data* problem caused by the underlying hidden processes of a CDHMM, it is not easy to compute the *predictive density* required in designing the BPC-based decision rules. To circumvent these difficulties we introduce two approximation procedures. The first one, called *quasi-Bayesian predictive classification* (QBPC), is based on the quasi-Bayesian approximation of the posterior probability density function (PDF) [1] to compute the predictive densities. The second one, called *Viterbi* BPC (VBPC), uses the joint predictive PDF of the observation sequence, the most likely state and mixture component sequences to approximate the predictive density. Details of the VBPC formulation and a case study on robust speaker independent recognition of isolated and connected digits in noise are given in a companion paper [3] for this conference. In this paper, we focus our study on the theoretical and implementation issues related to the QBPC approach. The viability of the techniques is confirmed in a series of comparative experiments using a 26-word English alphabet vocabulary and a 20-word English digit & command vocabulary.

2. BAYESIAN PREDICTIVE CLASSIFICATION

Let's view a *word* W and the associated acoustic observation $\mathbf{X}$ (usually, a feature vector sequence) as a jointly distributed random pair $(W, \mathbf{X})$. Depend on the problem of interest, *word* here could be any linguistic unit, such as a phoneme, a syllable, a word, a phrase, etc. Also note that for notational simplicity, in this paper, we always use the same symbol to denote both the random variable and the value it may assume. Suppose the *true* joint distribution of $(W, \mathbf{X})$ could be modeled by a *true parametric family* of PDF $p(W, \mathbf{X}) = p_\Lambda(\mathbf{X}|W) \cdot p_\Gamma(W)$, where $p_\Lambda(\mathbf{X}|W)$ is known as acoustic model with parameters Λ and $p_\Gamma(W)$ as language model with parameters Γ. Further suppose we have the full knowledge of the parameters (Λ, Γ) of the

0-8186-7919-0/97 $10.00

above distributions. Then, an optimal decoder (speech recognizer) which achieves *expected* minimum *word* recognition error rate is the following MAP decoder:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(W|\mathbf{X}) = \underset{W}{\operatorname{argmax}}\, p_\Lambda(\mathbf{X}|W) \cdot p_\Gamma(W) \quad (1)$$

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. Therefore, the above optimal speech recognizer will never be achievable, but we can only approximate it. A simple heuristic solution is first to assume some parametric form for $p(W, \mathbf{X})$ and then to estimate its parameters from some training data by using some parameter estimation techniques (e.g., maximum likelihood (ML), MAP, discriminative training, etc.). Then, we *plug in* the estimate $(\tilde{\Lambda}, \tilde{\Gamma})$ into the optimal but unavailable rule in equation (1) in place of the correct but unknown (Λ, Γ) to obtain a *plug in 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, and the degree of the mismatch between training and testing conditions. It is the last issue that motivates the consideration of other more conservative decision strategies.

According to the nature of the problem as stated at the beginning of the paper, one way to achieve performance robustness in unknown mismatch case is to adopt the *minimax principle* whose essence is to try and protect against the worst possible state of nature. Thus, minimax classification is the most conservative decision strategy. A case study of minimax classification for robust digit speech recognition was presented in [5]. In that study, a specific parametric uncertainty neighborhood surrounding the ML-trained HMM parameters was defined. The HMM parameters are assumed to have a uniform distribution in that neighborhood. So, the resulting minimax decision rule is equivalent to the *plug in MAP rule* in which the HMM parameters of each speech unit are replaced with their on-line constrained ML estimates from the testing utterance itself. Minimax strategy try to secure the decision in the worst case, thus usually do not perform nearly as well as in a less malign situation and/or those techniques which use some prior information of the possible mismatches.

A compromise between risky plug in MAP rule and overduely conservative minimax approach is a decision strategy which can somehow make use of the prior knowledge (albeit crude) about the *possible* mismatch, and at the same time take into account its uncertainty to plan accordingly for the possible severe mismatch. It is such an approach called Bayesian predictive classification approach that this paper focuses on. Suppose only acoustic models are adjusted in this study. We use a prior PDF $p(\Lambda|\varphi)$ to represent our knowledge about the uncertainty of the unknown parameters Λ (e.g. [1]). An *optimal Bayes solution* is to choose a speech recognizer which minimizes the *overall recognition error* when the average is taken both with respect to the sampling variation in the expected testing data and with respect to the uncertainty described by the prior distribution. Such a BPC rule is operated as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, \tilde{p}(W|\mathbf{X}) = \underset{W}{\operatorname{argmax}}\, \tilde{p}(\mathbf{X}|W) \cdot p_\Gamma(W) \quad (2)$$

where

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

is called the predictive PDF of the observation $\mathbf{X}$ given the word W. The computation of this predictive PDF is 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 predictive methods average over the uncertainty in parameters. We wish to draw the reader's attention to the work in [6] and [4]. We are actually using a very similar formulation as Nadas did in [6]. He was using a posterior PDF $p(\Lambda|\mathcal{X})$ derived from a *training set* $\mathcal{X}$ to serve as the prior PDF in predictive decision making and gave a simple example in which reproducing density existed. We start up where Nadas [6] left off, with an *empirical Bayes* method in which a specific parametric PDF $p(\Lambda|\varphi)$ is adopted to represent the prior PDF of the CDHMM parameters. Its hyperparameters φ could be estimated from some training data, or specified based on some empirical reasoning, or their combination [1]. We then use different approximation methods discussed in next section to compute the approximate predictive PDF and use the BPC rule in equation (2) to perform recognition. We can actually go one step further. By combing such decision strategy with the on-line model adaptation strategy [1, 2] to continuously update our prior knowledge about the uncertainty of the model parameters, we can approach a performance achieved by the plug-in MAP rule under a matched condition. Although Merhav and Ephraim [4] also started with Nadas's formulation, they finally used another so-called *approximate Bayesian decision rule* which was based on the generalized likelihood ratios computed from the available training and testing data.

3. APPROXIMATE BPC APPROACHES

In the CDHMM case, we have no closed form solution for the computation of the predictive PDF $\tilde{p}(\mathbf{X}|W)$. One way to compute an approximate predictive PDF is to use the Monte Carlo method. We can use the Monte Carlo simulation of the hidden processes (state sequence and mixture label sequence) of the CDHMM and then perform integration and averaging. We can also perform a double-fold Monte Carlo simulation of both the hidden processes and the HMM parameters, and then perform only averaging. Because it's computationally expensive, the Monte Carlo method has only of academic interest in the stage of performing speech recognition.

Another way to compute the approximate predictive PDF is to use the following *Laplace method for integrals*:

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

where $\Lambda_{MAP} = \underset{\Lambda}{\operatorname{argmax}}\, p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W)$, $\mathcal{M}$ is the number of HMM parameters involved in the integrand in equation (3), and V is the $\mathcal{M} \times \mathcal{M}$ modal dispersion matrix, i.e., $-V^{-1}$ is the Hessian matrix of second derivatives of $\log\{p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W)\}$ evaluated at $\Lambda = \Lambda_{MAP}$. This approximation is also known as the *normal approximation*

method in Bayesian community, because we are equivalently using a normal PDF $\mathcal{N}(\Lambda; \Lambda_{MAP}, V)$ to approximate the posterior PDF $p(\Lambda|\mathbf{X}, W)$. To compute V directly is still too computationally involved. So, we have to make further approximation. If we only consider the uncertainty of the mean vectors in CDHMM, we can use the QB algorithm in [1] or [2] to compute an approximate posterior PDF $\mathcal{N}(\Lambda; \Lambda_{MAP}, \tilde{V})$ and then replace V in equation (4) with $\tilde{V}$. We thus name the resultant BPC rule as QBPC rule.

A third way to compute the approximate predictive PDF is to use the following Viterbi approximation:

$$\tilde{p}(\mathbf{X}|W) \approx \max_{\mathbf{s},\mathbf{l}} \int p(\mathbf{X}, \mathbf{s}, \mathbf{l}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \qquad (5)$$

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 $\mathbf{X}$. A detailed algorithm to implement the above approximation is presented in another paper [3]. The resultant BPC rule is called VBPC rule.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

Two sets of speech recognition experiments are designed to examine the viability of the proposed QBPC algorithm. 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 TI46 corpora were used [1]. For speaker independent (SI) training and initial prior density estimation, the OGI ISOLET database produced by 150 speakers was used. For SI testing, the alphabet subset of the TI46 isolated word corpus produced by 16 speakers was used. Each person utters each of the letters 26 times. Among them, 8 of them were 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, transducer, recording environments and conditions, sampling rate and quantization resolution, etc. For the second set of experiments, 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. 20 English words subset (TI20) of the TI46 corpus was used. We train 2 sets of gender-dependent models (both CDHMMs and their initial prior PDFs) from 8 female and 8 male speakers by 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) by using about 16 tokens per word for each speaker.

Throughout the following experiments, each word is modeled by a left-to-right 5-state CDHMM with arbitrary state skipping and each state has 4 Gaussian mixture components with diagonal covariance matrices. The speech data in both corpora are down-sampled to 8 KHz. Each feature vector consists of 12 LPC-derived cepstral coefficients and utterance-based cepstral mean subtraction (CMS) is applied for acoustic normalization both in training and testing. The initial hyperparameters are estimated by using the method described in [1] where we normalize the importance of the initial prior knowledge to be comparable with the contribution from a single training token. In QBPC decoding, we can further set the refreshing coefficient rf (see [1] for the explanation) of 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. Note that in this study we only consider the uncertainty of the mean vectors of CDHMMs which is characterized by a set of Gaussian PDFs.

Table 1: Performance (word accuracy in %) comparison averaged over 8 female speakers of plug-in MAP, QBPC and minimax rules on English letter recognition task ($rf = 1.0$)

number of EM iterations	Decoding Methods		
	plug-in	QBPC	minimax
1	49.1	53.7	52.0
2	N/A	55.1	49.6
3	N/A	56.2	48.2

Table 2: Performance (word accuracy in %) comparison averaged over 8 female speakers on English letter recognition task by operating QBPC rule under different values of refreshing coefficient and numbers of EM iteration

number of EM iterations	Refreshing Weights				
	2.0	1.5	1.0	0.75	0.5
1	52.9	53.2	53.7	54.1	53.3
2	54.3	54.5	55.1	55.0	53.5
3	54.8	55.3	56.2	55.5	53.6

4.2. English Letter Recognition Results

Table 1 compares, the averaged recognition accuracy over 8 female speakers of the standard plug-in MAP decision rule to that of the QBPC and a modified minimax method with different EM iterations on SI English letter recognition task. For the minimax method adopted in this study, we just use $p(\mathbf{X}|\Lambda_{MAP}, W)$ in equation (4) to approximate the predictive PDF. In comparison with [5], we are using a more informative prior here 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.

Table 2 compares, the averaged recognition accuracies over 8 female speakers on English letter recognition task by operating QBPC rule under different values of refreshing coefficient and different numbers of EM iteration. It turns out that in a reasonably wide range of values of the control parameters, the QBPC method achieves improvement over that of conventional plug-in MAP method.

Table 3: Performance (word accuracy in %) comparison averaged over 8 male speakers on TI20 word recognition task by using female seed models and operating QBPC rule under different values of refreshing coefficient and numbers of EM iteration (the recognition rate is 40.5% by using standard plug-in method)

number of EM iterations	Refreshing Weights				
	2.0	1.5	1.0	0.5	0.25
1	47.9	49.4	51.5	54.4	53.7
2	49.3	50.7	52.4	54.3	53.9
3	49.7	51.4	53.5	54.6	54.9

Table 4: Performance (word accuracy in %) comparison averaged over 8 male speakers on TI20 word recognition task by using male seed models and operating QBPC rule under different values of refreshing coefficient and numbers of EM iteration (the recognition rate is 98.4% by using standard plug-in method)

number of EM iterations	Refreshing Weights				
	2.0	1.5	1.0	0.5	0.25
1	97.8	97.5	97.5	96.9	95.7
2	97.6	97.5	97.2	96.1	94.0
3	97.5	97.5	97.2	95.9	93.6

4.3. Experimental Results on TI20

Table 3 compares, the averaged recognition accuracies over 8 male speakers on TI20 word recognition task by using female seed models and operating QBPC rule under different values of refreshing coefficient and different numbers of EM iteration. The similar facts as the above are also observed here and the QBPC method achieved around 24% relative recognition error rate reduction over that of the standard plug-in method.

To examine the behavior of the QPBC method under the matched condition, we listed in Table 4 the experimental results averaged over 8 male speakers on TI20 word recognition task by using the male seed models. The results show that the QBPC method holds up the performance or only degrades slightly in matched training/testing condition under a reasonably wide range of control parameters.

5. DISCUSSION AND CONCLUSION

In this paper, we start with a revisit to the statistical formulation of the speech recognition problem, take a critical view of the two existing recognition strategies, namely *plug-in MAP* and *minimax*, and finally introduce a new decision strategy called *Bayesian predictive classification* for robust speech recognition where unknown mismatch between training and testing conditions exists. More specifically, we propose and focus on one of the approximate BPC approach called QBPC and show how it leads to considerable reduction of the error rate over the standard nonrobust scheme via a series of comparative experiments. The QBPC algorithm is relatively simple to implement and no big increase of the computational complexity. Generally speaking, in the case of less confusable vocabulary where the speech models are distinct enough and the mismatch is not so severe, to use a less *informative* prior distribution such as the uniform distribution we adopted in [3] will not cause any problem. On the other hand, it might even be beneficial when the mismatch neighborhood described by this prior distribution happens to be consistent with the real mismatch which is the case for additive Gaussian white noise in our study in [3]. So the effect of the VBPC decoding in [3] is especially pronounced in our experiments. However, the absolute recognition rate after QBPC or VBPC decoding in severely mismatched case is still far inferior to that of the matched testing results. How to bridge this performance gap is still a challenging topic for further research. If the application involves a recognition session which might consists of a number of testing utterances, then a combined BPC decoding and on-line adaptation of the HMM parameters will provide a good solution to enhance the robustness towards varying environments, microphones, channels, speakers, and other general mismatches or distortions. We will report those results elsewhere. We are also checking how the QBPC method works in other mismatch conditions such as different type of additive noises. More theoretical work is needed to include the uncertainty of the other HMM parameters than mean vectors into the QBPC framework. It will also be interesting to explore the possibility of applying BPC framework to utterance verification problem. As a final remark, like minimax, QBPC will encounter some difficulties while extending to the continuous speech recognition problem, but it can be easily operated under an N-best hypothesises re-scoring mode.

REFERENCES

[1] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on Speech and Audio Processing*, March 1997.

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

[3] H. Jiang, K. Hirose and Q. Huo, "Robust speech recognition based on Viterbi Bayesian predictive classification," *Proc. ICASSP-97*, 1997.

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

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

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