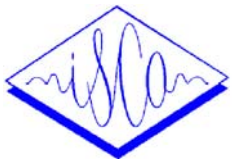

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

COMBINED ON-LINE MODEL ADAPTATION AND BAYESIAN PREDICTIVE CLASSIFICATION FOR 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
[‡]Multimedia Communications Research Lab, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA

ABSTRACT

In this paper, we study a class of robust automatic speech recognition problem in which mismatches between training and testing conditions exist but an accurate knowledge of the mismatch mechanism is unknown. The only available information is the test data along with a set of pre-trained speech models and the decision parameters. We try to compensate for the abovementioned mismatches by jointly adopting a dynamic system design strategy called on-line Bayesian adaptation to incrementally improve the estimation of the model parameters used in the recognizer, and a robust decision strategy called Bayesian predictive classification to average over the remaining uncertainty in model parameters. We report on a series of experimental results to show the viability and effectiveness of the proposed method.

1. INTRODUCTION

In the last two decades, much advance has been achieved in the area of automatic speech recognition (ASR). This is largely attributed to the use of a powerful statistical pattern recognition paradigm and the application of dynamic programming search over a structural network representation of acoustic and linguistic knowledge sources. 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 $(W, \mathbf{X})$. Depend on the problem of interest, *word* here could be any linguistic unit, such as a phoneme, a syllable, a word, a phrase, etc. Suppose the *true* joint distribution of $(W, \mathbf{X})$ could be modeled by a *true parametric family* of PDF (probability density function) $p(W, \mathbf{X}) = p_\Lambda(\mathbf{X}|W) \cdot p_\Gamma(W)$, where $p_\Lambda(\mathbf{X}|W)$ is known as 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 above distributions. Then, an optimal decoder (speech recognizer) which achieves the *expected* minimum *word* recognition error rate is the following MAP (maximum *a posteriori*) decoder:

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

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. 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 non-conservative 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 which may arise from inter- and intra- speaker variabilities, transducer, channel and other environmental variabilities, and many other phonetic and linguistic effects due to a task mismatch problem. It is the susceptibility of current ASR systems to even moderate acoustic mismatches that prevents the widespread deployment of the ASR systems in those applications where they can be most useful. Robust speech recognition in this context thus refers to the problem of designing an automatic speech recognizer that works well for different tasks and speakers over a wide range of unexpected and possibly adverse conditions.

There has been a great deal of effort aiming at improving speech recognition and hence enhancing performance robustness in the abovementioned mismatches. In the past few years, we have been adopting a Bayesian paradigm to address and formulate a class of robust speech recognition problem in which mismatches between training and testing conditions exist, but an accurate knowledge of the mismatch mechanism is unknown. The only available information is the test data along with a set of pre-trained speech models and the decision parameters. We've developed two sets of Bayesian techniques to cope with the acoustic mismatch problem for Gaussian mixture continuous density hidden Markov model (CDHMM) based speech recognition. The first type of algorithms are targeting those applications involving a recognition session which might consist of a number of testing utterances. Unlike those ASR systems which rely on a static design strategy that all the knowledge sources needed in a system are acquired at the design phase and remain fixed during use, we adopt a dynamic system design strategy where the new knowledge is acquired dynamically. New information is constantly collected during development and use of the ASR system, and is incorporated into the system using an adaptive learning algorithm, namely on-line Bayesian adaptive learning of the HMM parameters [1, 2]. For the second type of techniques, by modifying directly

The first author would like to thank Drs. Y. Yamazaki and Y. Sagisaka of ATR-ITL for their support of this work.
Email: qhuo@itl.atr.co.jp or chl@research.bell-labs.com.

the above plug-in MAP decision rule, we've developed a new robust decision strategy called *Bayesian predictive classification* (BPC) approach [3] so that part of the mismatch can be compensated and the decision performance can be improved.

The robustness of the ASR system can be further enhanced by integrating on-line adaptation of model parameters with BPC-based decoding. This is exactly what we want to present in this paper. The technical details of each component technique can already be found in [1, 2, 3]. In the remainder of the paper, we first give a summary of the basic principle of the algorithms and then report a series of experimental results to show the viability and effectiveness of the proposed method.

2. DYNAMIC SYSTEM DESIGN STRATEGY: ON-LINE BAYESIAN ADAPTATION

For this approach, we assume that our initial knowledge about HMM parameters Λ is contained in and represented by a known joint *a priori* PDF $p(\Lambda|\varphi^{(0)})$ with hyperparameters $\varphi^{(0)}$. Starting from this, when speech utterances $\mathcal{X}_1, \mathcal{X}_2, \cdots$ successively become available, repeated use of the following equation:

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

produces the sequence of densities $p(\Lambda|\mathcal{X}_1^1, \varphi^{(1)})$, $p(\Lambda|\mathcal{X}_1^2, \varphi^{(2)})$, and so forth, where $\mathcal{X}_1^n = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$ denotes n sets of independently obtained speech utterances, and $p(\mathcal{X}_n|\Lambda)$ is the likelihood function. This provides a basis of making formal recursive Bayesian inference of parameters Λ and thus a good solution for on-line HMM adaptation. However, there are some serious computational difficulties to directly implement this learning procedure. Consequently, some approximations are needed in practice. We've presented several solutions, e.g., [1, 2], in which we also show that the system performance can be consistently improved by using a *plug-in MAP* decision rule for recognition and on-line adaptation (OLA) for HMM parameter compensation.

3. ROBUST DECISION STRATEGY: BAYESIAN PREDICTIVE CLASSIFICATION

As noted before, the conventional *plug-in* MAP decision rule is known to achieve an optimal Bayes decision only if the assumed models and parameters of the rule were correct. Although OLA can continuously make the model parameters match the coming data, in the early stages of the OLA, the model parameters will not be good enough to warrant *plug-in* MAP rule a good performance, if severe mismatch exists initially. This motivates us to modify directly the plug-in MAP decision rule and develop a new robust decision strategy called BPC [3]. 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. More specifically, like in OLA, we use a prior PDF $p(\Lambda|\varphi)$ with hyperparameters φ to represent our knowledge about the uncertainty of the unknown parameters Λ. 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. Suppose only acoustic models are adjusted. Such a BPC rule is operated as follows:

$$\hat{W} = \arg\max_W \tilde{p}(W|\mathbf{X}) = \arg\max_W \tilde{p}(\mathbf{X}|W) \cdot p_\Gamma(W) \quad (3)$$

where

$$\tilde{p}(\mathbf{X}|W) = \int p(\mathbf{X}|\Lambda, W) p(\Lambda|\varphi, W) \mathrm{d}\Lambda \quad (4)$$

is called the predictive PDF of the observation $\mathbf{X}$ given the *word* W. Once again, for HMM, we have no closed form solution for the computation of this predictive PDF and some approximations are needed. We've developed several approximation procedures. One of them called *quasi-Bayesian predictive classification* (QBPC). We have shown in [3] that by using the QBPC alone (no OLA), we can improve the recognition performance in comparison with the plug-in MAP decision rule when mismatches between training and testing conditions exist.

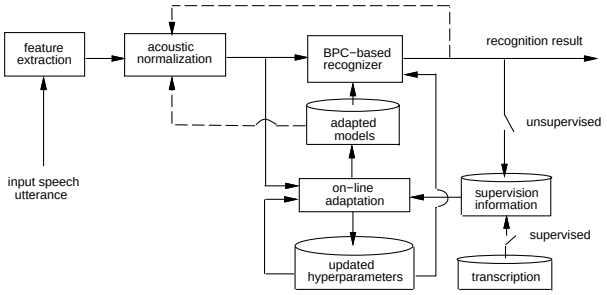

Figure 1: A block diagram of a robust HMM-based speech recognition system

4. COMBINED ON-LINE MODEL ADAPTATION AND BAYESIAN PREDICTIVE CLASSIFICATION

Because both OLA and BPC are formulated under a unified Bayesian paradigm to address respectively the model parameter inference problem and the decision problem, they can be seamlessly combined to produce an enhanced algorithm to cope with the robust ASR problem as described in the introduction section. Such a robust ASR system is schematically shown in Figure 1. Given a new block of input speech, feature extraction (usually spectral analysis) is first performed to derive the feature vector sequences used to characterize the speech input. It is followed by some kind of acoustic normalization to reduce the possible mismatch in the feature vector space. The processed feature vector sequences are then recognized based on the current set of HMMs by using BPC approach. After the recognition of the current block of utterances, the HMMs and the posterior distributions of the related speech units are adapted and the updated models are used to recognize future input utterance(s). In this way, we can get a better and better posterior/prior PDF (i.e., more and more accurate knowledge about the uncertainty of the model parameters), and this in turn makes the BPC-based recognition system approach a performance achieved by the plug-in MAP rule under a matched condition.

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup

Two sets of speech recognition experiments are designed to examine the viability of the proposed combined OLA and BPC method where we use the quasi-Bayes method as described in [1] for on-line adaptation of the independent CDHMMs and the QBPC method in [3] for recognition. The first set of experiments 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 of CDHMMs, the OGI ISOLET database produced by 150 speakers was used. For on-line condition & speaker adaptation and 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 tokens were used for adaptation and another 8 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 on-line speaker adaptation and testing. For each speaker, we have about 10 tokens per word for OLA and 16 tokens per word for testing.

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. Note that in this study, although we consider the uncertainty of all CDHMM parameters for OLA, we only consider, for QBPC, the uncertainty of the mean vectors of CDHMMs which is characterized by a set of Gaussian PDFs. Also note that all of the OLA experiments are performed in a supervised mode.

5.2. English Letter Recognition Results

Figure 2 shows the performance comparison averaged over 8 female speakers on English letter recognition task as a function of *total* number of OLA tokens (e.g., 26 means for every vocabulary word, we have one adaptation token) among several methods. Without any compensation, as expected, the cross-condition SI recognition rate is very low. With OLA and conventional plug-in MAP decoding (denoted as "Plug-in-MAP+OLA"), the performance is continuously improved with increasing amount

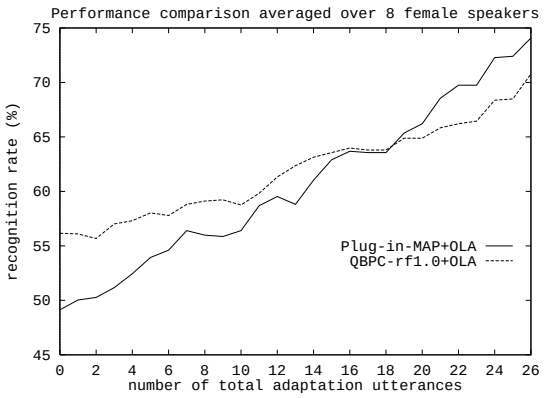

Figure 2: Performance (word accuracy in %) comparison averaged over 8 female speakers on English letter recognition task as a function of amount of adaptation data among methods by combining on-line adaptation with plug-in MAP decoding and QBPC decoding (3 EM iterations for both on-line adaptation and QBPC decoding)

of adaptation data. By combing OLA with QBPC decoding, the performance is further improved before certain point where a good enough model parameter estimation warrants the plug-MAP decision to surpass QBPC.

5.3. Experimental Results on TI20

Similar to Figure 2, Figure 3 shows the performance comparison averaged over 8 female speakers on TI20 task as a function of *total* number of OLA tokens among several methods. The similar facts as the above are also observed here. In QBPC decoding, we can further set the refreshing coefficient rf (see [1, 3] for the explanation) of the hyperparameters to control the degree of the uncertainty of the CDHMM parameters. So, in this figure, we also compare the effect of different rf values on the QBPC performance. The experimental results show that in a reasonably wide range of values of the control parameters (rf), the QBPC method works equally well, thus suggests that the manual tunning is not crucial.

In Figure 4, we further compare the performance of the QBPC algorithm with a modified minimax decoding method with different EM iterations (see [3] for the explanation). The experimental results show that the QBPC performs much better than the minimax method. We also observe that the minimax method is very sensitive to the different number of EM iterations. Less iterations perform better. However, the QBPC method is not so sensitive to the number of EM iterations, especially in TI20's case.

6. DISCUSSION AND CONCLUSION

The principle behind BPC approach is rather straightforward. Because we assume we have no knowledge about the possible mismatch, we thus rely on a quite general prior PDF to characterize the variability of the CDHMM parameters caused by the possible estimation errors and/or mismatches between training and testing conditions. We try to average out this variability while making decision with BPC. So, several factors will influence the efficacy of the BPC. The first one is the appropriateness of the prior PDF for the mismatch we are compensating. If the prior

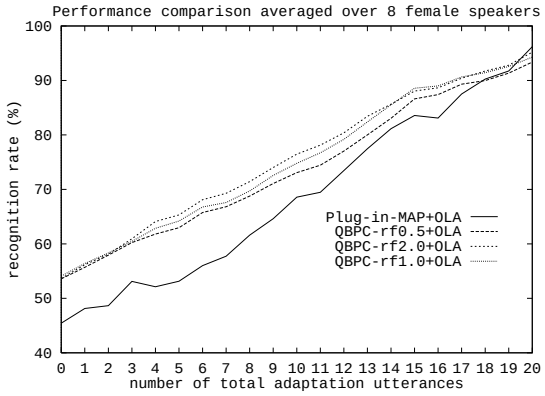

Figure 3: Performance (word accuracy in %) comparison averaged over 8 female speakers on TI20 task as a function of amount of adaptation data among methods by combining on-line adaptation with plug-in MAP decoding and QBPC decoding with different control parameters (3 EM iterations for both on-line adaptation and QBPC decoding)

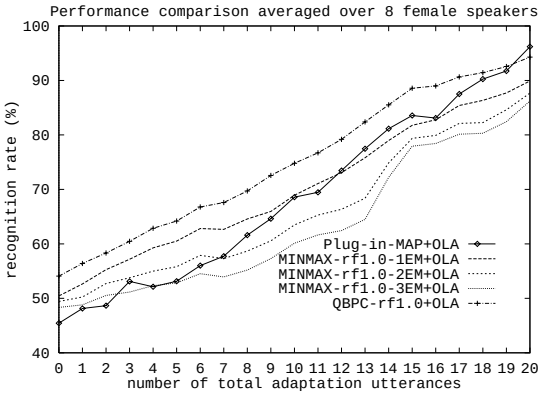

Figure 4: Performance (word accuracy in %) comparison averaged over 8 female speakers on TI20 task as a function of amount of adaptation data among methods by combining on-line adaptation with plug-in MAP decoding, QBPC decoding, and minimax decoding with different EM iterations (3 EM iterations for both on-line adaptation and QBPC decoding)

PDF fails to cover the variability reflected in CDHMM parameters, then BPC will not help much. However, because the BPC procedure does not make rigid assumptions about the possible distortions, consequently, it helps for many distortion types. On the other hand, if we have chance to access some testing data, by combining BPC with OLA, we can make the prior PDF more appropriate. The second factor which greatly influence the BPC performance is the confusability of the classes we are comparing. By using the prior PDF to model the parameter uncertainty, we are also making the classes more overlapping, and thus have the chance to lose some benefit of BPC. This is especially true for confusable vocabulary case which is evidenced by our experimental results. So, BPC will always helps more in a less confusable classes case because we have more chances to use a broader prior PDF to accommodate a higher degree of distortions. The third factor which might matter is the accurateness of the approximation method in QBPC procedure to compute the approximate predictive PDF for classification. The fourth concerns the fact if it is enough to only consider the uncertainty of the mean vectors of CDHMM. There are more theoretical work to do if we want to consider the uncertainty of the other parameters in BPC.

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

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 *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 structural model in structure-based compensation and/or a better *ignorant* model for *semi-blind* compensation, like the prior PDF in BPC. It is also believed to be crucial for efficient adaptation and compensation to formulate and develop the appropriate mathematical tools for discovering a good intrinsic structural model of speech in acoustic, phonetic and linguistic aspects.

The biggest challenge might come from those applications which only involve a couple of utterances, but every utterance involves a distinct "distortion channel" from the intended message to the received signal. How to reliably and efficiently recover and/or extract the interested message from this signal pose a big challenge for the so-called robust ASR in this context.

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*, Vol. 5, No. 2, pp.161-172, 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, 1996.

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